Universitat Politècnica de Catalunya
Department of Signal Theory and Communications

This thesis is submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy (PhD)

Data and Methods for a Visual Understanding of Sign Languages

by

Amanda Cardoso Duarte

Supervised by
Xavier Giró i Nieto & Jordi Torres Viñals

May 2022

Aos meus pais Maria Terezinha e Gilberto

por não medirem esforços em me apoiar

para que eu chegasse até aqui.

Abstract

Signed languages are complete and natural languages used as the first or preferred mode of communication by millions of people worldwide. However, they, unfortunately, continue to be marginalized languages. Designing, building, and evaluating models that work on sign languages presents compelling research challenges and requires interdisciplinary and collaborative efforts. The recent advances in Machine Learning (ML) and Artificial Intelligence (AI) have the power to enable better accessibility to sign language users and narrow down the existing communication barrier between the Deaf community and non-sign language users. However, recent AI-powered technologies still do not account for sign language in their pipelines. This is mainly because sign languages are visual languages, that use manual and non-manual features to convey information, and do not have a standard written form. Thus, the goal of this thesis is to contribute to the development of new technologies that account for sign language by creating large-scale multimodal resources suitable for training modern data-hungry machine learning models and developing automatic systems that focus on computer vision tasks related to sign language that aims at learning better visual understanding of sign languages.

Thus, in Part I we introduce the How2Sign dataset, which is a large-scale collection of multimodal and multiview sign language videos in American Sign Language. In Part II, we contribute to the development of technologies that account for sign languages by presenting in Chapter 4 a framework called Spot-Align, based on sign spotting methods, to automatic annotate sign instances in continuous sign language. We further present the benefits of this framework and establish *sign language recognition* baselines on the How2Sign dataset. In addition to that, in Chapter 5 we benefit from the different annotations and modalities of the How2Sign to explore *sign language video retrieval* by learning cross-modal embeddings. Later in Chapter 6, we explore *sign language video generation* by applying Generative Adversarial Networks to the sign language domain and assess if and how well sign language users can understand automatically generated sign language videos by proposing an evaluation protocol based on How2Sign topics and English translation.

Resum

Les llengües de signes són llengües completes i naturals que utilitzen milions de persones de tot el món com mode de comunicació primer o preferit. Tanmateix, malauradament, continuen essent llengües marginades. Dissenyar, construir i avaluar tecnologies que funcionin amb les llengües de signes presenta reptes de recerca que requereixen d'esforços interdisciplinaris i col·laboratius. Els avenços recents en l'aprenentatge automàtic i la intel·ligència artificial (IA) poden millorar l'accessibilitat tecnologògica dels signants, i alhora reduir la barrera de comunicació existent entre la comunitat sorda i les persones no-signants. Tanmateix, les tecnologies més modernes en IA encara no consideren les llengües de signes en les seves interfícies amb l'usuari. Això es deu principalment a que les llengües de signes són llenguatges visuals, que utilitzen característiques manuals i no manuals per transmetre informació, i no tenen una forma escrita estàndard. Els objectius principals d'aquesta tesi són la creació de recursos multimodals a gran escala adequats per entrenar models d'aprenentatge automàtic per a llengües de signes, i desenvolupar sistemes de visió per computador adreçats a una millor comprensió automàtica de les llengües de signes.

Així, a la Part I presentem la base de dades How2Sign, una gran collecció multimodal i multivista de vídeos de la llengua de signes nord-americana. A la Part II, contribuïm al desenvolupament de tecnologia per a llengües de signes, presentant al capítol 4 una solució per anotar signes automàticament anomenada Spot-Align, basada en mètodes de localització de signes en seqüències contínues de signes. Després, presentem els avantatges d'aquesta solució i proporcionem uns primers resultats per la tasca de *reconeixement de la llengua de signes* a la base de dades How2Sign. A continuació, al capítol 5 aprofitem de les anotacions i diverses modalitats de How2Sign per explorar la *cerca de vídeos en llengua de signes* a partir de l'entrenament d'incrustacions multimodals. Finalmet, al capítol 6, explorem la *generació de vídeos en llengua de signes* aplicant xarxes adversàries generatives al domini de la llengua de signes. Avaluem fins a quin punt els signants poden entendre els vídeos generats automàticament, proposant un nou protocol d'avaluació basat en les categories dins de How2Sign i la traducció dels vídeos a l'anglès escrit.

Resumen

Las lenguas de signos son lenguas completas y naturales que utilizan millones de personas de todo el mundo como modo de comunicación primero o preferido. Sin embargo, desgraciadamente, siguen siendo lenguas marginadas. Diseñar, construir y evaluar tecnologías que funcionen con las lenguas de signos presenta retos de investigación que requieren esfuerzos interdisciplinares y colaborativos. Los avances recientes en el aprendizaje automático y la inteligencia artificial (IA) pueden mejorar la accesibilidad tecnológica de los signantes, al tiempo que reducir la barrera de comunicación existente entre la comunidad sorda y las personas no signantes. Sin embargo, las tecnologías más modernas en IA todavía no consideran las lenguas de signos en sus interfaces con el usuario. Esto se debe principalmente a que las lenguas de signos son lenguajes visuales, que utilizan características manuales y no manuales para transmitir información, y carecen de una forma escrita estándar. Los principales objetivos de esta tesis son la creación de recursos multimodales a gran escala adecuados para entrenar modelos de aprendizaje automático para lenguas de signos, y desarrollar sistemas de visión por computador dirigidos a una mejor comprensión automática de las lenguas de signos.

Así, en la Parte I presentamos la base de datos How2Sign, una gran colección multimodal y multivista de vídeos de lenguaje la lengua de signos estadounidense. En la Part II, contribuimos al desarrollo de tecnología para lenguas de signos, presentando en el capítulo 4 una solución para anotar signos automáticamente llamada SPOT-ALIGN, basada en métodos de localización de signos en secuencias continuas de signos. Después, presentamos las ventajas de esta solución y proporcionamos unos primeros resultados por la tarea de *reconocimiento de la lengua de signos* en la base de datos How2Sign. A continuación, en el capítulo 5 aprovechamos de las anotaciones y diversas modalidades de How2Sign para explorar la *búsqueda de vídeos en lengua de signos* a partir del entrenamiento de incrustaciones multimodales. Finalmente, en el capítulo 6, exploramos la *generación de vídeos en lengua de signos* aplicando redes adversarias generativas al dominio de la lengua de signos. Evaluamos hasta qué punto los signantes pueden entender los vídeos generados automáticamente, proponiendo un nuevo protocolo de evaluación basado en las categorías dentro de How2Sign y la traducción de los vídeos al inglés escrito.

Acknowledgements

I would like to start by deeply thanking my dear advisors, Xavier and Jordi. Xavi, thank you for believing in me since day one, for saying yes to all the crazy ideas, for letting me explore research and find what I love to do, but also for always being there to guide me to the right path. Thank you for inspiring and shaping me to be the researcher I am today. Jordi, thank you for not hesitating to support our ideas and for making sure that I always had the resources I needed to develop my research. Thank you for believing in me and for choosing to support my research. I will always be grateful for it.

Thank you Prof. Alan Smeaton, Prof. Elisa Sayrol, Dr. Gregory Rogez and Dr. César Roberto De Souza for agreeing to take part in my thesis jury and for your valuable feedback that helped improving this document.

Thank you to the Barcelona Supercomputing Center and the Universitat Politècnica de Catalunya for proving me with all the resources and great environment necessary for my work. Thank you to Gisela and all the fellowship team from “la Caixa” and the HR team from BSC and UPC for all the support and exceptional work to make my PhD journey the best possible. Part of my doctoral studies and the How2Sign dataset was also funded by Facebook AI (now Meta AI) through the Caffe2 research award and a research gift. I would like to thank everybody that believed in this project, in special Deepti Ghadiyaram for all her help and support in this process.

My eternal gratitude, admiration and respect goes to the Deaf communities around the world, but in special to some very special people that I had the privilege, and the luck to learn from and work with during this journey. My most special thank you goes to Kenneth De Haan, from Gallaudet University, who taught me patiently everything I know about the Deaf culture and sign language. I thank him for guiding me and introduced me to so many lovely people that ended up being an important part of this project. Thank you so so much to Meg, Andrea, JoAnna, Brandon, Doug, Eric, Ashley M., Chris, Dave, Jeremy, Mary Jean, Caleb, Cheyanne, Michelle, Ashley W., Joshua, Rebecca, Kathleen and all the people that direct and indirectly contribute to the How2Sign dataset.

A big part of this work was just possible with the help and collaboration of the undergraduate and master's students at UPC with whom I have worked through these four years. I would like to thank all of you and also all my colleagues at BSC. Thank you for letting me be part of your journey and for making mine more fun and lighter working by my side. And an enormous thank you here goes to my Ph.D siblings Miriam and Victor, I will be forever grateful for our time together.

Thank you to all the members of the Grounded Sequence-to-Sequence Transduction team at JSALT 2018, especially to Lucia, Florian, Desmond and my sleepy peeps (Shruti, Ozan and Ramon) for the great time I had at JHU. Thank you also to the team at LTI at CMU, specially to Jessica for all her help during my stay and beyond. Thank you also to all the IMAGINE team at ParisTech for welcoming me and making my stay unforgettable.

A big thank you to all my internal and external mentors and collaborators, Gül Varol, Samuel Albanie, Marta Ruiz, Amaia Salvador, Didac Suris, Lucas Ventura, Miquel Tubau and Laia Tarres. Across late nights, failed experiments, all-too-soon deadlines, grant proposals, and endless redrafting of papers and rebuttals – in a very real way, this work exists because of you. Thank you for everything.

And behind the scenes, when reality knocked the door and hard times arrived (and, oh boy, they arrived), he was always there holding my hand and telling me that everything would be alright, and it is. Yannis, thank you not just for being my life partner and my biggest fan, but also for being the most loving and carrying person I ever met. Words will never be enough to express my gratitude for having you by my side, but they can start expressing how thankful I am for all the times you checked my drafts, helped me with ideas, and proofread my papers. Thank you for being my inspiration, for showing me that research can be fun and something to be passionate about. Thank you for sharing a home, a life and research interests with me. I ~~love~~ hate you!

E com certeza isso tudo não seria possivel sem as pessoas que me colocaram no mundo e me apoiaram desde sempre. Obrigada, Maria Terezinha and Gilberto, vocês são o motivo de eu estar aqui hoje. Palavras nunca serão o suficiente para descrever tudo o que vocês são para min, nem para agradecer por tudo que vocês me proporcionaram. Obrigada por tudo e principalmente por me entender e me apoiar quando eu decidi voar pra longe. Apesar de longe vocês são e sempre serão meu principal porto seguro e a razão por eu lutar a cada dia. Obrigada também a minha familia, Andréia, Marcos, Mariana, Tiago, Débora e Maria luisa por estar, mesmo que de longe, sempre presente na minha vida e me apoiando. Marcos, a nossa idea deu certo, que venha a próxima!

”la Caixa” Foundation

The project that gave rise to these results received the support of a fellowship from ”la Caixa” Foundation (ID 100010434) under the fellowship code LCF/BQ/DI18/11660029.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713673.

Important Note From the Author

I, the author of this dissertation, would like to discuss in this document an important part of this work that is not related to the technical contributions that this dissertation brings. To begin with, I would like to bring the awareness that sign languages are more than just languages themselves. I was fortunate to learn from the community that their language is part of their culture and how they proudly identify in the world, as a Deaf person. But at the same time, I also learned that, unfortunately, for many years, hearing people prohibited deaf and hard-of-hearing people from using their own language to communicate.

I, as a non-deaf person and not a member of the Deaf community, understand that I do not have the rights to simply use such cultural property for my own benefit. With this document, I would like to disclaim that such intention is far from what this work represents to me. My intention here was always to somehow help break, even that for just a little, the communication, accessibility and equality barrier that still exists in our society. I sincerely expect that this work can create opportunities for the members of the Deaf community can continue carrying out research that deals with their language and culture, which today seems to not be the reality in our field.

I believe that accessibility and inclusion should be topics more present, discussed and addressed in the fields of Computer Vision and Machine Learning, but *always*, with the inclusion of the members of the community in which the systems would be designed for, as countless times pointed out by the communities themselves. In this work, I prioritized the inclusion and participation of members of the Deaf community, most of them located in the United States, where ASL is used, and I could not be better welcomed by them.

Here, I would like to thank from the bottom of my heart everyone that directly or indirectly showed and patiently thought me part of their culture and language so I could understand better the problems that were tackled in this dissertation.

Contents

List of Figures

List of Tables

Acronyms

2D 2-dimensional

3D 3-dimensional

ASL American Sign Language

AI Artificial Inteligence

ASR Automatic Speech Recognition

BASL Black American Sign Language

BLEU Bilingual Evaluation Understudy

BSL British Sign Language

CM Cross-modal

CSL Chinese Sign Language

CSLR Continuous Sign Language Recognition

CV Computer Vision

DSGS Swiss-German Sign Language

DGS German Sign Language

FSL Finnish Sign Language

GT Ground Truth

GSL Greek Sign Language

HD High Definition

ISL Indian Sign Language

K-RSL Kazakh-Russian Sign Language

LSF French Sign Language

ML Machine Learning

NLP Natural Language Processing

SL Sign Language

SLP Sign Language Production

SLR Sign Language Recognition

SLT Sign Language Translation

SSL Swedish Sign Language

TİD Turkish Sign Language

VGT Flemish Sign Language

SR Sign Recognition

GAN Generative Adversarial Network

LSGAN Least-Squares Generative Adversarial Network

MFCC Mel-frequency Cepstral Coefficients

DTW Dynamic Time Warping

V2T Video-to-text retrieval

T2V Text-to-video retrieval

GPU Graphics Processing Unit

PDK Percentage of Detected Keypoints

PCK Percentage of Correct Keypoints

MOS Mean Opinion Score

Glossary

Annotations: Data annotation is the process of labeling different aspects of a data structure (it can be text, image, video, etc.) of a dataset.

Audism: It is an attitude based on pathological thinking that results in a negative stigma toward anyone who does not hear.

Bilingual Evaluation Understudy (BLEU) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another.

Dataset: Refers to a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer.

Co-articulation: is a naturally occurring situation in which a word (in the form of a speech sound or a sign) is influenced by and becomes more like the preceding or the following word.

Continuous signing: Specifies the nature of sign language datasets that contain long phrases or full sentences as opposed to single, isolated signs.

deaf: Refers the hearing status of a person. They have profound hearing loss, which implies very little or no hearing.

Deaf: Refers to members of any Deaf community.

Fingerspelling: Refers to the representation of an alphabet of a spoken language, where every letter in the alphabet has a corresponding (static or dynamic) sign.

Gloss: It is used in linguistics to transcribe signs using spoken language words.

Hard-of-hearing: Refers to people with hearing loss ranging from mild to severe.

Hearing loss: A person with hearing loss is not able to hear as well as someone with normal hearing (hearing thresholds of 20 decibels or better in both ears). It can affect one ear or both ears and leads to difficulty in hearing conversational speech or loud sounds.

Isolated signs: Specifies the nature of sign language datasets that only contains single signs as opposed to long phrases or full sentences.

Keypoints: Refers to the human joints represented by 2D or 3D coordinates. They can be captured using special equipment attached to a person or automatically from a video recording. In this thesis, we refer to skeleton the representation that was automatically extracted from the video recordings.

Lexicon: The complete set of meaningful units in a language.

MedR: Median Rank, lower is better

Pose: Refers to human pose information automatically extracted using computational systems. It is usually represented by 2D or 3D coordinates that refer to the human body joints.

R@K: Recall at rank K, higher is better

Sign language features/articulators: Each sign consists of a set of articulators or features. It consists of manual and non-manual features, *e.g.* Handshape, orientation, location and movement are the four manual parameters, while non-manual articulators include head and body posture, facial expression, eye gaze, and mouth patterns.

Signers: Refers to sign language users.

Skeleton: Refers to the wired visualization of the human body representation. This representation can be via 2D or 3D coordinates (or also called *Keypoints*) and can be captured using special equipment attached to a person or automatically from a video recording. In this thesis, we refer to the skeleton the representation that was automatically extracted from the video recordings.

Syntax: The arrangement of words and phrases to create well-formed sentences in a language.

Utterance: Refers to an uninterrupted sequence of a language.

Vocabulary: The set of unique signs (or words) that occur in a dataset.

1 Introduction

Hearing loss is the most common communication disorder affecting about 360 million people worldwide to different degrees, according to the World Health Organization [10]. A substantial part of these individuals use sign language (SL) to communicate. Although recent advancements like the internet, smartphones, and social networks have enabled people to instantly communicate and share knowledge at a global scale, deaf and blind people still have very limited access to large parts of the digital world. This thesis aims to contribute to the research field of sign language understanding by collecting and curating sign language data resources and employing it in the development of different computer vision tasks with such focus. We believe that new resources and technologies that focus on sign language understanding have the potential of removing or reducing the difficulties and barriers that deaf people encounter in their daily lives when interacting with non-sign language users.

In this thesis, we refer to "sign language understanding" as the *semantic* understanding of signing videos at the level required for solving basic vision and language tasks. We therefore use the term to refer to methods that are able to extract the linguistic cues necessary and sufficient to successfully perform tasks such as sign language video retrieval, sign spotting, sign language recognition, translation or production. This is a narrow use of the term "understanding" that of course does not reflect the richness and breadth of a highly complex visual languages such as sign languages.

Throughout the text, we use terminologies such as hearing loss, hard-of-hearing and, deaf that need to be defined and clarified for better understanding. The following definitions are established by the World Health Organization [10]. A person with *hearing loss* is not able to hear as well as someone with normal hearing (hearing thresholds of 20 decibels or better in both ears). Hearing loss may be mild, moderate, severe, or profound. It can affect one ear or both ears, and leads to difficulty in hearing conversational speech or loud sounds. *Hard-of-hearing* refers to people with hearing loss ranging from mild to severe. People who are hard-of-hearing usually use a mix of communication through spoken language and signed languages and can benefit from hearing aids, cochlear implants, and other assistive devices as well as captioning. A *deaf* [1] a person mostly has profound hearing loss, which implies very little or no hearing. Their primary means of communication is through sign language.

In the following sections we present the goals, motivations and contributions of this thesis followed by the outline of the rest of this document.

1.1 Goals

Our first goal is to create sign language resources, such as a large-scale multimodal dataset, suitable for training modern machine learning models. The creation of such datasets has the potential to instigate the advance in the area of research that involves Computer Vision (CV) tasks that focus on sign language, such as sign language recognition, translation and production and etc. Thus, in Part I we describe the multimodal data collection process that led to the creation of the *How2Sign* dataset.

Second, we aim at contributing to the development of automatic systems that account for sign language in their pipeline. Towards that, in Part II we tackle different tasks that include (i) automatic sign language data annotation, (ii) sign language recognition, (iii) sign language video retrieval, (iv) sign language video generation, and, (v) understanding if and how sign language users perceive automatic generated sign language videos. More specifically, we present in Chapter 4 a framework called Spot-Align, based on sign spotting methods, that is designed to automatically annotate sign instances across a broad vocabulary in continuous sign language. We further access the benefits of this framework and establish *sign language recognition* baselines on the How2Sign dataset. In Chapter 5 we benefit from the different annotations and modalities of the How2Sign to explore *sign language video retrieval* by learning cross-modal embeddings. Later in Chapter 6, we explore *sign language video generation* by applying Generative Adversarial Networks to the sign language domain in order to transfer the sign language motion from one individual into another. With this approach we aim to access the level of understanding sign language users have regarding automatically generated sign language videos and propose an evaluation protocol based on How2Sign topics English translation.

[1] We follow the recognized convention of using the upper-cased word Deaf which refers to the culture and describes members of the community of sign language users and the lower-cased word deaf describes the hearing status[11].

1.2 Motivations

The majority part of Deaf or hard-of-hearing individuals use sign language as their primary means of communication. Different countries and even different regions of the same country have developed their own sign language, that account for more than 300 different sign languages in use nowadays [12].

A person who is deaf from birth, usually acquires a sign language as their primary language, and eventually (but not necessarily), learns spoken language as their second or third language. In fact, in such cases, spoken language is a particularly hard-to-learn second language for deaf individuals– it must be learned based only on a set of written symbols and based on observations of highly ambiguous mouth patterns, without any auditory cues. As a consequence, many deaf children leave school with significant difficulties in writing and reading spoken languages [13].

For most deaf individuals, the interaction with non-sign language users in education, employment, healthcare, legal settings, entertainment, or even while watching online videos are usually challenging tasks. To overcome this communication barrier, an alternative is to employ a sign language interpreter to translate the spoken language into the sign language used in the situation. However, more than 80% of people who use sign language to communicate live in developing countries and have limited access to interpreters. In addition to that, when interpreters are available, it is important that they are trained, qualified, and certified to ensure quality standards, which comes with a high cost that not everybody can easily afford. In more casual circumstances, such as at restaurants, or daily interactions, most Deaf people often prefer to be independent, communicating with non-sign language users via gestures or writing.

When it comes to the digital world, some streaming platforms and broadcast services provide accessibility options such as captions or audio descriptions. However, these are available just for a part of the catalog and often in a limited amount of languages. When they are not available, volunteers or relatives may generate and distribute them through third-party platforms. However, a large portion of online videos are not from streaming or broadcast services but are generated by amateur users. As reported by a video stream platform, an average of 400 hours of videos are uploaded every day on a common video-sharing website. These users do not typically create any metadata for accessibility. Their intention is informal, addressed to a reduced audience and, produced in a very short time. The huge and growing amount of such online videos requires automatic methods capable of adapting these contents across modalities to make them more accessible to everybody.

To address such issues, researchers from different areas have recently started including sign language in the developments of new technologies. Sign Language understanding (also called Sign Language Processing by some authors [1, 14, 15]) is an emerging field of artificial intelligence that aims to automatically process, analyze and extract the semantic meaning of sign language content. It lies at the intersection of both Natural Language Processing (NLP) and computer vision, and frequently involves three base tasks: (i) Sign Language Translation (SLT), which targets an automatic translation from sign language representations (*e.g.* videos or poses) to spoken language text; (ii) Sign Language Production (SLP), which targets the generation of sign language representations (usually poses or SL annotations) or; (iii) Sign Language Recognition (SLR) where individual signs are just recognized and classified. However, we believe that the development of new challenging and interesting tasks that involve sign language, and the engagement of the research community on creating more resources and new methods are crucial and necessary next steps in order to overcome the challenges that sign language brings to the related research areas.

The development of resources and methods that account for signed languages in their pipeline can enable several real-world applications, such as: (i) systems that can better document endangered sign languages; (ii) the development of educational tools for sign language learners; (iii) tools that can query and retrieve information from signed language videos; (iv) the development of personal assistants that react to signed languages; (v) real-time automatic sign language translation and production, which can assist sign language users in daily interactions with non-sign language users, and more.

Although such applications can benefit different groups of people, it is important to remember that signed languages are and will always be an important part of the culture and lives of members of the Deaf communities. Here we would like to re-state that, when addressing this research area, researchers should work alongside and under the direction of Deaf communities, and prioritizing the benefit of the signing communities' interest above all [1, 14, 16, 17].

While investing in the development of resources and methods that address sign language can offer several benefits to society and challenging research topics to be addressed, signed languages present several technical and/or linguistic challenges that we discuss next. It is important to note that since sign languages are different languages among them, the challenges presented next can vary from language to language. We mostly present the ones we dealt with while working with American Sign Language, and British Sign Language but that are usually common to other sign languages.

From the challenges presented next, in this thesis, we explicitly focus on mitigating the first two challenges (lack of data and representation of sign languages), and deal implicitly with the rest by improving the visual understanding of sign languages via the proposed methods.

1.3 Challenges

Lack of data. With the availability of large amounts of labeled training data, deep learning models have shown excellent performance on some generic computer vision and natural language processing tasks such as human action detection/recognition and neural machine translation. However, when it comes to fields where specific and detailed data are required, the lack of reliable large-scale datasets is still a big challenge. This challenge is even bigger when dealing with low-resources languages, such as signed languages. A language is considered a low-resource when there is no accessible and public available documentation and data to be explored. In the case of signed languages, public available datasets are very scarce as it will be demonstrated in Section 2.2 with an extended overview of existing sign language datasets. When available, they unfortunately present several drawbacks, such as:

Small size: Sign language corpora needed to fuel the development of sign language-related tasks are still several orders of magnitude smaller than their spoken language counterparts, typically containing fewer than 100,000 articulated signs. (See Table 2.1 for detailed information about existing sign language datasets.)

Presence of continuous signing: Many datasets of sign languages only contain individual signs. While isolated signs may be important for certain scenarios, most real-world use cases of sign language tasks involve natural conversational with complete sentences.

Participation of native signers: Due to the difficulties of recruiting signers who are native in a specific signed language, many existing datasets allow people that are still learning the language (*i.e.* non-deaf students) to participate in the recordings or include data scraped from online sources (*e.g.* social media or video platforms) where information about the signer and theirs language skill are unknown. Other datasets also include professional interpreters, who are highly skilled but are often not native signers. Very frequently, interpreters can also change the execution of the language (*e.g.* by simplifying the narrative and vocabulary or signing slower for a better understanding) since this is a common practice in their daily jobs. Datasets of native signers are needed to build models that presents this core user group.

Limited signer variety: Signed languages are used by different groups of people and naturally differ from person to person. To accurately represent the signing population and assure realistic scenarios, sign language datasets should include a larger variety of signers with different: gender, ages, clothing, geography, culture, skin tone, body proportions, language fluency, video background, lighting conditions, camera quality, and camera angles. It is also very important to generate signer-independent datasets, that allow measuring the generalizability of models by training and testing on different signers.

Recorded in controlled settings: Current datasets are usually recorded in a laboratory or a TV broadcast setting, where illumination is constant and the usage of language is carefully chosen (by the fact that signers know they are being recorded). In addition to that, data from TV broadcasts are usually sign language interpretation, which as mentioned above, can change the execution of the language. Data that reflect real-world scenarios are needed to train models that can be used outside controlled scenarios.

Although the drawbacks of current datasets are noticeable and already pointed out in the literature [14, 18], creating larger, more representative and public sign language video datasets is far from being an easy task. Many Deaf signers are not comfortable with being recorded while signing, because of privacy and other personal reasons. Also, when subjects are willing to collaborate in the data collection, many find it a long and tedious process which makes the continuity of the task *hard, slow and, expensive.* Collection and annotation of sign language data can take up to 600 minutes for each minute of video data [19]. Moreover, annotation usually requires a specific set of knowledge and skills, which makes recruiting or training qualified annotators very challenging. Additionally, there is little existing signed language data in the wild in video platforms that can be used, especially from native signers that are not interpretations of speech data, as in weather forecast broadcasts. Therefore, data collection often requires massive efforts from data collectors as well as high costs of on-site recordings.

Representation of sign languages is a significant challenge when dealing with sign language data. Unlike spoken languages, signed languages have no widely adopted written form which prevents the easy adaptation of existing methods that work with spoken languages to also work with signed languages. Exploring a standard representation of sign languages can benefit the adaptation of existing models and instigate the improvement of recent developed technologies. Below we describe the current forms that sign languages are represented.

Video. Although videos are the most common and straightforward way of representing signed language, they are not an easy and convenient data type to work with. They are expensive to record, store, transfer, and encode and usually comes with unnecessary information that need to be processed out. In addition to that, when thinking about privacy, videos may also record the signers face and physical characteristics which can limit the possibility of making the video data publicly available [20].

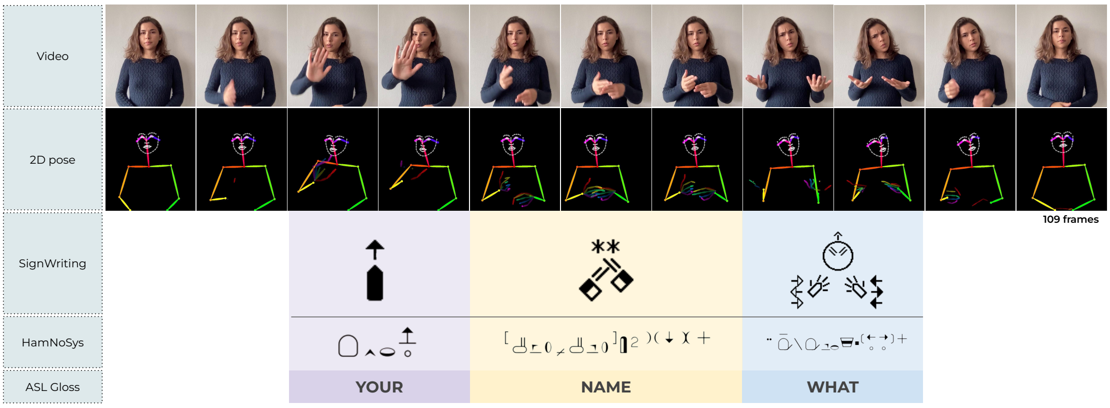


Figure 1.1: **Sign Language representations**. Different representations of a sentence in American Sign Language. Here we represent the sentence with video frames, 2D pose estimation, SignWriting, HamNoSys and glosses. The English translation is "What is your name?". Image adapted from [1].

Poses can be an alternative to reduce the computational cost of working with raw videos. They can be seen as a representation of the visual cues in a skeleton-like or mesh form that corresponds to the location of the human joints in the frame. These locations can be captured via special equipment or directly from video frames. While motion capture equipment can often provide better quality pose estimation, they are still very expensive and intrusive. An alternative to that are pose estimation methods from monocular videos [21–25]. Compared to video representations, *accurate* poses are lower in complexity and keep the signer's face anonymized, while observing relatively low information loss [26]. However, they remain a continuous, multidimensional representation that differs from the type of data that recent computer vision and natural language processing models expect. In addition to that, due to the fast movements of the signer's hand and the motion blur caused by that, state-of-the-art video pose estimation models [23, 25] often fail to predict the pose of the hands in sign language videos.

Written notation systems aim to represent signs as discrete visual features. While different notation systems have been proposed [27–29], there is still no standard widely adopted by any sign language community. The absence of such standard notaion system inhibits the exchange and unification of resources (such as sign language data) and models for different tasks. Figure 1.1 shows examples of two notation systems used by a few projects in sign language research: SignWriting [27], a two-dimensional pictographic system, and HamNoSys [28], a linear stream of graphemes that was designed to be readable by machines.

Glossing is a tool used by linguists to transcribe signed languages sign-by-sign using spoken language words together with sign language specific notations. While different gloss annotation guidelines have been recently provided [30–32], as in sign language writing notations systems, there is still no single standard way of transcribing signed language using gloss transcriptions. Gloss transcription has been adopted as an intermediate representation for machine learning solutions for translation and production. However, it is important to note that glosses are an incomplete way of representing signed language because they do not adequately capture all information expressed simultaneously through different cues (*i.e.* body posture, eye gaze and facial expressions) or depiction, and spatial relationship. This limitation leads to an inevitable information loss at the semantic level that affects downstream tasks, such as sign language translation [33].

Use of multiple articulators to convey information. Different from spoken languages, which primarily use the oral-auditory modality, signed languages use the visual-gestural modality to convey information. Signed languages rely on multiple manual and non-manual articulators such as the face, hands, body of the signer [34], and the space around them to create distinctions in meaning. Their use can be explicit or very subtle and all of them must be taken into account during their modeling and use in computational models to fully capture the meaning of a signed sentence. We present more information about the elements of sign languages in Section 2.1.

Co-articulation is a naturally occurring situation in which a word (in the form of a speech sound or a sign) is influenced by and becomes more like the preceding or a following word. Unlike in speech recognition [2], the influence of coarticulation in sign languages is over longer duration and simultaneously impacts different aspects of the sign in terms of the location, palm orientation, hand shape and movement. Due to this phenomenon, the appearance of a sign, especially the hand location at the beginning and end of the sign, can be significantly different under different sentence contexts making the recognition or natural production of signs in sequences (*e.g.* in a sentence) a hard task. In natural conversations, with the use of continuous signing and co-articulated signs, the speed at which a signer conveys information is also faster than when a signer performs an isolated sign. Thus, when recorded by a camera, continuous signing can also contain motion blur, which can be a problem specifically for computer vision techniques.

Sign variants. Sign languages can have several variants for the same meaning or word. The different variants may depend on the region that the language is being used, the signer's social background, the context in which the sign was used (formal or informal variants) etc. To train robust computational models, sign language datasets need to have a recorded amount of all the variants of each sign, which can be challenging to obtain.

Depiction is the act of representing or enacting information, an act, a dialogue, or a psychological event in sign language [3, 35]. This representations can also be done by the

[2] Coarticulation in speech recognition refers to changes in the speech articulation of the current speech segment due to neighboring speech.

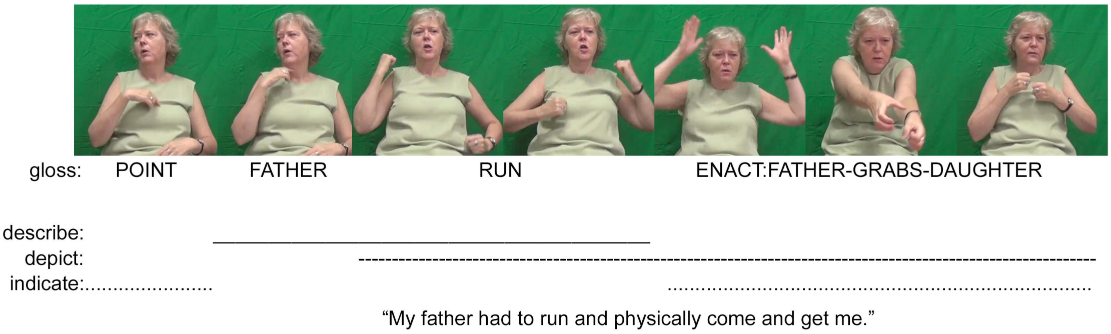

gloss:
POINT
FATHER
RUN
ENACT:FATHER-GRABS-DAUGHTER
describe:
depict:
indicate:
“My father had to run and physically come and get me.”

Figure 1.2: **Example of depiction in Norwegian Sign Language**. In this example from [2] the signer is telling a personal experience about her childhood where she narrates how her father would have to physically come and find her when she was out playing. We thank Ferrara and Halvorsen 2018 [3] for the example image.

use of *classifiers* [3]. As mentioned before, different sign languages present different components and can vary from language to language [4]. In Figure 1.2 we present an example of depiction in Norwegian Sign Language (NSL) where part of an informal conversation from [2] is illustrated. In this example, the signer is telling a personal experience about her childhood where she narrates how her father would have to physically come and find her when she was out playing because she could not hear his calls [5].

She begins with the signs “POINT FATHER”, indicating the “father” as the actor referent. Here, the pointing action serves to indicate that she is talking about her father. The signer then elaborates on her father’s actions by using the sign “RUN” to express how her father would have to run (and find her). She ends by enacting how her father would be physically grab her (and bring her close). Here, she indicates herself as a referent through eye gaze and meaningful use of space. In this example, apart from the use of depiction to illustrate the sentence’s verb, we can also notice how the signer used subtle shifts in her body positioning and eye gaze to indicate the referents. Correctly recognizing and identifying the occurrence of depiction, classifiers and the subtle non-manual components is crucial for an accurate sign language recognition, translation, and production system. However, understanding depiction requires exposure to Deaf culture and linguistics, which the communities driving progress in computer vision generally lack [16]. Another challenge is also how to create depiction annotations. Countless depictions can express the same concept, and annotation systems do not have a standard way to encode this richness in the way sign language is expressed.

[3] More information about classifiers can be seen in Section 2.1

[4] More information about different sign languages can be seen in Section 2.1

[5] This example was taken from the great work of Ferrara and Halvorsen 2018 [3]. We thank the authors for the great explanation and examples of depiction presented in their work

1.4 Contributions

In this Section, we list the publications and the releases of datasets and open-source software published during the course of this thesis. We detail the technical contributions within the main publications in Chapters 3, 4, 5 and 6 and Appendix A and B.

1.4.1 Publications

Here we list the **main publications** included in this thesis. All publications are published in the proceedings of peer-reviewed conferences.

- Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. **How2sign: a large-scale multimodal dataset for continuous american sign language**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021 [36] (Chapters 3 and 6)

- Amanda Duarte, Jordi Torres, and Xavier Giro-i Nieto. **Cross-modal neural sign language translation.** In *Proceedings of the 27th ACM International Conference on Multimedia (ACMMM) - Doctoral Symposium*, 2019 [37] (Chapter 3)

- Amanda Duarte, Samuel Albanie, Xavier Giro-i Nieto, and Gül Varol. **Sign language video retrieval with free-form textual queries.** *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022 [38] (Chapters 4 and 5)

- Didac Surís, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i Nieto. **Cross-modal embeddings for video and audio retrieval.** In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops - Sight and Sound Workshop*, 2018 [39] (Appendix A)

- Amanda Duarte, Francisco Roldan, Miquel Tubau, Janna Escur, Santiago Pascual, Amaia Salvador, Eva Mohedano, Kevin McGuinness, Jordi Torres, and Xavier Giro-i-Nieto. **Wav2pix: Speech-conditioned face generation using generative adversarial networks.** In IEEE *International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, 2019 [40] (Appendix B)

As a product of other research activities

Journal publication product of a collaboration during the course of this dissertation:

- Lucia Specia, Loic Barrault, Ozan Caglayan, Amanda Duarte, Desmond Elliott, Spandana Gella, Nils Holzenberger, Chiraag Lala, Sun Jae Lee, Jindrich Libovicky, et al. Grounded sequence to sequence transduction. *IEEE journal of selected topic sin signal processing*, 2020 [41]

List of works developed in the scope of this thesis by other students with my collaboration:

- Peter Muschick. Learn2Sign: Sign language recognition and translation using human keypoint estimation and transformer model. Master's thesis, Universitat Politecnica de Catalunya, 2020 [42]

- Pol Pérez Granero. 2d to 3d body pose estimation for sign language with deep learning. Bachelor's thesis, Universitat Politecnica de Catalunya, 2020 [43]

- Miquel Tubau. WAV2PIX: Enhancement and evaluation of a speech-conditioned image generator. Master's thesis, Universitat Politecnica de Catalunya, 2019 [44]

- Sandra Roca. Block-based speech-to-speech translation. Bachelor's thesis, Universitat Politecnica de Catalunya, 2018 [45]

- Janna Escur i Gelabert. Exploring automatic speech recognition with tensorflow. Bachelor's thesis, Universitat Politecnica de Catalunya, 2018 [46]

1.4.2 Datasets and open-source softwares

The How2Sign dataset. We have publicly released the How2Sign dataset (https://how2sign.github.io/) as part of the publication presented in Chapter 3 [36] in collaboration with Carnegie Mellon University, Facebook AI and Gallaudet University. How2Sign is the largest publicly available multimodal and multiview continuous American Sign Language (ASL) dataset to date, that includes sign language videos and a set of corresponding modalities such as speech, English transcripts, depth and other manual and automatic annotations. The dataset further includes a three-hour subset recorded in the CMU Panoptic studio [6] enabling detailed 3D pose estimation. How2Sign has the potential to impact a wide range of sign language processing tasks, such as sign language recognition, retrieval, translation and production, as well as wider multimodal and computer vision tasks like 3D human pose estimation.

The Youtubers dataset. We publicly released part of the Youtubers dataset (https://imatge-upc.github.io/wav2pix/) as part of the publication presented in

[6] http://www.cs.cmu.edu/~hanbyulj/panoptic-studio/

Appendix B [40]. The Youtubers dataset is an audio-visual dataset containing 168,796 seconds of speech with the corresponding video frames, and cropped faces from a list of 62 youtubers active during the past few years. The dataset is gender-balanced and manually cleaned keeping 42,199 faces, each with an associated 1-second speech chunk.

Software. The following works presented in this thesis were made open-source and publicly released:

- **Sign language video retrieval with free-form textual queries:** The code, model and annotations for sign language video retrieval will be released as part of the project presented in [38] (Chapter 5). The project page can be found at: `https://imatge-upc.github.io/sl_retrieval/`
- **Cross-modal embeddings for video and audio retrieval:** The code for video and audio retrieval is released as part of the project presented in [39] (Appendix A). The code can be found at: `https://github.com/surisdi/youtube-8m`
- **Wav2Pix:** The code for automatic face generation conditioned to raw speech signal is released as part of the project presented in [40] (Appendix B). The code can be found at: `https://github.com/imatge-upc/wav2pix`

1.5 List of Collaborations and Visits to other Research Groups

This section presents a list of collaborations and visits to other research groups pursued in the course of this dissertation:

- **Johns Hopkins University:** From June to August of 2018 I participate in the Fifth Frederick Jelinek Memorial Summer Workshop (JSALT) at Johns Hopkins University. I was part of the Grounded Sequence to Sequence Transduction team working under the supervision of Prof. Lucia Specia and Prof. Desmond Elliott on language grounding using multiple modalities. This collaboration lead to a journal publication title "Grounded sequence to sequence transduction" [41] mentioned under publications "product of other research activities" (Sec. 1.4.1).
  - Workshop website: `https://www.clsp.jhu.edu/workshops/18-workshop/`
  - Grounded Sequence to Sequence Transduction team website: `https://www.clsp.jhu.edu/workshops/18-workshop/grounded-sequence-sequence-transduction/`
  - Github: `https://github.com/lium-lst/nmtpytorch`

- **Carnegie Mellon University:** From February to August of 2019 I was a visiting student at the Language Technologies Institute (LTI) at Carnegie Mellon University (CMU) under the supervision of Prof. Florian Metze. During this visit, we collect the How2Sign dataset presented as one of the contributions of the this dissertation.
- **Gallaudet University:** During my visit at CMU and throughout this dissertation, I informally collaborated with Dr. Kenneth DeHaan from Gallaudet University.
- **École des Ponts, Univ Gustave Eiffel:** From June to November of 2021 I was a (virtual) visiting student at LIGM - École des Ponts at Univ Gustave Eiffel (ParisTech) under the supervision of Dr. Gül Varol and Dr. Samuel Albanie from Oxford University now at Cambridge University. During my visit we develop the sign language video retrieval with free-form textual queries work that is presented as one of the contributions of this dissertation.

1.6 Dissertation Outline

This thesis consists of seven chapters including this introduction. The main content is divided into two parts as well as two additional appendices.

Background. Chapter 2 presents an overview of the linguistic components of Sign Languages. This overview brings the basic important information that is necessary to understand the languages and the complexity of the topic tackled in this thesis. We also present an overview of public and non public sign language datasets up to date.

Part I: Collecting and Annotating Sign Language Data.

Chapter 3 presents the process of collecting and annotating the sign language video collection that compose the How2Sign. We first define and explain all the different the modalities that compose the How2Sign. Then we describe the video recording process that took place in two different studios followed by an explanation of the manual and automatic annotation process of the signing videos. We later present the final dataset statistics and a discussion on the privacy, bias and ethical consideration. We conclude by presenting our final remarks and the experience acquired by the process of collecting and annotating a large-scale and multimodal dataset.

Part II: Sign Language Meets Computer Vision.

Chapter 4 addresses the annotation scarcity problem and presents our proposed framework, called Spot-Align, that integrates multiple sign spotting methods to automatically annotate significant fractions of the How2Sign dataset with sign-level annotations.

Using the resulting automatic annotations, we further explore and establish sign language recognition baseline for the How2Sign dataset.

Chapter 5 proposes the task of sign language video retrieval with free-form textual queries by learning a joint embedding space between text and sign language videos. We establish baselines for the How2Sign in the proposed task and test our approach on the largely used Phoenix2014T dataset.

Chapter 6 explores the use of Generative Adversarial Networks in the context of sign language video generation and presents a user study that aims to understand if and how well sign language users understand automatically generated sign language videos.

Discussion. We conclude this thesis in Chapter 7 with a summary of contributions followed by a discussion of open problems and future work in the area of sign language understanding and beyond.

In addition to the technical contributions presented in the main body of this thesis, we append two other contributions that do not directly involve the main topic of this thesis (sign language understanding) but were developed as preliminary studies. They paved the way for the development of the main methods presented in Chapters 5 and 6.

Appendix A – Cross-modal audio and video retrieval presents our preliminary work on cross-modal audio-video retrieval with a simple and yet effective model that explores cross-modal embedding for retrieving videos given an audio file or vice-versa.

Appendix B – Sign Language Video Generation presents our approach towards cross-modal image generation, more specifically, the generation of facial images given a raw speech signal. We propose a new speech-conditioned generative adversarial network architecture and a new image-speech dataset, called *Youtubers*.

2

Background

The success of sign language understating systems not only relies on learning the complex linguistic aspects of sign languages, but also understanding the culture of the Deaf communities, making sure that these new technologies are aligned with their needs and desires.

Here, we summarize the background information presenting the basic linguistic components of sign languages (Section 2.1), the history of American Sign Language (Subsection 2.1.2), followed by a brief introduction to the Deaf culture and the marginalization of their languages (Subsection 2.1.3). In addition, we also present an extensive survey of the existing sign language datasets (Section 2.2), where we present the modalities and limitations of the publicly available sign language datasets up to date.

2.1 A Brief Introduction to Sign Languages

Sign languages are complex natural languages that have their own grammatical structure directed by linguistic rules [47] with distinct lexicons (vocabularies) of arbitrary signs (conventional symbols) that are constantly evolving. They should not be confused with gesturing or any sort of mime[1].

Given the complexity of sign languages and the focus of this thesis, here we provide a high-level overview of important elements that compose sign languages, followed by

[1] Although sign languages have been used for thousands of years they have been marked as inferior and not recognized as languages by hearing people throughout history. Until circa of 1980 sign languages used in Deaf communities was wrongly considered to be merely a system of mime and ungrammatical gestures [48].

a few information about American Sign Language and the marginalization of the Deaf communities and their languages. It is important to note that the following information is by no means comprehensive and detailed, but it is intended to provide an overall understanding of the level of difficulty of the problems tackled in this thesis, as well as an overall understanding of sign languages. Furthermore, the notations discussed here can differ between signed languages and may not apply to all of them.

2.1.1 Linguistic Components of Sign Languages

Similar to spoken languages, signed languages can appear in different grammatical levels, such as individual letters (called *fingerspelling*), single sign/word (or also called *isolated sign*) or complete sentences (*continuous signing*). The nomenclature in *italics* is used throughout this dissertation.

Use of multiple articulators. Sign languages are visual-gestural languages that convey information via *manual* and *non-manual markers* [49]. Manual articulators are composed of four basic visual-gestural units, which are: hand shape, hand location, hand movement, and palm orientation [34]. Apart from the hands, non-manual articulators such as the head (nod/shake/tilt), mouth (mouthing), eyebrows, cheeks, face (called facial grammar or facial expressions), and eye gaze are also used and equally important for sign language communication [50]. Below we provide more details about important manual and non-manual articulators and their roles in ASL.

Hand shape refers to the distinctive configurations that the hands take as they are used to form signs [51]. Hand shape is one of five components of a sign, along with location, orientation, movement, and facial-body expression. Different sign languages use of different hand shapes. American Sign Language uses 18 hand shapes for ordinary signs, plus a few marginal hand shapes taken from the American Manual Alphabet for finger spelling [52].

Head movement. The movement of the head supports the semantics of sign language. Questions, affirmations, denials, and conditional clauses are communicated with the help of the signer's head movement.

Facial grammar. Facial grammar does not only reflect a person's emotions, but also constitutes a large part of the grammar and linguistic information in sign languages [53]. In American Sign Language (ASL), for example, the eyebrows are used to indicate whether the sentence ends with a question mark, exclamation mark, or period. When asking a W-question (*e.g.* questions that have WHO, WHAT, WHERE, WHEN, WHY, and

WHICH) the eyebrows move downward to indicate curiosity or inquiry, while when asking YES or NO questions, the eyebrows are raised. Another purpose of facial expressions in ASL grammar is to express the emotion of the sign. For example, when signing HAPPY, SAD, or MAD, the facial expression must match the sign. If the sign HAPPY is signed with a sad face, the sign will be grammatically incorrect. Facial expressions are also used to add emphasis to a sign. In English, if a person wants to express the importance of a point, it is common to add the word "very" before it, for example, to provide emphasis. To show emphasis in ASL, facial expressions are added instead of an additional sign. Furthermore, ASL also has a tone to the signs. For example, we can use the sign "FINE" and create different meanings, *e.g.* FINE (happy), FINE (annoyed), or FINE (angry). These tones are created by adding a happy, annoyed, or angry face, in addition to emphasizing the sign FINE [54].

Mouth morphemes (mouthing). Mouth movement or mouthing is used to convey an adjective, adverb, or another descriptive meaning in association with an ASL word. Some ASL signs have a permanent mouth morpheme as part of their production. For example, the ASL sign NOT-YET requires a mouth morpheme (TH) whereas LATE has no mouth morpheme. These two are the same sign but with a different non-manual signal. These mouth morphemes are used in some contexts with some ASL signs, not all of them.

Simultaneity. Both manual and non-manual markers usually appear simultaneously within the signs but can also be used at a sentence level to express a grammatical change. For example, in American Sign Language, a positive sentence can be turned into a negative sentence by adding a headshake while signing or a signer can use eyebrow movements to transform a statement into a question.

Signing space. Sign language users utilize not just their body to communicate but also the space around them[2]. The signer's use of space is an important aspect of signed languages used not just to articulate signs but also to represent different mapping functions and interactions [55, 56]. For example, in ASL, the signing space can be used to: convey the noun-verb relationships; identify nouns and pronouns; refer to individuals, objects, buildings, and places; illustrate time sequencing; reflect spatial relationships; express distance between locations; compare and contrast; or to indicate the location of objects.

The most common way of using the signing space is by "placing" people, objects, etc. in a conversation. One example of placement can be seen as the following: the signer introduces who they will be talking about by signing their names and placing them

[2] The signing space extends from above the signer's head to the waist vertically and from elbow to elbow horizontally. Signs are articulated on or in front of the signer's body.

somewhere in the signing space. If we want, for example, to describe characteristics of Peter and Julia, we will sign “P-E-T-E-R” and place it on our right, then sign “J-U-L-I-A” and place it on our left. During the whole conversation, we will then refer back to Peter and Julia as “he” or “she” by just pointing to, looking at, or signing towards where each of them were placed instead of signing their names again. The placement can also be used to demonstrate a verb. For example, if we want to sign that Peter gave something to Julia, this action can be expressed by moving the sign “to give” from right to the left, or the opposite if we want to express that Julia gave something to Peter.

Classifiers are important yet complex elements of sign languages. They are designated handshapes that are associated with specific categories (classes) of things, size, shape, or usage. They can help to clarify the message, highlight specific details, and provide an efficient way of conveying information [57]. However, it is important to note here that the same handshape can be used to represent different objects or situations; this means that disambiguation can only come via *context*. For example, the handshape classifier “1” (CL:1) can be used to represent a person walking or a thin object such as a knife, pen, stick. Yet, it can also be used to explain how skinny an object is. Another commonly used classifier is the classifier “3” (CL:3). It is typically used to represent vehicles (*e.g.* cars, trucks, motorcycles, etc). For example, if signing *CAR* one can use the CL:3 classifier to explain how the vehicle was moving (direction, speed, etc) [3].

Depiction. The use of classifiers are usually related to the use of depiction, where the person uses their body to depict an action (*e.g.* showing how one would fillet a fish), a dialogue, or psychological events [35]. When a signer is depicting something, subtle shifts in body positioning and eye gaze can be used to indicate a referent, as explained above in the signing use of space.

Fingerspelling is often used to convey concepts from a spoken language that do not have a corresponding sign, or if the person does not know the sign for it. It can also be used to introduce a person who has not yet been assigned a name sign[4]. It is based on the alphabet of a spoken language, where every letter in the alphabet has a corresponding (static or dynamic) sign. Fingerspelling is also not shared between sign languages and can differ the way it is presented. For example, in ASL, the sign letters are performed using one hand, while in BSL two hands are used.

The **vocabulary and syntax** of sign languages also differ from the spoken languages of the same geographical area. They are independent of the spoken languages around them

[3] More information about handshapes and classifiers can be found at https://www.lifeprint.com/asl101/pages-signs/classifiers/classifiers-main.htm

[4] People within the Deaf community will often assign a unique and personal "name sign" as a way to identify someone without fully spelling out their name. These names often reflect the person’s character and are usually devised by someone within the Deaf community.

and keep evolving in the Deaf [5] communities where new signs emerge within the user's interaction. Signs from other sign languages can also be borrowed, similar to loanwords in spoken languages. In this case, the sign is still part of an established lexicon. However, SL users can also create an ad hoc sign (productive lexicon) to describe or convey a piece of a specific information. For example, if one wants to sign "a man walking on long legs" in Flemish Sign Language, instead of signing "MAN", "WALK", "LONG" and "LEGS", the hands can be used as classifiers to imitate the man walking [58]. Both the established and productive lexica are considered parts of the languages.

2.1.2 American Sign Language

While sign languages share some important characteristics, they are **not universal**. Similar to spoken languages, different countries and even different regions have developed their own sign language with distinct grammar and lexicons that brings with them the culture and identity of a community.

Although all Deaf communities and their sign languages are important and should have appropriated accessibility and inclusion in our society, for logistics reasons, in this work, we mainly focus on American Sign Language. ASL is the predominant language of the Deaf communities in the United States and most of the English-speaking parts of Canada. Dialects of ASL and ASL-based sign languages are also used in many countries around the world, including a big part of West Africa and parts of Southeast Asia. Despite its wide use, no accurate count of ASL users has been taken since hardly signed languages are included in language census [59].

ASL was originated in the early 19th century when Thomas Gallaudet and Laurent Clerc together opened the first American school for the deaf in Hartford (Connecticut, USA) after having traveled to Paris to study the French method of teaching [6] [60, 61]. Thus, unlike the spoken languages, American Sign Language has more in common with French Sign Language (LSF) than with British Sign Language (BSL) [60]. Although sign language had been used by various communities in the United States, ASL was just formalized as a language in 1960 by William C. Stokoe. His contributions revolutionized the understanding of the language in the United States and sign languages throughout the world [34].

Even though most Deaf communities of a country adopt a sign language to be used, this can vary depending on the geographic region and culture. For example, in Spain,

[5] We follow the recognized convention of using the upper-cased word Deaf which refers to the culture and describes members of the community of sign language users and the lower-cased word deaf describes the hearing status [11].

[6] The first school for the deaf was established in Paris during the 18th century.

there exists Spanish Sign Language, which differs from Catalan Sign Language (used in Catalonia), and Galician Sign Language (used in Galicia), similarly as their spoken counterparts. In addition to that, sign languages can also have dialects that vary according to region, social, and ethnic groups. Very often signers in different geographic regions or from different social groups show systematic differences in the way they use the language. For example, ASL users from the Northeast and Southern regions of the United States often use different signs for the same object, and deaf African Americans use the Black American Sign Language (BASL) dialect within the Black Deaf community [62].

2.1.3 Deaf Culture and the Marginalization of Signed Languages

A culture has different components such as language, values, traditions, norms and identity [63]. Sign language users form cultural minorities, that are connected by all these five sociological aspects. Deafness is not seen as a disability, but as a cultural identity [64] with many advantages [65]. Sign languages are a central component of Deaf cultures, and consequently, the development of systems that account sign language is highly sensitive, and must do the language justice to gain adoption [14].

Although signed languages are complete and natural languages used as the first or preferred mode of communication by millions of people worldwide [10], they unfortunately continue to be marginalized languages. It is worth noting that the marginalization of the Deaf community extends way beyond language and is highly similar to the long history of oppression experienced by minority groups in our society [66]. The suppression of sign language communication has been a major form of oppression against the Deaf community known as "audism" [67, 68]. In 1880, an international congress of a majority of hearing educators declared that sign language should not be used in the education process of deaf children, and spoken language should be used instead [69].

Consequently, oralism was widely enforced by training students to lip-read and speak. Since then, Deaf communities have fought to use sign languages in schools, work, and public life (*e.g.* [70]). Although this historical struggle can make development of sign language technology particularly sensitive in the Deaf community, we believe and reinforce that providing easy access to sign languages, including in education, public services and access to information is a critical current missing point for the human rights of deaf people [71].

2.2 Sign Language datasets

There are a number of sign language datasets publicly available that can be used for computer vision and natural language processing tasks, referred as *sign language understanding* tasks throughout this thesis. However, such datasets have been mostly collected for linguistic purposes, disconsidering most of the features that modern deep learning models require.

Benchmarks of *isolated signs* and *continuous signing* have been proposed for American (ASL) [6, 72–76], British (BSL) [77–79], Chinese (CSL) [80, 81], Finnish (FSL) [82], Flemish (VGT) [83], German (DGS) [19, 84, 85], Greek (GSL) [86], Indian (ISL) [87], Irish [88], Kazakh-Russian (K-RSL) [89], Swiss-German (DSGS) [83], Swedish (SSL) [90], Turkish (TİD) [91] Sign Languages.

Isolated signs datasets normally contains videos of people performing single signs (lexicon) together with its corresponding spoken word[7]. They are usually collected and provided by linguists and researchers from different areas as a dictionary database or in smaller sizes as a collection of specific signs [6, 72–74, 86, 87, 89–91]. The signs in these datasets are usually performed at a slower speed (for clarity), and the hands of the signer regularly starts and ends from a neutral pose, also called resting position. Although such data may be important as sign dictionary, or as a resource for those who are learning a sign language, most real-world use cases of sign language involve natural conversational with complete sentences (*i.e.* continuous sign language).

Continuous signing datasets can contain different types of data (also called modalities in this thesis) together with videos of people signing in a continuous way. Different from isolated signs, continuous signing presents two or more signs performed one after the other, with a context, usually forming a sentence. This represents a more natural way one would use the language in a daily live, *e.g.* in a conversation, in common activities, or also translating from spoken languages to sign languages.

Despite the large effort towards having large-scale sign language datasets in recent years[8], the lack of suitable datasets is still one of the biggest challenges in the area of computer vision applied to sign language research [14]. It is important to note that the collection and annotation of such data is a *laborious* and *expensive* task. It requires a team of sign language linguistics and native speakers (*e.g.* a Deaf person) working together to acquire reliable data. Sign language videos can sometimes be harvested from online sources (*e.g.*

[7] Depending on the purpose of the dataset, the corresponding spoken language word may not be available.

[8] Datasets released after the publication of How2Sign are marked with a * in Table 2.1

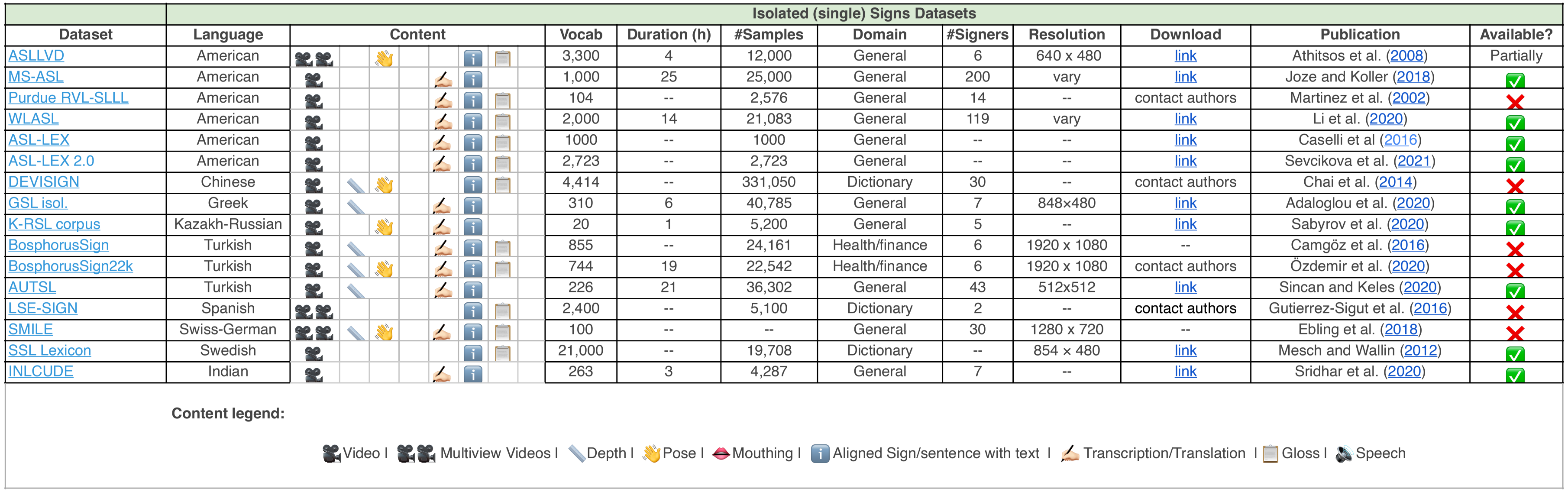
Table 2.1: **Summary of isolated and continuous sign language datasets publicly available**. At the time of its publication, How2Sign was the largest publicly available Sign Language dataset across languages in terms of vocabulary, and still the largest American Sign Language (ASL) dataset in terms of video duration. We also see that How2Sign is the dataset with the most parallel modalities. A detailed explanation of each modality can be found in the Section 3.2. *denotes datasets that were published after the How2Sign. *Mult.* stands for Multiview and it is consider when the recordings contain 2+ cameras. *Align.* stands for Alignment and it is consider when there is an automatic or manual alignment between the sign videos and the text translation (*Trans.*).

Name	**Lang.**	**Vocab.**	**Duration (h)**	**Signers**	**Modalities**						
					Mult.	**Trans.**	**Gloss**	**Pose**	**Depth**	**Alig.**	**Speech**
					Isolated signing						
K-RSL corpus [89]*	K-RSL	200	1	5	✗	✓	✗	✓	✗	✓	✗
AUTSL [91]*	TİD	226	21	43	✗	✓	✗	✗	✓	✓	✗
INCLUDE [87]*	ISL	263	3	7	✗	✓	✗	✗	✗	✓	✗
GSL isol. [86]*	GSL	310	6	7	✗	✓	✗	✗	✓	✓	✗
MSASL [6]	ASL	1,000	25	200	✗	✓	✗	✗	✗	✓	✗
WLASL [72]*	ASL	2,000	14	119	✗	✓	✓	✗	✗	✓	✗
ASL-LEX 2.0 [73]*	ASL	2,723	–	–	✗	✓	✓	✗	✗	✓	✗
ASLLVD [74]	ASL	3,300	4	6	✓	✗	✓	✓	✗	✓	✗
SSL Lexicon [90]	SSL	19,708	–	–	✗	✗	✓	✗	✗	✓	✗
					Continuous signing						
Public DGS Corpus [19]	DGS	–	50	327	✓	✓	✓	✓	✗	✓	✗
Video-Based CSL [80]	CSL	178	100	50	✗	✓	✗	✓	✓	✓	✗
GSL SD [86]*	GSL	310	10	7	✗	✓	✓	✗	✓	✓	✗
SIGNUM [84]	DGS	450	55	25	✗	✓	✓	✗	✗	✓	✗
S-pot [82]	FSL	1,211	9	5	✗	✓	✗	✗	✗	✗	✗
SWISSTXT-WEATHER [83]*	DSGS	1,248	1	–	✗	✓	✗	✓	✗	✓	✗
CSL-Daily [81]*	CSL	2,000	23	10	✗	✓	✗	✗	✗	✗	✗
BOBSL [78, 79]*	BSL	2,281	1,467	39	✗	✓	✗	✓	✗	✗	✗
RWTH-Phoenix-2014T [85, 92]	DGS	2,887	11	9	✗	✓	✓	✗	✗	✓	✗
BSL Corpus [77]	BSL	5,000	–	249	✗	✓	✓	✗	✗	✗	✗
VRT-NEWS [83]*	VTG	6,875	9	–	✗	✓	✗	✓	✗	✓	✗
SWISSTXT-NEWS [83]*	DSGS	10,561	9	–	✗	✓	✗	✓	✗	✓	✗
Boston104 [75]	ASL	104	8.7 (min)	3	✗	✓	✓	✗	✗	✗	✗
NCSLGR [76]	ASL	1,866	5.3	4	✓	✓	✓	✗	✗	✗	✗
How2Sign (ours)	ASL	15,686	79	11	✓	✓	✓	✓	✓	✓	✓

video platforms or SL learning resources), but still need to be annotated by sign language experts.

An extended overview including non public isolated and continuous sign language datasets are presented in Figures 2.1 and 2.2 respectively. We present detailed information about the signed language, the content or modalities included in the dataset, the vocabulary size, the number of samples, the domain of discourse, the number of signers, the resolution of the videos or images, the link for download (if available), the related publication and if the dataset is publicly available or not. An online version of both lists can also be found at https://how2sign.github.io/related_datasets.html. We intend to keep both lists updated in order to assist future work on this research area.

Table 2.1 presents summary of the dataset overview containing just publicly available datasets grouped by isolated and continuous signing.

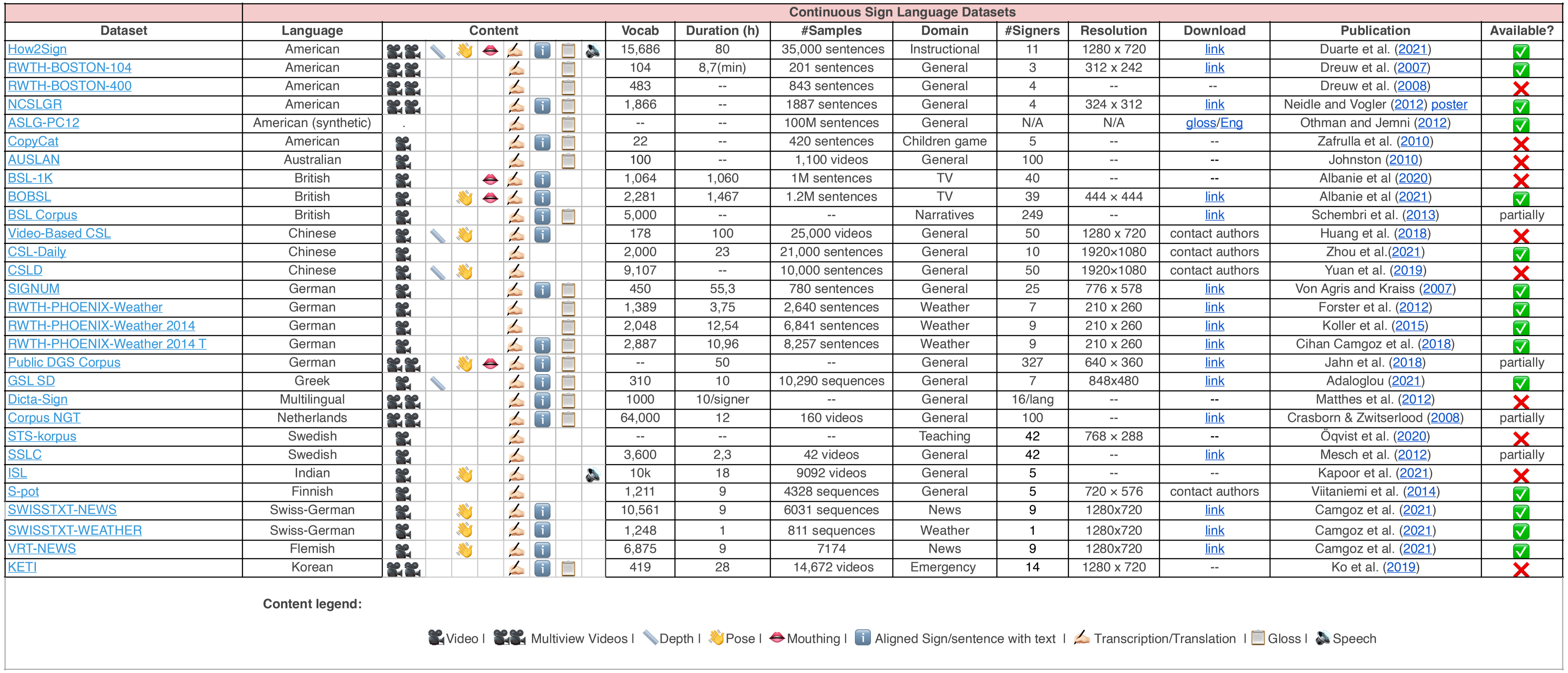


Dataset	Isolated (single) Signs Datasets									
	Language	Content	Vocab	Duration (h)	#Samples	Domain	#Signers	Resolution	Download	Publication	Available?
ASLLVD	American	🎥🎥 👋 ℹ 📋	3,300	4	12,000	General	6	640 x 480	link	Athitsos et al. (2008)	Partially
MS-ASL	American	🎥 ✍ ℹ	1,000	25	25,000	General	200	vary	link	Joze and Koller (2018)	✓
Purdue RVL-SLLL	American	🎥 ✍ ℹ 📋	104	--	2,576	General	14	--	contact authors	Martinez et al. (2002)	✗
WLASL	American	🎥 ✍ ℹ 📋	2,000	14	21,083	General	119	vary	link	Li et al. (2020)	✓
ASL-LEX	American	🎥 ✍ ℹ 📋	1000	--	1000	General	--	--	link	Caselli et al (2016)	✓
ASL-LEX 2.0	American	🎥 ✍ ℹ 📋	2,723	--	2,723	General	--	--	link	Sevcikova et al. (2021)	✓
DEVISIGN	Chinese	🎥 ✏ 👋 ℹ 📋	4,414	--	331,050	Dictionary	30	--	contact authors	Chai et al. (2014)	✗
GSL isol.	Greek	🎥 ✏ ✍ ℹ	310	6	40,785	General	7	848×480	link	Adaloglou et al. (2020)	✓
K-RSL corpus	Kazakh-Russian	🎥 👋 ✍ ℹ	20	1	5,200	General	5	--	link	Sabyrov et al. (2020)	✓
BosphorusSign	Turkish	🎥 ✏ ✍ ℹ 📋	855	--	24,161	Health/finance	6	1920 x 1080	--	Camgöz et al. (2016)	✗
BosphorusSign22k	Turkish	🎥 ✏ 👋 ✍ ℹ 📋	744	19	22,542	Health/finance	6	1920 x 1080	contact authors	Özdemir et al. (2020)	✗
AUTSL	Turkish	🎥 ✏ ✍ ℹ	226	21	36,302	General	43	512x512	link	Sincan and Keles (2020)	✓
LSE-SIGN	Spanish	🎥🎥 ℹ 📋	2,400	--	5,100	Dictionary	2	--	contact authors	Gutierrez-Sigut et al. (2016)	✗
SMILE	Swiss-German	🎥🎥 ✏ 👋 ✍ ℹ 📋	100	--	--	General	30	1280 x 720	--	Ebling et al. (2018)	✗
SSL Lexicon	Swedish	🎥 ℹ 📋	21,000	--	19,708	Dictionary	--	854 x 480	link	Mesch and Wallin (2012)	✓
INLCUDE	Indian	🎥 ✍ ℹ	263	3	4,287	General	7	--	link	Sridhar et al. (2020)	✓

Content legend:

🎥 Video | 🎥🎥 Multiview Videos | ✏ Depth | 👋 Pose | 👄 Mouthing | ℹ Aligned Sign/sentence with text | ✍ Transcription/Translation | 📋 Gloss | 🔊 Speech

Figure 2.1: **Isolated sign language datasets.** We present an extended overview and detailed information of the public and non-public isolated sign language datasets up to date.



	Continuous Sign Language Datasets																	
Dataset	Language	Content								Vocab	Duration (h)	#Samples	Domain	#Signers	Resolution	Download	Publication	Available?
How2Sign	American	🎥🎥	📏	👋	👄	✍	ℹ	📋	🔊	15,686	80	35,000 sentences	Instructional	11	1280 x 720	link	Duarte et al. (2021)	✅
RWTH-BOSTON-104	American	🎥🎥				✍		📋		104	8,7(min)	201 sentences	General	3	312 x 242	link	Dreuw et al. (2007)	✅
RWTH-BOSTON-400	American	🎥🎥				✍		📋		483	--	843 sentences	General	4	--	--	Dreuw et al. (2008)	❌
NCSLGR	American	🎥🎥				✍	ℹ	📋		1,866	--	1887 sentences	General	4	324 x 312	link	Neidle and Vogler (2012) poster	✅
ASLG-PC12	American (synthetic)	.				✍		📋		--	--	100M sentences	General	N/A	N/A	gloss/Eng	Othman and Jemni (2012)	✅
CopyCat	American	🎥				✍	ℹ	📋		22	--	420 sentences	Children game	5	--	--	Zafrulla et al. (2010)	❌
AUSLAN	Australian	🎥				✍		📋		100	--	1,100 videos	General	100	--	--	Johnston (2010)	❌
BSL-1K	British	🎥			👄	✍	ℹ			1,064	1,060	1M sentences	TV	40	--	--	Albanie et al (2020)	❌
BOBSL	British	🎥		👋	👄	✍	ℹ			2,281	1,467	1.2M sentences	TV	39	444 x 444	link	Albanie et al (2021)	✅
BSL Corpus	British	🎥				✍	ℹ	📋		5,000	--	--	Narratives	249	--	link	Schembri et al. (2013)	partially
Video-Based CSL	Chinese	🎥	📏	👋		✍	ℹ			178	100	25,000 videos	General	50	1280 x 720	contact authors	Huang et al. (2018)	❌
CSL-Daily	Chinese	🎥				✍				2,000	23	21,000 sentences	General	10	1920×1080	contact authors	Zhou et al.(2021)	✅
CSLD	Chinese	🎥	📏	👋		✍				9,107	--	10,000 sentences	General	50	1920×1080	contact authors	Yuan et al. (2019)	❌
SIGNUM	German	🎥				✍	ℹ	📋		450	55,3	780 sentences	General	25	776 x 578	link	Von Agris and Kraiss (2007)	✅
RWTH-PHOENIX-Weather	German	🎥				✍		📋		1,389	3,75	2,640 sentences	Weather	7	210 x 260	link	Forster et al. (2012)	✅
RWTH-PHOENIX-Weather 2014	German	🎥				✍		📋		2,048	12,54	6,841 sentences	Weather	9	210 x 260	link	Koller et al. (2015)	✅
RWTH-PHOENIX-Weather 2014 T	German	🎥				✍	ℹ	📋		2,887	10,96	8,257 sentences	Weather	9	210 x 260	link	Cihan Camgoz et al. (2018)	✅
Public DGS Corpus	German	🎥🎥		👋	👄	✍	ℹ	📋		--	50	--	General	327	640 x 360	link	Jahn et al. (2018)	partially
GSL SD	Greek	🎥	📏			✍	ℹ	📋		310	10	10,290 sequences	General	7	848x480	link	Adaloglou (2021)	✅
Dicta-Sign	Multilingual	🎥🎥				✍	ℹ	📋		1000	10/signer	--	General	16/lang	--	--	Matthes et al. (2012)	❌
Corpus NGT	Netherlands	🎥🎥				✍	ℹ	📋		64,000	12	160 videos	General	100	--	link	Crasborn & Zwitserlood (2008)	partially
STS-korpus	Swedish	🎥				✍				--	--	--	Teaching	42	768 x 288	--	Öqvist et al. (2020)	❌
SSLC	Swedish	🎥				✍				3,600	2,3	42 videos	General	42	--	link	Mesch et al. (2012)	partially
ISL	Indian	🎥		👋		✍			🔊	10k	18	9092 videos	General	5	--	--	Kapoor et al. (2021)	❌
S-pot	Finnish	🎥				✍				1,211	9	4328 sequences	General	5	720 x 576	contact authors	Viitaniemi et al. (2014)	✅
SWISSTXT-NEWS	Swiss-German	🎥		👋		✍	ℹ			10,561	9	6031 sequences	News	9	1280x720	link	Camgoz et al. (2021)	✅
SWISSTXT-WEATHER	Swiss-German	🎥		👋		✍	ℹ			1,248	1	811 sequences	Weather	1	1280x720	link	Camgoz et al. (2021)	✅
VRT-NEWS	Flemish	🎥		👋		✍	ℹ			6,875	9	7174	News	9	1280x720	link	Camgoz et al. (2021)	✅
KETI	Korean	🎥🎥				✍	ℹ	📋		419	28	14,672 videos	Emergency	14	1280 x 720	--	Ko et al. (2019)	❌

Content legend:

🎥 Video I 🎥🎥 Multiview Videos I 📏 Depth I 👋 Pose I 👄 Mouthing I ℹ Aligned Sign/sentence with text I ✍ Transcription/Translation I 📋 Gloss I 🔊 Speech

Figure 2.2: **Continuous sign language datasets.** We present an extended overview and detailed information of the public and non-public continuous sign language datasets up to date.

Modalities As mentioned before, sign language datasets can contain different types of data [9]. Regarding the different modalities available in the public datasets, we can observe that: (i) the majority of the datasets does not include *multiview* recordings (*e.g.* more than one camera at different view points), which can be helpful to disambiguate signs and have a better estimation of the body position; (ii) all continuous signing datasets include the *translation* in one or more of its spoken language counterpart; (iii) in addition to the translation, some of the datasets also provide *gloss* annotations, that is, a text-based transcription of the signs that can serve as a proxy in translation tasks; (iv) a few provide 2D or 3D pose estimation, usually extracted using out-of-the-box human pose estimation methods (*e.g.* OpenPose [23]); (v) only five datasets used a *Depth* sensor to record the signers in addition to the RGB camera; (vi) some of the datasets provide manually or automatic *alignment* between the sign language videos and the spoken language translation – datasets that do not provide such alignment usually maintain the alignment with the audio coming from the spoken language translation, when available; (vii) although other datasets have the translation coming from the spoken language audio, How2Sign is the only sign language dataset aligned with a *speech* track.

Limitations. The publicly available datasets present one or more of the following weaknesses: (i) they have a limited number of signers – for example, GSL SD [86] and S-pot [82], have seven or fewer signers; (ii) they have a limited vocabulary size – for example, Video-Based CSL [80], SIGNUM [84] and Boston104 [75] only have a few hundred signs or spoken language words; (iii) they are limited in total duration – for example the popular RWTH-Phoenix-2014 [85] dataset contains only 11 hours of content; (iv) they represent natural continuous signs but cover a limited domain of discourse or contain sign language interpretation[10] – for example, the videos in RWTH-Phoenix-2014 [85], SWISSTXT-WEATHER [83], SWISSTXT-NEWS [83], VRT-NEWS [83] and BOBSL [78] are from weather broadcasts or TV shows. If we focus on American Sign Language (ASL), BOSTON-104 [75] only contains 8.7 minutes of grayscale video, while NCSLGR [76] is larger, but still an order of magnitude smaller than How2Sign. We present details about the How2Sign dataset including its statistics, modalities, and information on the video recordings and annotations in the next Chapter.

[9] We summarize and describe the most common types of data available in continuous SL datasets in Section 3.2.

[10] Language interpretation is defined by the International Standards Organization (ISO) as the following: "Rendering a spoken or signed message into another spoken or signed language, preserving the register and meaning of the source language content." [93]

Part I

Collecting and Annotating Sign Language Data

3

The How2Sign dataset

In this Chapter we introduce the How2Sign dataset and present in detail the modalities it is composed of (Section 3.2), followed by the description of the data collection process. In particular, we describe in Section 3.3 how the sign language videos were recorded, and in Subsections 3.3.2 and 3.3.3 we detail the how the manual (English translation re-alignment, gloss annotations and videos categories) and automatic (2D/3D pose estimation) annotations where collected. We also include a discussion about the privacy, bias and ethical considerations that are important for a better understanding of the data (Section 3.4).

3.1 Introduction

Promising recent works in sign language understanding tasks have shown that modern computer vision and machine learning architectures may be able to help breaking down the communication barriers that sign language users face in their daily lives [85, 94–98] .

However, in order to successfully train such data-hungry models, large amounts of annotated data is needed. The availability of public large-scale datasets suitable for sign language understanding tasks is very limited, as presented in the previous Chapter, especially when it comes to *continuous sign language* datasets, *i.e.*, where the data is segmented and annotated at the sentence level. Before the publication of the dataset presented in this thesis, there was no ASL dataset large enough to be used with current deep learning approaches.

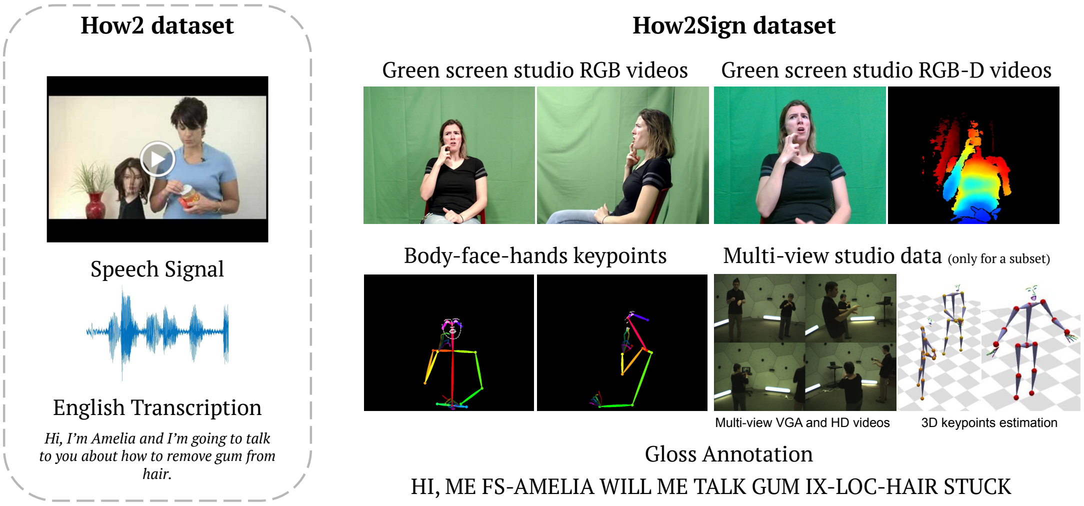

How2 dataset
Speech Signal
English Transcription
Hi, I'm Amelia and I'm going to talk to you about how to remove gum from hair.
How2Sign dataset
Green screen studio RGB videos
Green screen studio RGB-D videos
Body-face-hands keypoints
Multi-view studio data (only for a subset)
Multi-view VGA and HD videos
3D keypoints estimation
Gloss Annotation
HI, ME FS-AMELIA WILL ME TALK GUM IX-LOC-HAIR STUCK

Figure 3.1: **Sample data of the How2Sign dataset.** Our dataset consists of over 80 hours of multiview sign language videos and different modalities.

In order to tackle this problem, we collected and we present in this Chapter the *How2Sign* dataset. In the following sections we describe in details the process of collecting sign language video and annotations and provide insights and information that can facilitate future data collection efforts. Figure 3.1 shows samples of the data that compose the How2Sign. The dataset is publicly available for research purposed and can be downloaded at: http://how2sign.github.io/.

3.2 Dataset Modalities

As mentioned in Section 2.2, continuous sign language datasets usually include different types of data (called here modalities) together with the sign language videos. The How2Sign dataset is composed of six different modalities that were either borrowed from the source language dataset or were manually or automatically collected.

As a starting point, we selected a set of instructional videos from the existing *How2 dataset* [99] that were originally available in spoken English. The How2 is a publicly available multimodal dataset for vision, speech and natural language understanding, with utterance-level time alignments between the speech and the ground-truth English transcription. The original dataset consists of 79,114 instructional videos (2,000 hours in total, with an average length of 90 seconds) downloaded from Youtube with English transcription (subtitles) manually added by the users. A subset of 13,662 videos (300 hours) was manually verified as part of the first release and translated into Portuguese and called by the authors *How2-300h*. From the How2-300h subset, we selected 2,192 videos (60 hours) from the training set and the complete validation (115 videos) and test sets (149 videos) to be part of the How2Sign.

Using the selected videos, we record videos of ASL signers translating the original spoken English content into American Sign Language translations. In addition to the videos, we also collect a set of different modalities that are described next.

Multiview. All sign language videos were recorded from multiple angles. This allows the signs to be visible from multiple points of view, reducing occlusion and ambiguity, especially in the hands. Specifically, the sign language videos recorded in the Green Screen studio contain two different points of view (frontal and lateral), while the Panoptic studio recordings consist of recordings of more than 500 cameras allowing for a high quality estimation of 3D keypoints [100]. We detail the recordings in both studios in Section 3.3.

Transcriptions. The English speech transcriptions (or also called English translation) originates from the subtitles track of How2 original videos. The transcriptions were manually produced by the uploader of the instructional video in form of text, that was loosely synced with the video's speech track. As subtitles are not necessarily fully aligned with the speech, transcriptions were time-aligned at the sentence-level as part of the How2 dataset [99]. The transcriptions were later re-aligned with the sign language videos as part of the How2Sign dataset. We explain this process in Subsection 3.3.2.

Gloss. As mentioned in Section 1.3, gloss is used in linguistics to transcribe signs using spoken language words. Although gloss is not a true translation, it is the form of text that is closest to sign language and it has been used by a number of approaches as an intermediate representation for sign language understanding [85, 94, 97, 98, 101]. An example of gloss annotation is shown on the bottom right of Figure 3.1.

Pose. Human pose information, *e.g.* body, hand and face keypoints were extracted for all the recorded sign language videos in the full resolution – 1280 x 720 pixels. For the Green Screen studio data, the two-dimensional (2D) pose information was automatically extracted using OpenPose [23]. In total, each pose consists of 25 body keypoints, 70 facial keypoints and 21 keypoints for each hand. We provide pose information for both frontal and side view of the Green Screen studio data. A sample of the pose information extracted can be seen on the bottom row in the left side of Figure 3.1. For the Panoptic studio data, we provide high quality 3-dimensional (3D) pose information estimated by the Panoptic studio internal software [100] that can be used as ground-truth for a number of 3D vision tasks.

Depth data. For the Green Screen studio data, the sign language videos were also recorded using a Depth sensor (Creative BlasterX Senz3D) from the frontal viewpoint.

The sensor has high precision facial and gesture recognition algorithms embedded and is able to focus on the hands and face, the most important human parts for sign language.

Speech. The speech track comes from the instructional videos as part of the How2 dataset [99]. They are audio files containing what the person in the video is speaking. As part of the How2 dataset, a set of features were extracted and distributed.

3.3 Data collection

In this section we detail the data collection process, *i.e.* the process we followed for collecting the sign language videos as well as for the manual and automatic annotations. We later present the data statistics of the dataset.

3.3.1 Sign language video recordings

All recordings were performed in a supervised setting at two different locations: the *Green Screen studio* and the *Panoptic studio*, both described below. We recorded the complete dataset (80 hours) in the green screen studio. Note that the 60 hours of videos selected from the How2 dataset translated to 80 hours of videos in the How2Sign due the different speed in which both English and ASL are conveyed, as well the followed recording pipeline (we explain this process below). We then choose a small subset of videos (approx. 3 hours) from the validation and test splits and recorded them again in the Panoptic studio. After recording, we trimmed the RGB frontal and lateral sign language videos and divided them in *sentence-level clips*, each annotated with a corresponding English transcript, and the modalities presented in Section 3.2.

The **Green Screen studio** was equipped with a depth and a high definition (HD) camera placed in a frontal view of the signer, and another HD camera placed at a lateral view. All three cameras recorded videos at 1280x720 resolution, at 30 fps. Figure 3.2 shows the studio and its setup. Samples of data recorded in this studio are shown in the top row of Figure 3.1.

The **Panoptic studio** [100] is a singular system equipped with 480 VGA cameras, 30 HD cameras and 10 RGB-D sensors, all synchronized. All cameras are mounted over the surface of a geodesic dome[1], providing redundancy for weak perceptual processes (such as pose detection) and robustness to occlusion. In addition to the multiview VGA and HD videos, the recording system can further estimate high quality 3D keypoints of the interpreters, also included in How2Sign. Figure 3.3 shows the inside of the dome and

[1] http://www.cs.cmu.edu/~hanbyulj/panoptic-studio/

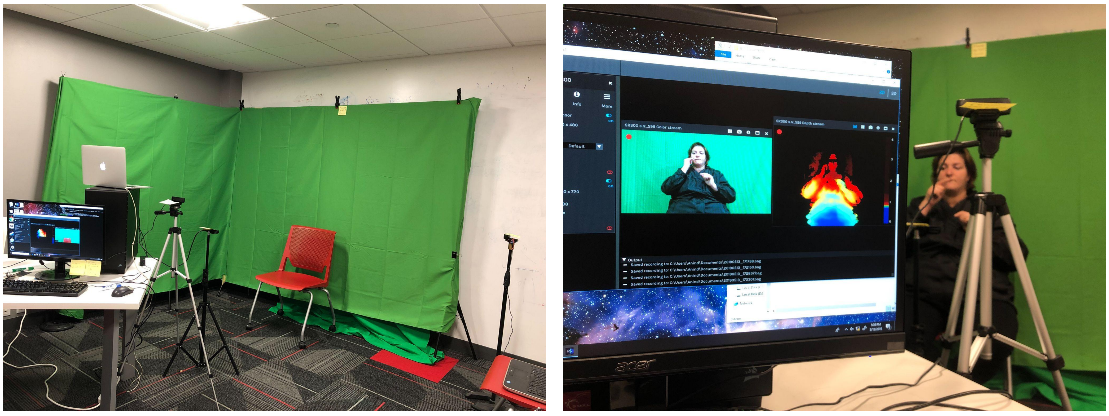


Figure 3.2: **Green Screen studio.** On the left we can see the studio setup with the green background, the cameras on the tripods in front and on the side of the chair where the signers would remain sitting during the recordings as well as the screen (grey laptop) where they would watch the video that would be translated. On the right we can see a signer during a recording session from the point of view of the person recording it. The image show the output of the Depth camera.

its setup. Samples of data recorded in this studio are shown on the bottom-right of Figure 3.1.

Recording pipeline. Before recording the videos, the signers were asked to position themselves in front of a screen where the videos were presented to them. In the Green Screen studio they remained sitting with a green screen background behind them, while in the Panoptic studio they stayed standing in the middle of the dome. Before starting the video recordings the signer would first watch the video with the transcript as subtitles in order to become familiar with the overall content; this enables them to perform a richer translation. After becoming familiar with the content, they were asked to watch the video again but now translating the English sentences from the subtitles into ASL while being recorded. During the recordings the signers were watching the original video at a slightly slower-than-normal (0.75) speed to facilitate the translation.

Importance of providing the original video to the signer before the recordings. During the preliminary design phase of the data collection, signers were asked to perform English to ASL translation when given: (1) just text without reading it beforehand; (2) the video and text together but without seeing it previously and (3) text and video together and allowing them to watch it before the recording. The conclusions for each case were: (1) signers found it hard to understand and follow the lines at the same time, causing lots of pauses and confusion; (2) signers found it easier to understand and translate but still with some pauses and (3) the understanding and flow improved. With these conclusions we decide to proceed with (3) during the rest of the recordings.

Verification process. After the recordings, a set of around 750 videos (30%) were manually verified by ASL users that did not participate into the recordings. They were

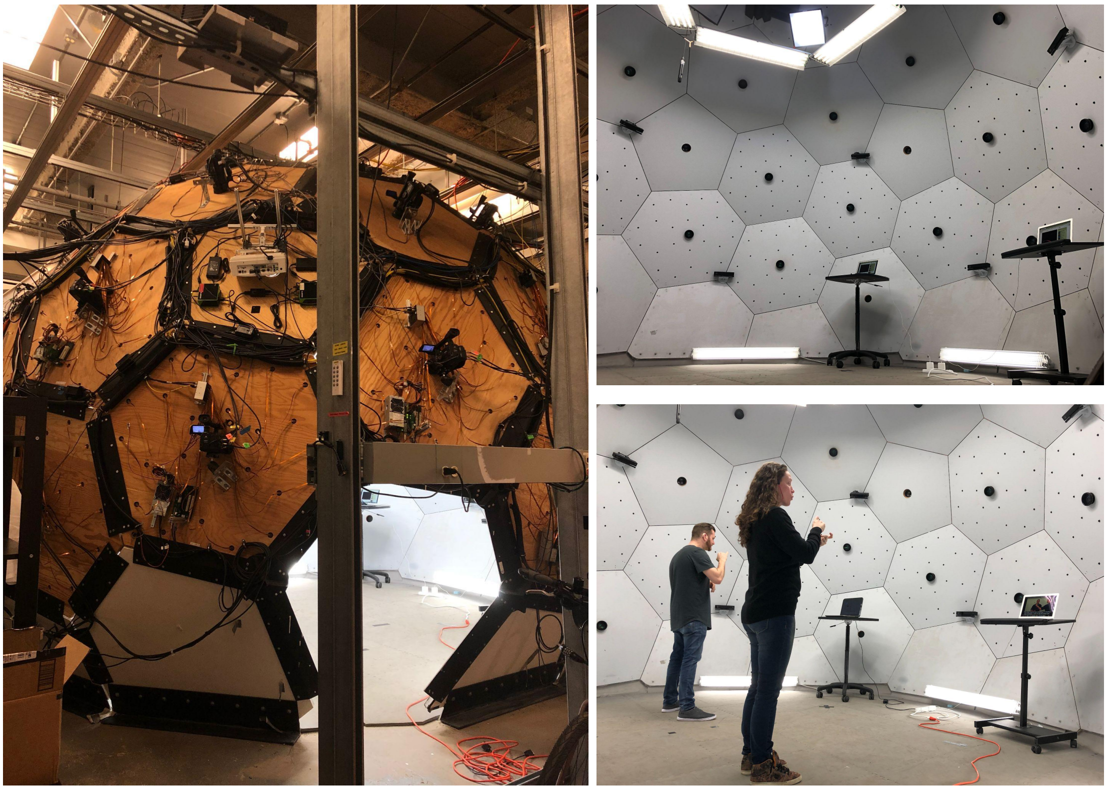


Figure 3.3: **Panoptic studio.** On the left we can see the external structure of geodesic dome where the 500+ cameras are mounted in different view points. The whole system is equipped with 480 VGA cameras, 30 HD cameras and 10 RGB-D sensors. On the top right of the image we can see the internal structure of the dome with the cameras and the illumination system. On the bottom right we can see two signers during a recording session.

asked to watch the original video with the subtitles that were presented to the signers as well as the ASL translation and asked if the translation conveyed by the signer was correct or not. This verification process assure us that the signers translated the content correctly as requested.

Signers. In total, 11 people appear in the sign language videos of the How2Sign dataset; we refer to them as *signers*. Of the 11 signers, 5 self-identified as hearing, 4 as Deaf and 2 as hard-of-hearing. The signers that were hearing were either professional ASL interpreters (4) or ASL fluent. Figure 3.4 show all the 11 signers that participated in the recordings of the How2Sign dataset. From the 11 signers, four of them (signers 1, 2, 3 and 10) participated in both the Green Screen studio and the Panoptic studio recordings. Signers 6 and 7 participated only in the Panoptic studio recordings, while signers 4, 5, 8, 9 and 11 participated only in the Green Screen recordings. The signer ID information of each video is made available with the dataset.

Duration and cost of data collection. The videos in the dataset were collected across 65 days over 6 months. For each hour of video recorded, the preparation, recording and

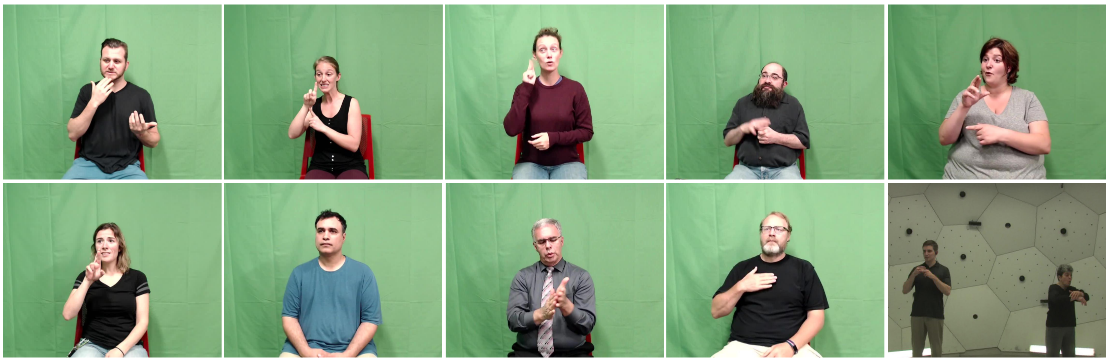


Figure 3.4: **The 11 signers that appear in the How2Sign dataset videos**. On the top row, we can see signers 1-5 (from left to right) in the Green Screen Studio, while on the bottom row we can see signers 8-11 (again left to right) in the Green Screen Studio. The rightmost figure on the bottom row shows signers 6-7 in the Panoptic studio.

video review took approximately 3 hours on average. All the videos were then manually verified and trimmed by professionals. This post-process step took approximately 4 months. After being verified and trimmed, the English transcriptions were realigned with the ASL videos (see Section 3.3.2 for more information) as well as part of the gloss annotation was collected. This process took approximately 8 months [2]. All participants including signers, video editing professionals and ASL linguists were compensated accordingly to their tasks rounding the total cost of the data collection to $35,000 US dollars [3].

3.3.2 Manual Annotations

Beyond the video recordings we further collected a number of manual annotations for the sign language videos. Here we present the re-aligment of the English transcriptions, followed by the gloss annotations and the classification of the video categories. All manual annotations are available with the dataset.

Re-alignment of the English translation. The English translation originates from the subtitles in the How2 dataset (also called English transcriptions). The transcriptions were manually produced by the uploader of the instructional video in form of text, that was loosely synced with the video's speech track. As subtitles are not necessarily fully aligned with the speech, transcriptions were time-aligned at the sentence-level as part of the How2 dataset release. However, when translating the content from English to American Sign Language, the original alignment between the speech and the subtitles is not transferable to the ASL videos. English and ASL are different languages and do

[2] Part of the gloss annotations are still being collected.

[3] Video recordings, manually verification and post-processing step cost a total of $15,000 US dollars while the subtitle realignment and gloss annotation cost a total of $20,000 US dollars.

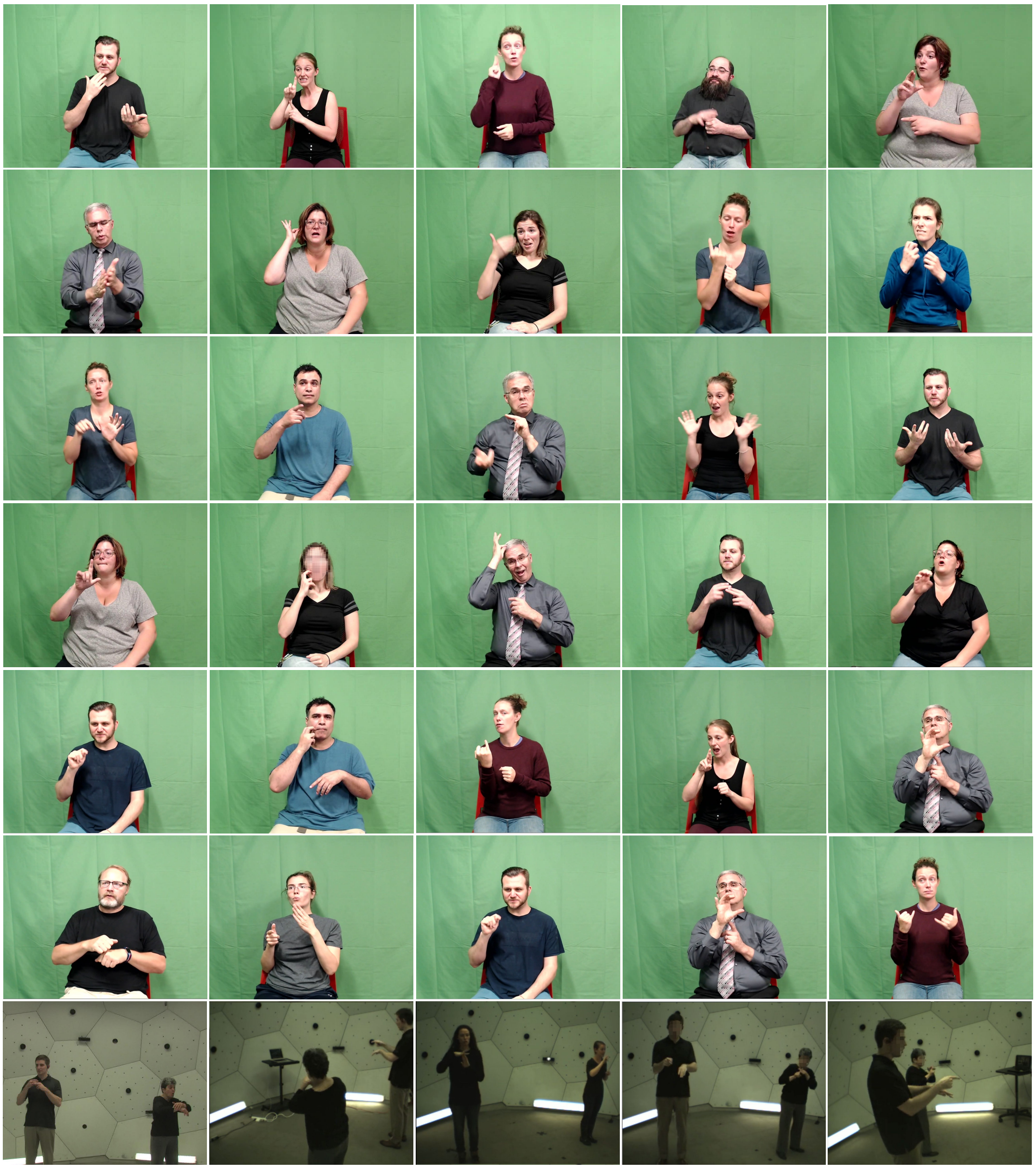


Figure 3.5: **Samples of the RGB videos from How2Sign**. the six first rows show samples of the videos recorded by the frontal view RGB camera in the Green Screen studio. Last row shows samples from the RGB cameras from the Panoptic studio. Given the overall structure of both studios, for each recording session, we recorded one person at the time on the Green Screen studio, and two people at the time in the Panoptic studio. At the Panoptic studio it is possible to easily separate and choose the data of each signer to be used by selecting the camera view (in case of the RGB videos) or the person ID (in the estimated 3D keypoints data).

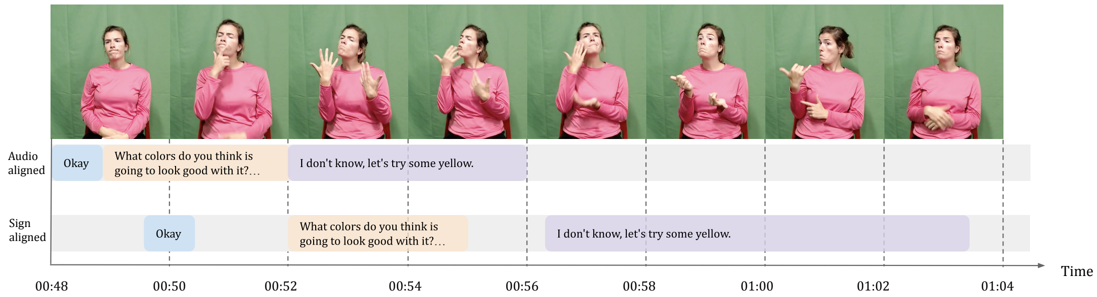

Audio aligned
Okay
What colors do you think is going to look good with it?...
I don't know, let's try some yellow.
Sign aligned
Okay
What colors do you think is going to look good with it?...
I don't know, let's try some yellow.
Time
00:48
00:50
00:52
00:54
00:56
00:58
01:00
01:02
01:04

Figure 3.6: **Re-alignment of the English translation.** We illustrate the differences in the temporal alignment between the ASL video sequences (top) and the English translation when the text is aligned with the audio from the original video from How2 (middle), and when it is re-aligned to match the signing video (bottom). In order to align the ASL videos and the English translation, we manually re-aligned all sentences in the How2Sign dataset. The timestamps of both alignments are provided with the dataset. Image adapted from [4].

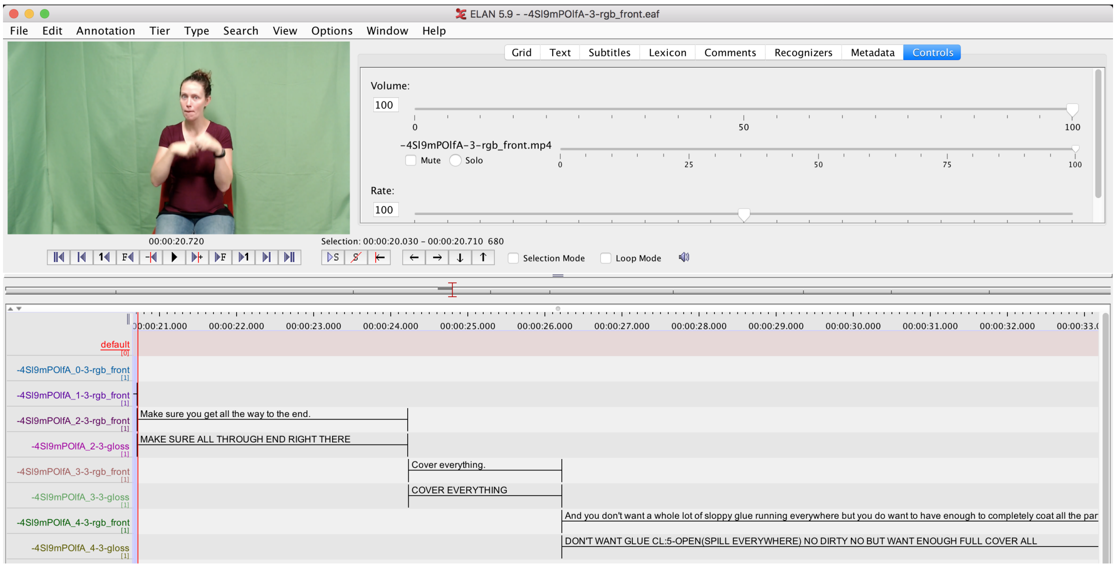
not have the same structure. In addition to that, during the ASL translation and sign language video recording process, any temporal alignment is naturally lost. We illustrate this problem in Figure 3.6.

In order to correct that, we asked ASL users to re-align the English subtitles with the ASL videos so that each sentence in ASL have a correspondent English translation. The re-alignment process was manually performed using ELAN [102], an annotation software for audio and video recordings, specifically enhanced for sign language annotations. Figure 3.7 shows a screenshot of the software. In ELAN, each sentence of the subtitle is allocated in a different line, called “tiers”, which is time-aligned to the video file. Figure 3.7 shows different tiers in the bottom part of the image. To adjust the time-alignment, the sentence boundaries can be moved from left to right (and vice-versa) so the start and end of each English text sentence coincides with the corresponding ASL sentence in the video.

It should be noted that *not all sentences* in the How2Sign have a corresponding English translation. This is due the fact that the subtitles were first automatically time-alignment with the speech in the How2 dataset and some sentences were removed in this process. More specifically sentences that were too short, that contained music components or that did not form a complete sentence.

Gloss annotations are used in linguistics to transcribe signs using spoken language words and sign language specific notations. We provide more details about gloss annotations in Section 3.2. Note that since gloss annotations are used specifically as a linguistic tool, sign language users are not necessarily familiar with them.

Thus, to annotate part of the sign language videos with gloss, it was necessary to employ ten sign language linguists. A total of 4 hours (around 128 videos) of the How2Sign was annotated with gloss. The annotation process was performed using ELAN, over



Selection: 00:00:20.030 – 00:00:20.710 680

default [0]	
-4SI9mPOIfA_0-3-rgb_front [1]	
-4SI9mPOIfA_1-3-rgb_front [1]	
-4SI9mPOIfA_2-3-rgb_front [1]	Make sure you get all the way to the end.
-4SI9mPOIfA_2-3-gloss [1]	MAKE SURE ALL THROUGH END RIGHT THERE
-4SI9mPOIfA_3-3-rgb_front [1]	Cover everything.
-4SI9mPOIfA_3-3-gloss [1]	COVER EVERYTHING
-4SI9mPOIfA_4-3-rgb_front [1]	And you don't want a whole lot of sloppy glue running everywhere but you do want to have enough to completely coat all the par
-4SI9mPOIfA_4-3-gloss [1]	DON'T WANT GLUE CL:5-OPEN(SPILL EVERYWHERE) NO DIRTY NO BUT WANT ENOUGH FULL COVER ALL

Figure 3.7: **Samples of gloss annotations.** We provide an example of the ELAN files used to collect the gloss annotation.

the re-aligned English translation. In order to prepare the data to be annotated, the ELAN files needed to be created one-by-one and prepared so the linguists would have the appropriate files to work on. This process took on average 5 minutes per video.

The gloss annotation was done sentence-by-sentence, where the annotator would first watch the ASL video, identify the start and end of the sentence, and transcribe each sign in that sentence into glosses. This process took in average one hour for every *90 seconds* of video, becoming an extreme difficult, time consuming, and expensive part of the annotation step that was initially not expected.

Below we describe the convention used to annotate the How2Sign with glosses. Most symbols are standards in ASL glossing, but some adaptations were proposed by the linguists who annotate our data to better adjust the data to our needs. A complete list is available on the dataset website [4].

- *Capital letters.* English glosses are written using capital letters. They represent an ASL word or sign. It is important to remember that gloss is not a translation. It is only an approximate representation of the ASL sign itself, not necessarily a meaning.
- A *hyphen* is used to represent a single sign when more than one English word is used in gloss (*e.g.* STARE-AT).

[4] https://how2sign.github.io/

- The *plus sign (+)* is used in ASL compound words (*e.g.* MOTHER+FATHER – used to transcribe parents). It is also used when someone combines two signs in one (*e.g.* YOU THERE will be glossed as YOU+THERE).
- The *plus sign (++)* at the end of a gloss indicates a number of repetitions of an ASL sign (*e.g.* AGAIN++ – the word “again" was signed two more times meaning “again and again").
- *FS:* represents a fingerspelled word (*e.g.* FS:AMELIA).
- *IX* is a shortcut for “index", which means to point to a certain location, object, or person.
- *LOC* is a shortcut for “locative", a part of the grammatical structure in ASL.
- *CL:* is a shortcut for “classifier". Classifiers are signs that use handshapes that are associated with specific categories (classes) of things, size, shape, or usage. They can help to clarify the message, highlight specific details, and provide an efficient way of conveying information[5]. In our annotations, classifiers will appear as: “CL:classifier(information)". For example, if the signer signs “TODAY BIKE" and uses a classifier to show the bike going up the hill, this would be glossed as: “TODAY BIKE CL:3 (going uphill)").

A number of unexpected events including budget, time consumed, and lack of annotators lead to an interruption of the gloss annotation collection. Here we give a perspective of how much time and budget would be needed to finish collecting the annotations for the whole How2Sign according to our estimations and previous data. As mentioned above, ELAN files need to be created and adapted to accommodate the gloss annotations. Considering the total amount of videos on the How2Sign, a total of 200 hours would be needed to finish the creation of such files. An automatic strategy could help accelerate this process. In addition to that, given the total amount of hours remaining to be annotated, a total of around 3000 hours would be needed to finish the gloss collection. For this the budget for each step would vary accordingly to the annotators' cost as well as the group of researchers and assistants conducting the data collection.

Video Categories. Although the How2 dataset provides automatically extracted “topics” for all videos using Latent Dirichlet Allocation [103], we found that the automatic annotations were in general very noisy and not properly characterizing the selected videos. In order to better categorize the videos, we manually selected 10 categories from the

[5] More info about handshapes and classifiers can be found at: https://www.lifeprint.com/asl101/pages-signs/classifiers/classifiers-main.htm

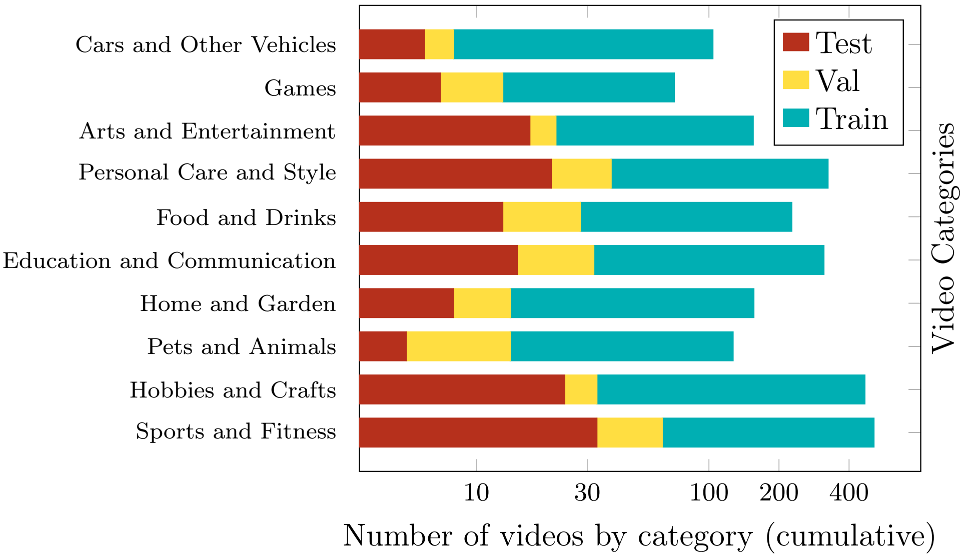


Figure 3.8: **Video categories.** Cumulative number of videos per category.

instructional website Wikihow [6] and manually classified each How2Sign video in a single category. The categories are: *Personal Care and Style, Games, Arts and Entertainment, Hobbies and Crafts, Cars and Other, Vehicles, Sports and Fitness, Education and Communication, Food and Drinks, Home and Garden and Pets and Animals.* The distribution of videos across the ten categories can be seen in Figure 3.8.

3.3.3 Automatic Annotations

Green Screen 2D Pose Estimation. We extract 2-dimensional (2D) pose information (*e.g.* body, hand, face keypoints) for the RGB videos collected in the Green Screen studio (frontal and lateral view points) using OpenPose [23]. OpenPose is a multiple-person detection library that associates human body parts with individuals in a image. In total, the full body pose estimation consists of 25 body keypoints, 70 facial keypoints and 21 keypoints for each hand. Figure 3.9 shows the kinematic tree of the predicted keypoints of the body, face and hand [7].

The 2D pose estimation was extracted for all the HD videos (frontal and lateal RGB videos) recorded in the Green Screen studio using the full resolution– 1280 x 720 pixels. A sample of the pose information extracted from the How2Sign videos can be seen in the bottom row in the left side of Figure 3.1.

Confidence of the extracted keypoints. We compare the confidence of OpenPose detections for two different video definitions in order to estimate the quality of the

[6] https://www.wikihow.com/Special:CategoryListing

[7] More information can be found at the OpenPose documentation: https://cmu-perceptual-computing-lab.github.io/openpose/web/html/doc/md_doc_02_output.html

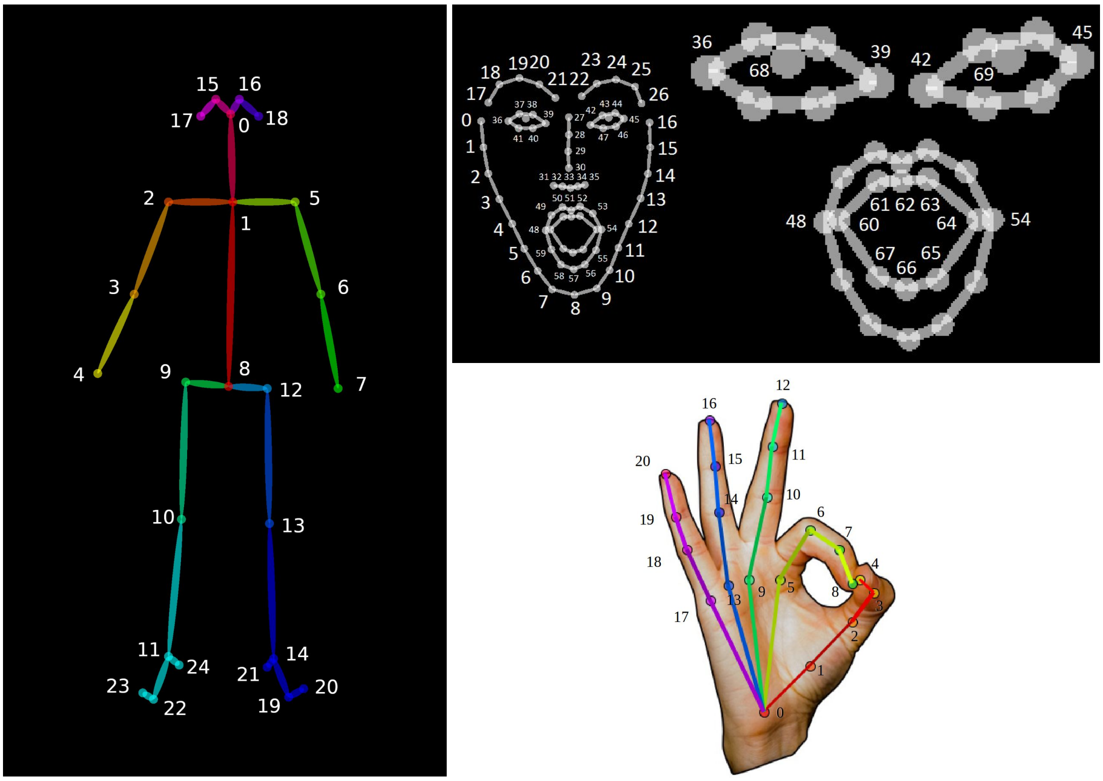


Figure 3.9: **Kinematic tree of OpenPose.** On the left image we can see the kinematic tree representation of the output of the COCO body model used to extract the keypoints of the How2Sign videos. The facial (top) and the hand (botton) kinematic tree representation are shown on the right.

Table 3.1: **OpenPose confidence scores.** Average of confidence score of OpenPose on high resolution (1280 x 720) compared with low resolution (210 x 260) videos of the How2Sign dataset.

	Body	Right hand	Left hand	Face	Total
High resolution	0.39	0.42	0.47	0.84	0.53
Low resolution	0.40	0.24	0.30	0.73	0.42

automatic extracted 2D poses. Extracting keypoints from 1280 x 720 videos provided a 53.4% average confidence versus a 42.4% confidence for a lower resolution (210 x 260). This difference is more prominent when different parts of the body are analyzed. Table 3.1 show the different average confidence scores when OpenPose is extracted using high and low resolution videos. We see that both hands cause the most harm when low resolution is used.

Panoptic 3D Pose Estimation. For the Panoptic studio data, we provide high quality 3-dimensional (3D) pose information estimated by the Panoptic studio internal software [100] which may be used as ground-truth for a number of 3D vision tasks.

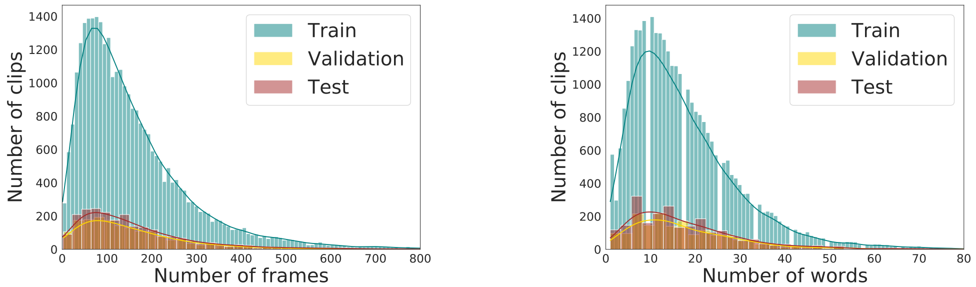


Figure 3.10: **Clip-level statistics.** Distribution of the number of frames (left) and words (right) over sentence-level clips. Each clip has on average 162 frames (5.4 seconds) and 17 words.

Sign Spotting. Using two different sign spotting techniques, we collect mouthing-based [79] and dictionary-based [104] sign spotting annotations for the How2Sign, enabling its use in sign language recognition and retrieval tasks. The automatic production of these pseudo-labels is a contribution of this thesis and it is detailed in Section 4.

3.3.4 Dataset statistics

In Table 3.2 we show detailed statistics of the How2Sign dataset. To the best of our knowledge, How2Sign is the largest publicly available sign language dataset across languages in terms of vocabulary, as well as an order of magnitude larger than any other ASL dataset in terms of video duration. We see that How2Sign is also the dataset with the most parallel modalities, enabling multimodal learning. The bottom section presents the statistics of the automatically extracted human pose annotations.

Number of videos. A total of 2,456 videos from the How2 [99] were used to record the sign language videos– 2,192 from the trainning set, 115 from the validation set and 149 from the test set. We maintain the videos in the same partion as they appear in the How2 dataset. Some of the videos were recorded more than once by a different signer in the Green screen studio – 21 videos from the training set, 17 videos from the validation set and 35 videos from the test set. Those videos can be identified by the video ID provided in the dataset. In total the How2Sign consists of 2,529 videos.

Number of sentences. All the HD-RGB recorded videos were split into sentence-level clips. Each clip has on average 162 frames (5.4 seconds) and 17 words. The distribution of frames (right) and words (left) over all the clips for the 3 splits of the dataset can be seen in Figure 3.10. The collected corpus covers *more than 35k sentences* with an English vocabulary of more than 16k words, where approximately 20% of it is finger spelled.

Number of signers. The videos were recorded by 11 different signers distributed across the splits. The test set contains 26 duplicated videos that were recorded by a

Table 3.2: **How2Sign statistics**. Some of the videos were recorded more than once by a different signer in the Green screen studio (see second row vs. first row). The RGB videos from frontal and side view were split into sentence-level *clips*.

	Green screen studio				**Panoptic studio**		
	train	**val**	**test**	**Total**	**val**	**test**	**Total**
How2 [99] videos	2,192	115	149	2,456	48	76	124
Sign language videos	2,213	132	184	2,529	48	76	124
Sign language video Duration (h)	69.62	3.91	5.59	79.12	1.14	1.82	2.96
Number of frames (per view)	6.3M	362,319	521,219	7.2M	123,120	196,560	319,680
Number of clips	31,128	1,741	2,322	35,191	642	940	1,582
Camera views per SL video	1 HD + 1 RGB-D (frontal) + 1 HD (side)				480 VGA + 30 HD + 10 RGB-D		
Sentences	31,128	1,741	2,322	35,191	642	940	1,582
Vocabulary size	15,686	3,218	3,670		1807	2360	3260
Out-of-vocabulary	–	413	510				
Number of signers	8	5	6	9	3	5	6
Signers not in train set	–	0	1		2	2	
		2D keypoints				*3D keypoints*	
Body pose		25				25	
Facial landmarks		70		137		70	
Hand pose (two hands)		21 + 21				21 + 21	
Automatic sign-level annotations	31,128	1,741	2,322	35,191	642	940	1,582

Table 3.3: **Statistics of the *Green Screen studio* data by signer.** We present the number of videos recorded by signer (Videos), together with the total duration of the recorded videos in hours (Hours) and the number of utterances (Utterances) of each signer.

	Signer 1	**Signer 2**	**Signer 3**	**Signer 4**	**Signer 5**	**Signer 8**	**Signer 9**	**Signer 10**	**Signer 11**	**Total**
					Train					
Videos	50	22	163	24	899	994	18	-	43	**2213**
Hours	1.89	0.82	3.80	0.82	31.59	28.28	0.67	-	1.72	**69.59**
Utterances	892	422	1859	398	12102	14596	292	-	486	**31047**
					Test					
Videos	16	16	37	-	47	42	-	26	-	**184**
Hours	0.51	0.53	1.05	-	1.67	1.08	-	0.71	-	**5.55**
Utterances	224	243	538	-	621	449	-	268	-	**2343**
					Validation					
Videos	17	19	27	-	37	32	-	-	-	**132**
Hours	0.57	0.68	0.65	-	1.20	0.79	-	-	-	**3.89**
Utterances	276	270	306	-	454	433	-	-	-	**1739**

signer that is not present in the training set; this subset of 26 videos can be used for measuring *generalization across different signers*. In total, 9 signers participated in the Green Screen studio recordings, and 6 signers in the Panoptic studio recordings. Table 3.3 presents detailed statistics of the videos from the How2Sign dataset recorded in the *Green Screen studio* grouped by signer.

3.4 Privacy, Bias and Ethical Considerations

In this section we discuss some metadata that we consider important for understanding the biases and generalization of the systems trained on our data.

Privacy. Recording sign language data may feel very personal since it is necessary to record videos of the person signing (at a minimum, the upper body including face and torso). Because of the small size of the deaf community, it may be possible to personally identify signers in videos even though if their names are not available. Signer's usually have privacy concerns when contributing to video datasets. This concerns include misuse of videos, being recognized, showing their surroundings when the video is recorded at their personal space, signing personal content when asked to tell stories about themselves, discomfort about looking presentable/attractive or concerns regarding their signing abilities [18]. One way of mitigating this problems would be by anonymizing the signer's face. However, facial expressions are a crucial component for generating and/or translating sign languages. With this in mind, before the video recordings, each signer attended an explanation session where the whole recording procedure was explained as well as how the data would be used and shared for research purposes. All the explanation was also delivered in a written document. After that, if they agreed with the procedure and were comfortable on being recorded and having their data shared, they were asked to sign a consent form stating their agreement. The collection of the dataset followed a set of procedures approved by the Carnegie Mellon University Institutional Review Board (IRB Registration Number: IRB00000352, Federalwide Assurance Number: FWA00004206) including a Social & Behavioral Research training completed by the author of this thesis.

Audiological status and language variety. The majority of the participants identified American Sign Language as the main language used during the recordings and sometimes making use of "contact signing" (Pidgin Sign English - PSE) . It is noteworthy that differences in audiological status can be correlated with different language use. The Deaf were likely to identify ASL as the main language used in the recording process. In contrast, the hearing were likely to identify a mix of contact signing and ASL as the main language use in the recording process. More information about PSE and ASL can be found in [105].

Geographic. All participants were born and raised in the United States of America (most of them in Pittsburgh, PA), and learned American Sign Language as their primary or second language at school time.

Signer variety. Our dataset was recorded by signers with different body proportions. Six of them were self-identified male and five self-identified female. The dataset was collected across 65 days during 6 months, which gives a variety of clothing and accessories used by the participants.

Data bias. Our data does not contain large diversity in race/ethnicity, skin tone, background scenery, lighting conditions and camera quality.

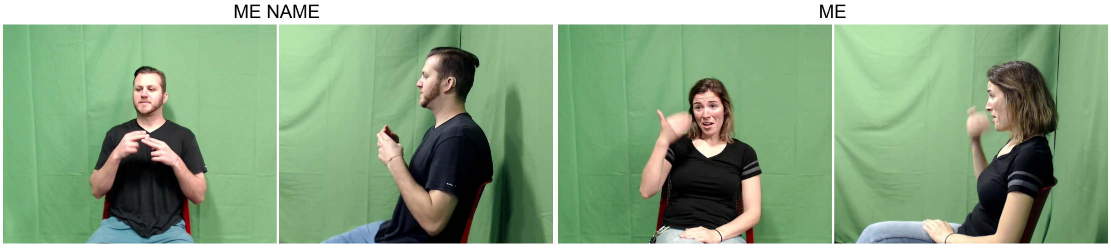


Figure 3.11: **Sample of language variety on our dataset.** Both signers were translating the sentence "I am". We can see that the signer on the left used the casual approach of signing it (ME NAME) while the signer on the left used the formal approach (ME).

Factors that may impair accurate automatic tracking. During the recording, signers were requested to not use loose clothes, rings, earrings, watches, or any other accessories that might impair accurate automatic tracking. They were also asked to wear solid colored shirts (that contrasted with their skin tone).

Out-of-vocabulary and signer generalization. Although not specifically designed for this, the How2Sign dataset can be used for measuring generalization with respect to both out-of-vocabulary words and signers. The dataset contains 413 and 510 out-of-vocabulary words, *e.g.* words that occur in validation and test, respectively, but not in training. It further contains duplicate recordings on the test set by a signer that is *not present in the training set*; these recordings can be used for measuring generalization across different signers and assessing how well the models can recognise or translate the signs given an out of the distribution subject.

Language variety. As discussed previously, our dataset contains variations in the language used by each signer during the recordings. In addition to that, we also would like to mention that sign language users can also use different signs or different linguistic registers (*i.e.*, formal or casual) to express the same given sentence. As we can see in Figure 3.11, two signers from our dataset used two different signs in a linguistic register to express the phrase "I am". The signer on the left used the casual approach of signing (ME NAME) while the signer on the left used the formal approach (ME).

Intra-sign variety. In addition to the multiple variety of signs and linguistic registers, it is also common to notice differences in the way that different people would perform the same sign. For example, we can see on Figure 3.12 two signers from our dataset signing the word "hair". In this sign, as described by its gloss annotation (IX-LOC-HAIR) the signer should point to their own hair location. This pointing is not specified and can vary depending of the person or even depending on the position that the person is in the moment. We can observe in this situation that both signers choose different locations to point to their hair. This is a common practise in signed languages that can bring challenges especially for computer vision-based models.

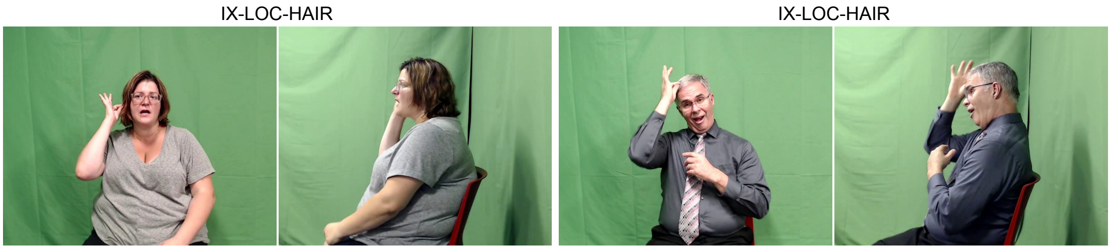


Figure 3.12: **Sample of intra-sign variety.** In this case, both signers are signing the word “hair" (IX-LOC-HAIR). We can see that the signer on the left chose to point to her hair with a different gesture from the signer on the right.

3.5 Final Remarks

In this Chapter, we have presented the full process of collecting and building the How2Sign dataset. The design, management, collection, processing, annotation, storage and manipulation of the How2Sign is not just one of the biggest contributions of this thesis, but also the process with more challenges and learning moments of my entire researcher career until now.

The collection of the How2Sign challenged me to leave my personal and professional comfort zone and deal with situations that were not common for me until that time neither were expected during this PhD journey. The first challenge was for sure putting in practice all the leanings about ASL, management and data collection to be able to recruit and manage more than thirty people including, Deaf signers, interpreters, ASL users, linguists, technical and financial support. The How2Sign was a product of the dedication and contribution of several people that worked hard for this project to succeed.

After the data collection, the post-processing phase brought different challenges, such as the manipulation and annotation of a large-scale collection of data. In total, the How2Sign counts with six different types of data, that accounts for more than twenty terabytes of raw data. Most part are videos that contain about seven million frames for each camera recording (in total we have 4 streams of videos), and an large amount of different folders and files that needed to be organized, processed, annotated and made available in a manageable way. A storage and data manipulation strategy (in special videos) is something that needs to be well planned and executed to avoid data loss.

The annotation phase, brought to us unexpected challenges that could have being avoided with a better design and testing phase. As explained before, annotating such large amounts of data specially with specific annotations that requires specialists can become a extreme hard task. In our case, we did not properly account the difficulties in finding specialized ASL linguists neither the large amount of time that this task can take and the different annotation pace in which different people work. This miscalculation regarding the time of annotation and people management as well as resources needed led to a interruption of the data collection after several months of unsuccessful efforts. With such process we have learn the importance of a well planned and calculated design phase with several rounds of annotation testing to proper estimate the time, efforts and human and financial resources needed.

Later, in the final phase, the manipulation, storage and distribution of the data taught us the technical difficulties of the process. Having a proper server to store and distribute date is an expensive and challenging task that should be taken into account and planned to be available for a long period of time. Accounting such expense and future technical support is something that should be included in the final budget.

In summary, for future data collection efforts, the lesson learned with this process was to plan, calculate and account methodically every possible problem and leave margin for errors, that they will naturally appear during the process.

Part II

Sign Language Meets Computer Vision

Introduction

Although sign languages have been investigated and studied by linguists for many decades, only recently they have caught the attention of computer linguistics and computer vision researchers. Vision-based sign language research, has mainly focused on building systems that can understand, recognize and translate sign languages to spoken/written languages or vice versa, aiming at creating a more natural way of communication between signers and non signers. However, most research to date has mainly focused on isolated sign recognition and spotting, neglecting the underlying rich grammatical and linguistic structures of sign language that differ from spoken language.

Tackling complex tasks like Continuous Sign Language Recognition (CSLR) or Sign Language Translation (SLT) and Production (SLP) has only recently become a tangible goal for the computer vision research community. With advancements in deep learning modeling, the first annotated datasets and weakly-supervised algorithms have recently emerged [8, 85, 98, 101, 104, 106–108]. However, the area is still in its infancy: most datasets available are very small and/or restricted to a specific domain lacking annotated data and researchers mainly focus on the same tasks mentioned above. We believe that such practices, although important, hinders the progress in algorithmic design and benchmarking.

In this Part of the thesis, we present our contribution to the development of new tasks, approaches to existing problems and annotation tools that include sign language in their pipeline. In the following Chapters we explore different computer vision tasks developed thanks to the How2Sign dataset presented in Chapter 3. In particular, we have addressed sign language recognition (Chapter 4), sign language video retrieval (Chapter 5) and sign language video generation (Chapter 6).

4
Sign Language Recognition

4.1 Introduction

Sign language Recognition (SLR) is the task of recognizing individual signs in a video sequence, and potentially classify them into individual spoken words or glosses. Sign Language videos can contain letters (fingerspelling), isolated or a full sequence of signs. Depending on the levels of syntactic hierarchy, the name of this task can appear as: (i) *fingerspelling recognition*– when the task aims to recognise just individual letters; (ii) *sign language recognition*– when the task aims to recognise isolated signs or; (iii) *continuous sign language recognition*– when the task aims to go one step further and recognise a full sentence or sequence of signs. It is important to note that in the former case, the underlying linguistic rules of sign language are not taken into account and for this reason not considered a full translation task.

Early works in sign recognition date back in 1983 where Grimes [109] propose a glove-based system that used sensors to detect the movement of the hand, twisting and flexing of the wrist and flex of the finger joints that was able to recognize 26 different signs. Since then, hundreds of works have been proposed using different approaches, including recently the successfully use of deep neural networks.

Deep architectures, however, are known to be “data hungry” and require a large amount of data in order to be successfully trained. To apply such models for sign language recognition, it is necessary to have datasets with sign-level annotations. This is usually not the case, since such fine-grained annotations are extremely costly to obtain. In order to address this problem, in this Chapter, we propose an automatic framework to generate sign-level pseudo-labels. It employs sign spotting techniques that can produce sparse annotations over continuous sign language data. We use this framework to automatically annotate the How2Sign dataset and use a sign recognition model based on a 3D convolutional neural network to establish a baseline for this task.

4.2 Related Work

Sign Language Recognition was one of the first sign language research directions explored by the computer vision community [109, 110]. Early works tackled this task focusing just on a few of the manual markers of sign language by computing hand-crafted features for hand shape and motion [82, 111–113]. Later, additional markers such as the upper body and the hand pose where incorporated in the pipeline [114–116]. Although non-manual features are as important as the manual ones, face [92, 113, 117], and mouth patterns [118–120] are relatively less considered. In order of modeling sequence of signs, Hidden Markov Models [111, 121–123], and later Long short-term memory (LSTM) models [124, 124–126], have been used. More recently the community has started to employ convolutional neural networks to consider and model the signer's whole appearance [127]. In particular, the I3D architecture, originally developed for action recognition [128], has proven to be effective for sign recognition [6, 72, 79, 129–131]– we similarly employ this model in sign recognition pipeline. A broader survey of works that tackle isolated and continuous sign language recognition are presented by Koller et al [132] and Rastgoo et al [133].

Sign spotting. Although the sign spotting problem has been formulated some time ago [82, 134], it was just recently used by [131] to localise signs in continuous news footage and improve a proposed isolate sign classifier. Since then this idea has been used, similarly to our work, to weakly annotate continuous sign language data and improve the performance of sign language tasks such as recognition [79, 104, 129].

Automatic annotation of sign language with auxiliary cues. The abundance of audio-aligned subtitles in broadcast data with sign language interpreters has motivated a rich body of work that has sought to use them as an auxiliary cue to annotate signs. Cooper and Bowden [135] propose to use a priori mining to establish correspondences between subtitles and signs in news broadcasts. Alternative approaches investigate the use of Multiple Instance Learning [114, 136, 137]. Other recent contributions leverage words from audio-aligned subtitles with keyword spotting methods based on mouthing cues [79], dictionaries [104] and attention maps generated by transformers [8] to annotate large numbers of signs, as well as to learn domain invariant features for improved sign recognition through joint training [138].

Similarly to these works, we also aim to automatically annotate sign language videos by making use of audio-aligned subtitles. To this end, we make use of prior keyword spotting methods [79, 104]. However, differently from all the other methods mentioned above we propose an *iterative* approach, SPOT-ALIGN, that alternates between repeated sign spotting (to obtain more annotations) and jointly training on the resulting annotations together with dictionary exemplars (to obtain better features for spotting).

4.3 Enhancement of Sign Language Annotations

As mentioned in Chapter 1, a key challenge for machine learning-based sign language methods is the lack of data, more specifically, datasets that are fine-grained annotated. To the best of our knowledge, there are no large-scale public datasets of *continuous* signing with corresponding sign-level annotations in ASL. This is due the fact that such annotations are extremely expensive and time consuming to produce manually.

To address this challenge, in this Chapter, we present SPOT-ALIGN, which is an automatic and iterative method to obtain sign-level annotations for continuous sign language videos based on different sign spotting techniques. Sign spotting methods aim at detecting and recognizing isolated signs, from a set vocabulary, in a given sequence of continuous signing. Such methods usually use different sign language related cues (*e.g.* mouthing, isolated signs from a dictionary collection) to guide the recognition and classification. However, as mentioned before, when dealing with continuous sign language data, co-articulation and the visual domain gap between isolated and continuous signing data are challenges to be addressed.

To that end, we tackle both problems by first obtaining a collection of sign-level annotations for the How2Sign dataset using two different sign spotting techniques. Such methods were proposed in recent works and employ mouthing cues [79] and dictionary examples [104]. We improve the sign language visual representations and obtain better visual features and iteratively increase the amount of dictionary-based annotations following a re-train and re-query framework. Next, we describe each of these steps.

4.3.1 Sign spotting methods

Mouthing-based sign spotting [79] First, we use the mouthing-based sign spotting framework of [79] to identify sign locations corresponding to words that appear in the written How2Sign translations. This approach, which relies on the observation that signing sometimes makes use of mouthings in addition to head movements and manual gestures [48], employs the keyword spotting architecture of [139] with the improved P2G phoneme-to-grapheme keyword encoder proposed by Momeni et al. [140]. We search for keywords from an initial candidate list of 12K words that result from applying text normalisation [141] to words that appear in How2Sign translations (to ensure that numbers and dates are converted to their written form, e.g. "7" becomes "seven") and filtering to retain only those words that contain at least four phonemes. Whenever the keyword spotting model localises a mouthing with a confidence over 0.5 (out of 1), we record an annotation. With this approach, we obtain approximately 37K training annotations from a vocabulary of 5K words. We filter these words to those that appear in the vocabulary of either WLASL [5] or MSASL [6] lexical datasets. Both are largely adopted and related ASL isolated signs datasets that were collected from a large range of videos on the internet. The resulting 9K training mouthing spottings (M) annotations cover a vocabulary of 1079 words, which consists of our initial vocabulary for training a sign recognition model.

Dictionary-based sign spotting [104]. Next, we employ an exemplar-based sign spotting method similar to [104]. This approach considers a handful of video examples per sign which are used as visual queries to compare against the continuous test video. The location is recorded as an automatic annotation for the queried sign at the time where the similarity is maximised. Such similarity measure between the query and the test videos requires a joint space. In [104], a complex two-stage contrastive training strategy is formulated. For our framework, we opt for a simpler mechanism in which we jointly train a sign recognition model with an I3D backbone on the set of query videos (which are often from an isolated domain such as lexical dictionaries) and sign-annotated videos from our search domain (i.e. How2Sign sparse annotations obtained from the previous step of mouthing-based spotting). The latent features from this classification model (which are now approximately aligned between the two domains) are then used to compute cosine similarities.

Similarly to the mouthing method, we select candidate query words for each video based on the subtitles. However, when employing dictionary spotting, we look for both the original and the lemmatised (removing inflections) forms of the words, since the sign language lexicons we employ usually contain a single version of each word (e.g. 'run' instead of 'running').

As the source of sign exemplars from which we construct queries, we make use of the training sets of WLASL [5] and MSASL [6], two datasets of isolated ASL signing, with 2K and 1K vocabulary sizes, respectively. For joint training, we select samples from their training subsets that occur in the 1079-sign vocabulary from our previous mouthing

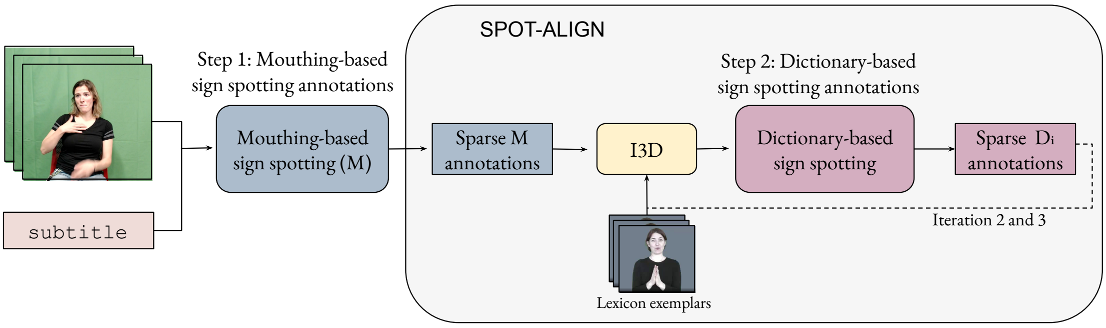

SPOT-ALIGN
Step 1: Mouthing-based sign spotting annotations
Step 2: Dictionary-based sign spotting annotations
subtitle
Mouthing-based sign spotting (M)
Sparse M annotations
I3D
Dictionary-based sign spotting
Sparse Di annotations
Iteration 2 and 3
Lexicon exemplars

Figure 4.1: **Spot-Align overview:** We propose Spot-Align, a framework for iteratively increasing annotation yield to obtain better visual sign language features. At each iteration i, the current I3D model is trained for sign language recognition (sign classification) jointly on the How2Sign annotations from iteration $i-1$ and lexicon exemplars from the WLASL [5] and MSASL [6] datasets. The resulting improved embedding is then used to obtain a new set of sign spottings by re-querying How2Sign videos with lexicon exemplars.

annotations. However, we use the full training sets for querying, allowing us to automatically annotate signs outside of the initial 1079 signs. We record all annotations where the maximum similarity (over all exemplars per sign) is higher than 0.75 (out of 1), resulting in 59K dictionary sign spottings (D_1) training annotations from an expanded vocabulary of 1887 signs. We initialise the I3D classification from the pretrained British Sign Language recognition model released by the authors of [78].

4.3.2 Spot-Align: An iterative sign language annotation framework

From the previous two methods, we obtain an initial set of automatic annotations. However, as mentioned before, the yield of the dictionary-based spotting method is heavily limited by the *domain gap* between the videos of How2Sign and the datasets used to obtain the exemplars. It is therefore natural to ask whether we can improve the output of the dictionary-based spotting by achieving a better feature alignment between the dictionary exemplar and How2Sign domains. To this end, we introduce a retrain-and-requery framework, which we call Spot-Align, to increase the amount of dictionary sign spottings (D) and obtain better aligned sign language visual features. We describe the followed procedure next.

At iteration i, we employ the I3D latent features obtained by joint training between WLASL-MSASL lexicons and How2Sign automatic annotations provided by iteration $i-1$ as illustrated in Figure 4.1. We observe a significant increase in the output annotations (*e.g.* 160K annotations in iteration D_2 vs 59K annotations in iteration D_1) despite using the *same exemplars* and *same subtitles* to construct our queries. The key difference is then the better aligned visual features with which we compare the exemplar from WLASL-MSASL and test videos from the How2Sign.

In Figure 4.2, we illustrate the resulting sparse annotations over a continuous timeline for sample videos where we observe that the density of annotations significantly increases with SPOT-ALIGN iterations. We denote with D_i, the set of automatic training annotations after applying iteration i.

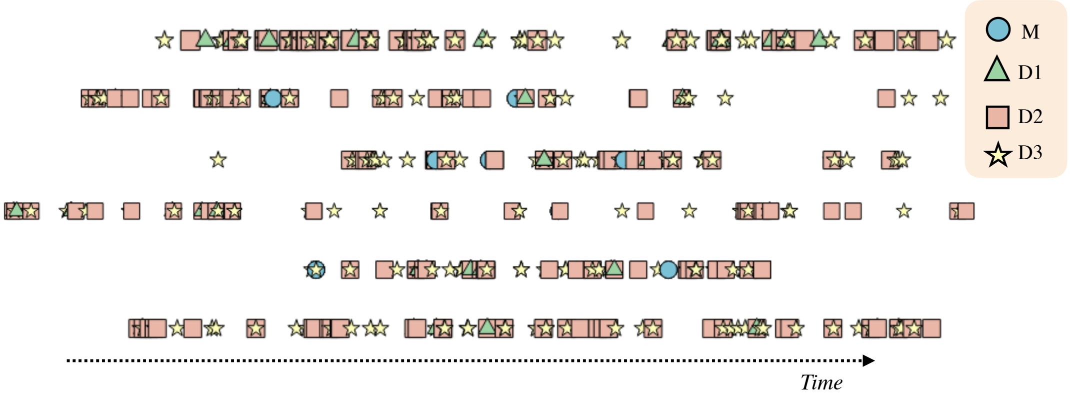


Figure 4.2: **Iterative enhancement of automatic annotations:** We illustrate sparse annotations generated by different iterations of the SPOT-ALIGN framework on six different video segments (rows) over a fixed-duration interval of 50 seconds each (x-axis).

4.4 Experiments

Automatic sign annotations. We train our sign recognition model using the automatic sparse annotations produced by the SPOT-ALIGN framework. Summary statistics obtained across multiple iterations of sign spotting are illustrated in Figure 4.3. To enable evaluation of sign recognition performance, we construct a manually verified test set. This is done by providing annotators proficient in ASL with sign spotting candidates using the VIA annotation tool [142]. This results in a labelled test set of 2212 individual sign video-category pairs, which have been made public as part of the How2Sign dataset.

Sign Recognition model and implementation details. The automatic annotations are used to train a sign recognition model that takes a multiple-frame video as input and outputs class probabilities over sign categories. This model is composed of a 3D convolutional neural network instantiated with an I3D [128] architecture pretrained on the BOBSL dataset [78]. We employ this architecture due to its success on action recognition benchmarks, as well as its recently observed success on sign recognition datasets [6, 72, 79].

We finetune this model on the How2Sign dataset using our automatic sign spotting annotations. In the final setting with mouthing (M) and dictionary (D_3) spottings from a vocabulary of 1887 signs, we have 206K training video clips, each corresponding to a

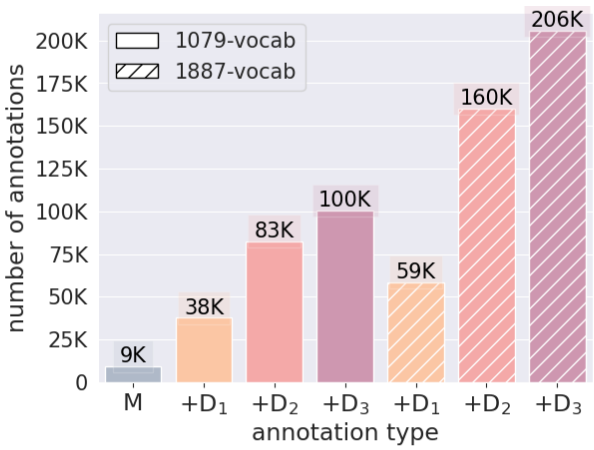


Figure 4.3: **Iteratively increasing the sign annotations:** Starting from a small set of mouthing annotations, we apply sign spotting through dictionaries several times, by retraining our I3D backbone on the previous set of automatic annotations. The left plot demonstrates the significant increase in the number of annotations, for both the restricted (1079) and the full (1887) set of categories.

single sign. Since the spottings represent a point in time, rather than a segment with beginning-end times, we determine a fixed window for each video clip. For mouthing annotations, this window covers a length of 0.8 seconds and it is defined as 15 frames before the annotation time and 4 frames after ($[-15, 4]$). For dictionary annotations, the window covers 1 second and it is similarly defined as 3 frames before the annotation time and 22 frames after ($[-3, 22]$). During training, we randomly sample 16 consecutive frames from this window, such that the RGB video input to the network becomes of dimension $16 \times 3 \times 224 \times 224$. We apply a similar spatial cropping randomly from 256×256 resolution. We further employ augmentations such as colour jittering, resizing and horizontal flipping.

We perform a total of 25 epochs on the training data, starting with a learning rate of 1e-2, reduced by a factor of 10 at epoch 20. We optimise using SGD with momentum (with a value of 0.9) and a minibatch of size 4. At test time, we apply a sliding window averaging in time, and center cropping in space. The sign recognition performance on the manually verified test set are reported in Table 4.1.

Limitations. The annotations produced by the Spot-Align method are limited to the vocabulary size of the queried videos. With this methodology we are not able to discover signs that do not appear in the vocabulary of the lexicon queried videos. We believe that other techniques, such as label propagation and the use of the available subtitles can be a possibility to expand the vocabulary of the automatic annotations. We leave this new approach for future work.

Training	Vocab	per-instance		per-class	
Annotation		top-1	top-5	top-1	top-5
M	1079	15.1	27.0	12.5	21.9
M+D1	1079	61.9	77.8	44.3	65.6
M+D2	1079	66.1	81.3	51.8	71.3
M+D3	1079	64.7	81.0	49.3	70.7
M+D1	1887	55.4	74.6	40.1	61.6
M+D2	1887	59.5	78.9	44.5	68.7
M+D3	1887	58.4	77.7	42.2	66.4

Table 4.1: **Sign recognition results:** We report the accuracy (%) results of individual sign recognition (1079-way and 1887-way classification) on the manually verified test set.

4.5 Final Remarks

In this Chapter we have demonstrated the feasibility of using sign spotting techniques to automatically annotate continuous sign language videos with signing-aligned subtitles. We utilize two sign spotting techniques, and a re-train and re-querying methodology that incorporates iterative rounds of sign spotting and feature alignment, and presented the Spot-Align framework that can be used to automatic sparse annotate sign language data with low annotation budget.

Such framework is an important tool for collecting automatic sign-level annotations that can later be used to improve more complex tasks such as sign language translation and production. We also provide a sign language recognition baseline for the How2Sign dataset that can be further used as a starting point for training more complex models. The automatic annotations are further used in the next Chapter for the novel task of sign language video retrieval.

5

Sign Language Video Retrieval

5.1 Introduction

Recent developments in automatic speech recognition (ASR) for spoken languages [143–146] have enabled automatic captioning of vast swathes of video content hosted on platforms such as YouTube. In addition to rendering the videos more accessible, this captioning yields a second important benefit: it allows the content of the videos to be indexed and efficiently searched with text queries. By contrast, the same automatic captioning capability (and hence searchability) does not exist for sign language content. Indeed, recent work has drawn attention to the pressing need to develop systems that can index archives of sign language videos to render them searchable [14]. Without these tools, sign language video creators must manually introduce the spoken language translation of their content if they want to reach the same discoverability as their spoken language counterparts.

One solution might appear to be to use sign language translation systems to perform video captioning, analogous to ASR cascading in spoken content retrieval [147]. Unfortunately, while promising translation results have been demonstrated in constrained domains of discourse (such as weather forecasts) [85, 106, 148], it has been widely observed that these systems are unable to achieve functional performance across diverse topics [8, 14, 108] required for open-vocabulary video indexing.

An alternative solution would be to employ existing methods for *sign spotting* to perform keyword search, as explored in Chapter 4. However, such approaches are fundamentally brittle—they work best when the user knows exactly which signs of interest were used in

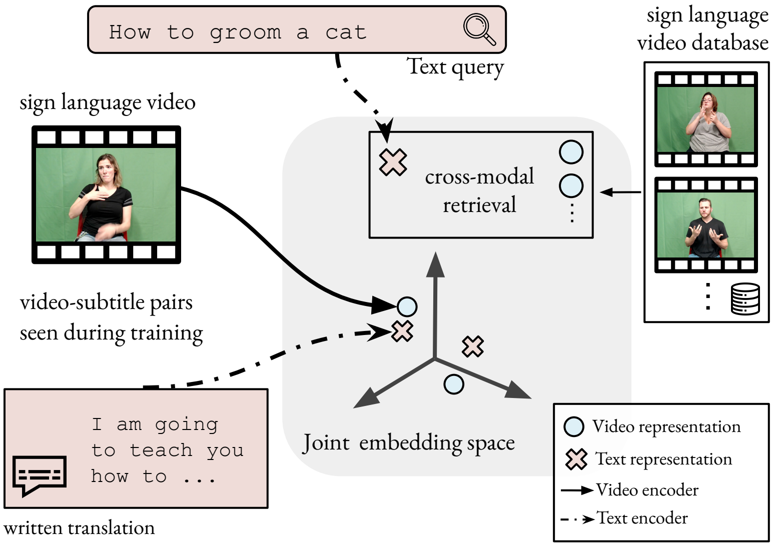


Figure 5.1: **Text-based sign language video retrieval:** In this Chapter we introduce *sign language video retrieval with free-form textual queries*, the task of searching collections of sign language videos to find the best match for a free-form textual query, going beyond single keyword search.

the video. Moreover, to build an accurate index of such signs using recent sign spotting techniques [79, 104, 149] requires a list of appropriate query candidates, which to date have often been obtained from subtitles corresponding to speech transcriptions of the translation, for example from an ASR engine. Our focus is on sign language videos produced by and for signers, that do not contain any speech track, so producing such transcriptions is not an option.

To address this gap, in this Chapter, we present the task of sign language retrieval with free-form textual queries: given a written query (e.g., a sentence) and a large collection of sign language videos, the objective is to find the signing video in the collection that best matches the written query. The terminology "natural language query" is commonly used to describe unconstrained textual queries in spoken languages. However, since sign languages are also natural languages, we adopt for the term "free-form textual query" instead.

We tackle this task by learning a joint embedding space between text and video as illustrated in Figure 5.1. Cross-modal embeddings target only the task necessary to enable search (i.e. ranking a finite pool of sign language videos), rather than the more involved task of full sign language translation. As we demonstrate through experiments, this renders their practical application even across multiple topics. Moreover, cross-modal embeddings enable extremely efficient search with the potential to scale up to collections of billions of videos thanks to mature approximate nearest neighbour algorithms for embedding spaces [150].

The task of sign language video retrieval is challenging for several reasons: (1) Translation mappings between sign languages and spoken languages are highly complex [48], with differing modalities and grammar structures (ordering is typically not preserved between signed and spoken languages, for example); (2) In contrast to the datasets used to train text-video retrieval models (millions of paired examples of videos with corresponding sentences [151, 152]) sign language datasets are orders of magnitude smaller in scale; (3) In addition to a paucity of paired data, the annotated data available for learning robust sign embeddings is also extremely scarce (with sign recognition datasets also considerably smaller than their counterparts for action recognition [153, 154], for example).

In order to address the first and second challenges highlighted above, we construct cross-modal embeddings that leverage pretrained language models to reduce the burden of data required to learn the mapping between signing sequences and sentences. To address the third challenge (annotation scarcity), we used the Spot-Align framework, described in Chapter 4, to automatically annotate significant fractions of the How2Sign dataset. By training on the resulting annotations, we obtain more robust sign embeddings for the downstream retrieval task.

5.2 Related Work

Text-video embeddings for video retrieval. Recently, there has been extensive research interest in enabling video content search with textual queries via cross-modal embeddings. Following the seminal DeViSE model [155] that demonstrated the strength of this approach for images and text, a wide array of text-video embeddings have been explored [152, 156–165]. Differently from these works which target the retrieval of *describable events*, our work focuses on retrieving *signing content* that matches a spoken language query formulated with text.

Sign language video retrieval. The task of sign language video retrieval has primarily been investigated under the query-by-example search paradigm, in constrained domains and with small datasets. In this formulation, a user query consists of an example of the sign(s) of interest, similarly how most keyword-based search engines deal with text databases. Two particular variants of this problem have received attention for sign language video retrieval: searching visual dictionaries of isolated signs, and searching continuous sign language datasets, both discussed next.

Sign language dictionaries are video repositories with recordings of individual signs suitable for learners. To search such videos, Athitsos et al. [166] coupled hand motion cues with Dynamic Time Warping (DTW) to enable signer-independent search of an American Sign Language (ASL) dictionary containing 3k signs and testing with 921 queries.

For *continuous sign language datasets*, the goal is retrieving all occurrences of a demonstrated query sign in a target video. Different techniques have been proposed for this purpose, including hand features with CRFs [167], hand motion with sequence matching [168], hand and head centroids [169], per-frame geometric features coupled with HMMs [170], and non-face skin distribution matching [82].

As an alternative to querying by example, a number of works have investigated sign spotting with learned classifiers. Ong et al. [134] tackled this problem with HSP-Trees, a hierarchical data structure built upon Sequential Interval Patterns. Later work combined human pose estimation with temporal attention mechanisms to detect (but not localise) the presence of a set of glosses among signing sequences [171]. This work was later extended to enable search for individual words [172] and further extended to additionally incorporate hand-shape features, improving performance [173]. More recently, Jiang et al. [149] showed the effectiveness of the transformer architecture for the sign spotting task, achieving promising results on the BslCorpus [77] and Phoenix2014 [132] datasets.

However, to the best of our knowledge, no prior sign language retrieval literature has considered the task that we presented in this Chapter, namely *retrieving sign language videos with free-form textual queries.*

5.3 Sign Language Retrieval

In this section, we first formulate the task of sign language video retrieval with free-form textual queries (Subsection 5.3.1). Next, in Subsection 5.3.2 we describe the cross-modal (CM) learning formulation considered in this work, and in Subsection 5.3.3 we describe our text-based retrieval through sign language recognition (SLR) model that uses the model described previously in Chapter 4.

5.3.1 Task formulation

Let $\mathcal{V}$ denote a *dataset* of sign language videos of interest, and let t denote a free-form textual user query. The objective of the *sign language video retrieval with textual queries* task is to find the signing video $v \in \mathcal{V}$ whose signing content best matches the query t. We use *text-to-sign-video* (T2V) as notation to refer to this task. Analogously to the symmetric formulations considered in the existing cross-modal retrieval literature [158, 174], we also consider the reverse *sign-video-to-text* (V2T) task, in which a signing video, v, is used to query a collection of text, $\mathcal{T}$.

5.3.2 Cross modal retrieval embeddings

To address the retrieval task defined above, we assume access to a parallel corpus of signing videos with corresponding written translations. We aim to learn a pair of encoders, ϕ_{V} and ϕ_{T}, which map each signing video v and text t into a common real-valued embedding space, $\phi_{\text{V}}(v), \phi_{\text{T}}(t) \in \mathbb{R}^C$, such that $\phi_{\text{V}}(v)$ and $\phi_{\text{T}}(t)$ are close if and only if t corresponds to the content of the signing in v. Here C denotes the dimensionality of the common embedding space.

To learn the encoders, we adopt the cross modal ranking learning objective proposed by Socher et al. [175]. Specifically, given paired samples $\{(v_n, t_n)\}_{n=1}^N$, we optimise a max-margin ranking loss:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1, i \neq j}^{B} [\eta_{ij} - \eta_{ii} + m]_+ + [\eta_{ji} - \eta_{ii} + m]_+ \tag{5.1}$$

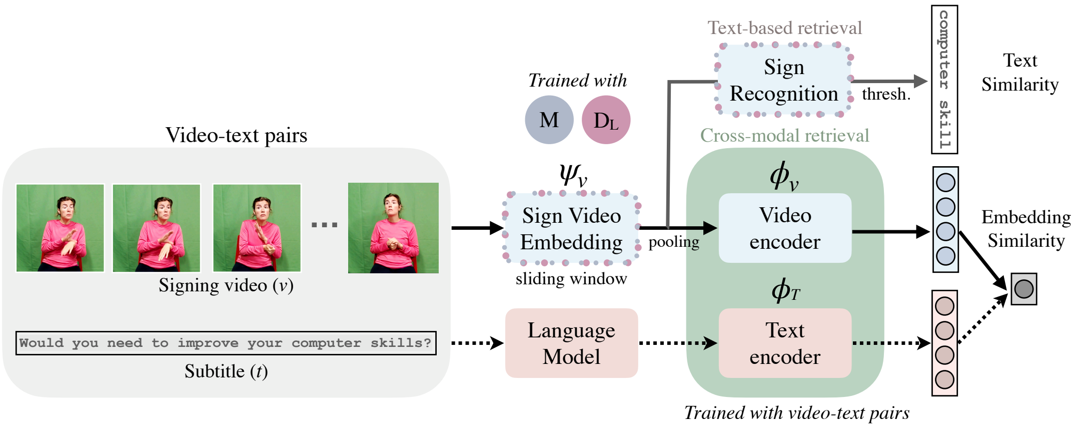
where m denotes the margin hyperparameter, $[\cdot]_+$ denotes the hinge function $\max(\cdot, 0)$, B denotes the size of minibatch sampled during training, and η_{ij} denotes the cosine similarity between signing video v_i and text t_j.

Once learned, the embeddings can be applied directly to both the T2V and V2T tasks. For the former, inference consists of simply computing the cosine similarity between the text query t and every indexed signing video $v \in \mathcal{V}$ to produce a ranking (and vice versa for the V2T task).

Encoder architectures. The sign video encoder, ϕ_{V} consists first of an initial *sign video embedding*, ψ_v, which we instantiate as an I3D neural network [128] over clips of 16 frames (motivated by its effectiveness for sign recognition tasks [5, 6, 79]). The output of ψ_v is temporally aggregated to a fixed size vector, and then projected to the C-dimensional cross modal embedding space, $\phi_{\text{V}}(v) \in \mathbb{R}^C$.

To implement ϕ_{T}, each text sample, t, is first embedded through a language model that has been pretrained on large corpora of written text. The resulting sequence of word embeddings are then combined via NetVLAD [176] and projected via Gated Embedding Unit following the formulation of [158] to produce a fixed-size vector, $\phi_{\text{T}}(t) \in \mathbb{R}^C$.

Here, we pay particular attention to the initial sign video embedding, ψ_s, which, as we show through experiments in Section 5.4, has a critical influence on performance. We also



(b) Sign language video retrieval

Figure 5.2: **Method overview:** To perform cross modal retrieval, we employ ψ_v, together with a language model, to produce embeddings of videos and text. These are passed to a video encoder and text encoder, respectively, which are trained to project them into a joint space such that they are close if and only if the text matches the video. The embedding produced by ψ_v is additionally passed to a sign recognition model, providing the basis for text-based similarity search.

conduct experiments to evaluate suitable candidates for both the temporal aggregation mechanism on ϕ_{V}, and the language model employed by ϕ_{T}. As shown in Figure 5.2, this embedding underpins the sign video encoder, ϕ_{V}, of our cross modal embedding, and is also used to classify individual signs to enable text-based retrieval, described next.

5.3.3 Text-based retrieval by sign recognition

The individual sign recognition model presented in Chapter 4 can naturally be used to obtain a *sequence* of signs if applied in a sliding window manner on the long signing videos from v. While the performance of this model is not expected to be high (due to a lack of temporal modelling stemming from the lack of continuous annotations), the output list of predicted sign categories gives us a set of candidate words which can be used to check the overlap with the query text. This is analogous to cascading ASR for spoken content retrieval [147], except that sign recognition is significantly more difficult than speech recognition (in part, due to a lack of training data [14]). Since the order of signs do not necessarily follow the order of the words in the translated text, we simply compute an Intersection over Union (IoU) to measure similarity between a query text and the recognised signs. Before we compute the IoU, we lemmatise both the query words and predicted words. We constrain the set of recognised signs by removing duplicates and removing classifications that have probabilities below a certain threshold (0.5 in our experiments). In the next section, we show that this text-based retrieval approach, while performing worse than the cross-modal retrieval approach, is complementary and can significantly boost overall performance.

5.4 Experiments

In this section we present the data and evaluation protocols used in our experiments (Subsection 5.4.1). Next, we provide retrieval results on How2Sign dataset, conducting ablation studies to evaluate the influence of different components of our approach (Subsection 5.4.3). Then, we establish baseline retrieval performances on the Phoenix2014T [85] dataset (Subsection 5.4.5). Finally, we present a qualitative analysis and discuss limitations and the societal impact (Subsection 5.4.6).

5.4.1 Data and evaluation metrics

Datasets. We use the videos and the temporally aligned subtitles of the How2Sign for training and evaluating the retrieval model, taking subtitles as textual queries. There are 31075, 1739 and 2348 video-subtitle pairs in training, validation and test sets, respectively. Note that, we remove a small number of videos from the original splits, where the subtitle alignment is detected to fall outside the video duration. We use the validation set to tune parameters (*i.e.* training epoch), and report all results on the test set.

We also evaluate our sign language retrieval method and provide baselines on the commonly used Phoenix2014T dataset [85], (although this is not our central focus due to its restricted domain of discourse). Phoenix2014T contains German Sign Language (DGS) videos depicting weather forecast videos. The dataset consists of 7096, 519 and 642 training, validation and test video-text pairs, respectively. The benchmark is primarily used for sign language translation where promising results can be obtained due to the restricted vocabulary size of 3K German words. Here, we re-purpose it for retrieval, providing baselines using both our cross-modal embedding approach, and a text-based retrieval by sign language translation [106].

Evaluation metrics. To evaluate retrieval performance, we follow the existing retrieval literature [158, 159, 163] and report standard metrics R@K (recall at rank K, higher is better) and MedR (median rank, lower is better). For cross modal embedding ablations (for which the sign video embedding ψ_v is frozen, and only the text encoder, ϕ_{T}, and video encoder, ϕ_{V}, are trained), we report the mean and standard deviation over three randomly seeded runs.

5.4.2 Implementation Details

Text embedding. We consider several text embeddings in this work. When conducting experiments on the How2Sign dataset, we explore the following English language embeddings:

GPT [177] is a 768-dimensional embedding that uses a Transformer decoder which is trained on the BookCorpus [178] dataset.

GPT-2-xl [179] is a 1600-dimensional embedding (employing 1558M parameters, also in a Transformer architecture [180]) that is trained on the WebText corpus (containing millions of pages of web text).

Albert-XL [181] is a 2048-dimensional embedding that builds on BERT [182] to increase its efficiency. It is trained with a loss that models inter-sentence coherence on the BookCorpus [178] and Wikipedia [182] datasets.

W2V [183] is a 300-dimensional word embedding, trained on the Google News corpus [1].

GroVLE [184]. This is a 300-dimensional embedding that aims to be vision-centric: it is adapted from Word2Vec [183]

For experiments on the Phoenix2014T dataset, we use a German language model:

German GPT-2 [185] (based on the original GPT-2 architecture of [177]) is a 768-dimensional embedding. The model is trained on the OSCAR [186] corpus, together with a blend of smaller German language data [2].

Sign video embedding. We employ the I3D recognition model (presented in Chapter 4) to instantiate our sign video embedding. More specifically, we use the outputs corresponding to the spatio-temporally pooled vector before the last (classification) layer. This produces a 1024-dimensional real-valued vector for each 16 consecutive RGB frames. We extract these features densely with a temporal of stride 1 from How2Sign sign language sentences to obtain the sequence of sign video embeddings.

Text-based retrieval. At test time, we obtain the predicted class for each 16-frame sliding window (with a stride of 1 frame), and record the corresponding word out of the 1887-vocabulary if the probability is above the 0.5 threshold. The resulting set of words are merged in case of repetitions, and are compared against the queried text to obtain an intersection over union (IOU) score, used as the similarity.

[1] we use the `GoogleNews-vectors-negative300.bin.gz` model from https://code.google.com/archive/p/word2vec/

[2] We use the parameters made available at https://huggingface.co/dbmdz/german-gpt2

Cross-modal retrieval. The dimensionality of the shared embedding space (denoted by the variable C in Subsection 5.3.2) used is 512. The margin hyperparameter, m, introduced in Equation. 5.1 is set to 0.2, following [158]. All cross modal embeddings are trained for 40 epochs using the RAdam optimiser [187] with a learning rate of 0.001, a weight decay of $1E-5$ and a batch size of 128. For each experiment, the epoch achieving the highest geometric mean of R@1, R@5 and R@10 on the validation set was used to select the final model for test set evaluation. The NetVLAD [176] layer employed in the text encoder uses 20 clusters.

Spot-Align iterations. In Figure 5.3 we provide, as a reminder, a detailed sketch of the Spot-Align framework and how we obtain our Mouthing and Dictionary-based annotations for the How2Sign dataset, explained in the previous Chapter. On the left ([b], [d], [f]), we show the iterations for the joint training between How2Sign, WLASL [72] and MSASL [6], which is used for Dictionary-based sign spotting. On the right ([a], [c], [e], [g]), the training is only performed on the How2Sign dataset, which provides sign video embeddings for retrieval.

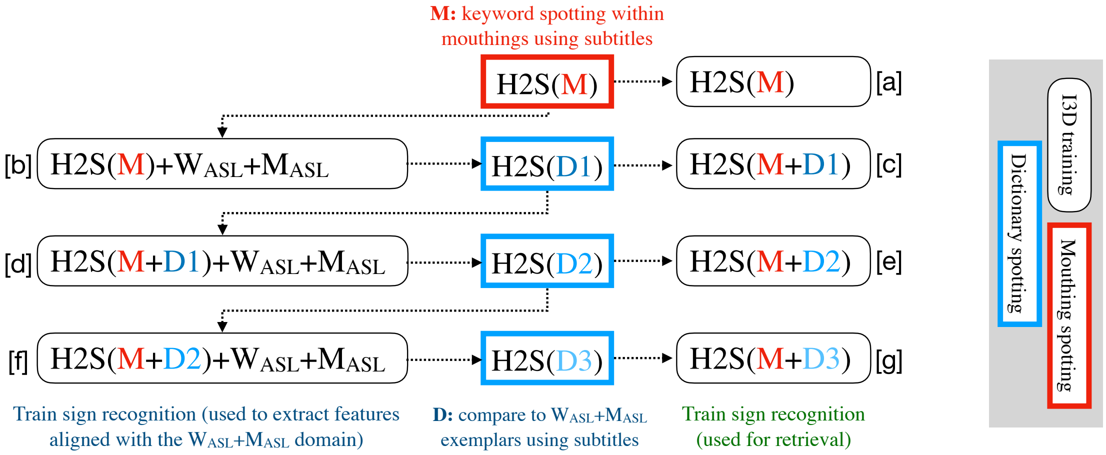

M: keyword spotting within mouthings using subtitles
H2S(M)
H2S(M)
[a]
[b]
H2S(M)+W$_{ASL}$+M$_{ASL}$
H2S(D1)
H2S(M+D1)
[c]
[d]
H2S(M+D1)+W$_{ASL}$+M$_{ASL}$
H2S(D2)
H2S(M+D2)
[e]
[f]
H2S(M+D2)+W$_{ASL}$+M$_{ASL}$
H2S(D3)
H2S(M+D3)
[g]
Train sign recognition (used to extract features aligned with the W$_{ASL}$+M$_{ASL}$ domain)
D: compare to W$_{ASL}$+M$_{ASL}$ exemplars using subtitles
Train sign recognition (used for retrieval)
I3D training
Dictionary spotting
Mouthing spotting

Figure 5.3: **Pipeline sketch of Spot-Align iterations:** [a] We use a Mouthing-based sign spotting to obtain an initial set of automatic sign-level annotations on the How2Sign (H2S) dataset which we call here H2S(M). [b] Using the automatic annotations obtained, we jointly train on the continuous signing examples from H2S(M) and the dictionary-style signing videos from WLASL and MSASL, in order to obtain a feature space aligned between the two domains. A Dictionary-based sign spotting approach is then used to obtain a new set of sign spottings (D1) by re-querying How2Sign videos with lexicon exemplars. The process is then iterated with the new spottings, as described in the main paper.

5.4.3 Retrieval results on How2Sign

In this section, we present ablation studies experimenting with: (i) different sign video embeddings, (ii) different initialisations, (iii) video embedding aggregation mechanisms, and (iv) text embeddings. We further study (v) the probability threshold hyperparameter for text-based retrieval via sign recognition. We also highlight (vi) the importance of having a sign language aligned subtitle data by experimenting with using the original speech-aligned timings provided by [36]. Finally, we demonstrate the advantages of (vii) combining our cross-modal embedding similarities with text-based similarities via sign recognition.

(i) Comparison of sign video embeddings. Our main results on sign language retrieval are summarised in Table 5.1. Here, we assess the quality of our end-to-end video classification model to obtain sign video embeddings from the last layer of the I3D model. We report both the sign language retrieval from the sign recognition outputs (using text-to-text matching, as described in Subsection 5.3.3) on the left, and the learned cross-modal embeddings (text-to-video matching) on the right.

We first observe that our cross-modal embeddings (which can potentially capture cues beyond the limited categories of the sign recognition model) perform significantly better than their text-based counterparts. Next, we compare various choices of backbone sign video embeddings to evaluate the effectiveness of our proposed Spot-Align framework. As a first baseline, we experiment with using standard Kinetics [128] training—we observe that this produces video embeddings that (as expected) perform poorly for our task. We also include as a baseline the model from [78] (pretrained on the BOBSL data) that was used to initialise our I3D sign video embedding. Next, we investigate the sensitivity of our model to different initialisations.

Table 5.1: **Effect of sign video embeddings:** The iterative increase of sign annotations with mouthing- (M) and dictionary-based (D) spotting improves the performance for sign video retrieval tasks with both sign recognition and cross-modal embeddings. The embeddings for the last seven rows are obtained from How2Sign trainings, pretrained on BOBSL (second row), which itself was pretrained on Kinetics (first row).

		Text-based retrieval				Cross-Modal retrieval			
Sign-Vid-Emb	Vocab	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
Kinetics [128]	-	-	-	-	-	$1.0_{0.1}$	$4.4_{0.4}$	$6.9_{0.6}$	$296.8_{12.5}$
BOBSL [78]	-	-	-	-	-	$17.2_{0.6}$	$32.5_{0.7}$	$39.5_{1.3}$	$30.5_{2.2}$
M	1079	0.6	2.3	4.4	1174.5	$16.4_{1.2}$	$31.1_{0.8}$	$38.2_{0.8}$	$32.7_{3.1}$
M+D$_1$	1079	10.2	21.2	26.5	136.3	$20.6_{1.1}$	$36.7_{0.6}$	$43.3_{0.9}$	$22.0_{2.6}$
M+D$_2$	1079	15.6	29.0	33.9	92.0	$21.8_{0.4}$	$38.0_{0.6}$	$44.6_{0.8}$	$18.2_{2.0}$
M+D$_3$	1079	16.7	29.1	33.3	95.3	$21.9_{1.2}$	$38.2_{0.7}$	$44.8_{0.5}$	$18.7_{0.6}$
M+D$_1$	1887	14.1	26.1	31.4	88.0	$20.4_{0.6}$	$36.4_{0.3}$	$43.5_{0.7}$	$20.0_{1.0}$
M+D$_2$	1887	18.3	31.3	35.8	69.8	$23.7_{0.5}$	$\mathbf{40.8}_{0.1}$	$\mathbf{47.1}_{0.2}$	$\mathbf{14.7}_{0.6}$
M+D$_3$	1887	**18.4**	**32.2**	**36.6**	**68.0**	$\mathbf{24.5}_{0.2}$	$40.7_{1.1}$	$46.7_{0.7}$	$15.7_{1.5}$

While strongly outperforming Kinetics training, this model remains substantially weaker than the end-to-end ASL sign recognition training on How2Sign enabled by Spot-Align. We observe improvements from each of our Spot-Align iterations, instantiated from mouthing-only (M) annotations, expanded first in the number of annotations within the same vocabulary size of 1079, then expanded in the number of sign categories with 1887-way classification. In light of their superior performance, we use sign video embeddings trained with M+D_3 annotations from the 1887 large-vocabulary for the rest of our experiments on How2Sign.

(ii) Sensitivity to Initialisation. We provide in Table 5.2 comparisons for training with annotations from Mouthing (M) and the first iteration of Dictionary (D1) ([a] and [c] in Figure 5.3) spottings from four different initialisations: I3D weights pretrained on BOBSL [78], BSL-1K [79], WLASL [72], or randomly initialised. Note that all BOBSL, BSL-1K and WLASL models are also initialised from Kinetics. Here, we rerun the Dictionary-based sign spotting to obtain different sets of D_1 annotations by initialising from WLASL-pretrained and random weights (instead of BSL-1K model from [8] in the rest of the experiments). While random initialisation significantly hurts performance, the WLASL-pretrained model performs slightly worse than [78], demonstrating that our method can work provided a reasonable initialisation. Assuming access to WLASL is realistic since we use it in step [b].

Table 5.2: **Sensitivity to initialisation:** We investigate the effects of different initialisation for our sign video embedding. We experiment with random and WLASL initialisation. $D_{1,\text{BSL1K},ft(H_M\text{WM})}$ means obtaining D_1 by pretraining the [b] model (see Fig. 5.2) on BSL-1K [8] and finetuning jointly on H2S mouthing annotations and WLASL/MSASL exemplars.

		#tr.	Acc.	Text-based retrieval				Cross-modal retrieval			
Sign-Vid-Emb	Init [a][c]	ann.	top-5	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
M	BOBSL	9K	27.0	0.6	2.3	4.4	1174.5	$16.4_{1.2}$	$31.1_{0.8}$	$38.2_{0.8}$	$32.7_{3.1}$
M+$D_{1,\text{BSL1K},ft(H_M\text{WM})}$	BOBSL	38K	77.8	10.2	21.2	26.5	136.3	$20.6_{1.1}$	$36.7_{0.6}$	$43.3_{0.9}$	$22.0_{2.6}$
M	BSL-1K	9K	25.1	0.6	2.4	4.4	1174.5	$18.0_{0.7}$	$32.4_{0.6}$	$39.3_{0.7}$	$27.8_{1.6}$
M+$D_{1,\text{BSL1K},ft(H_M\text{WM})}$	BSL-1K	38K	77.7	10.4	22.6	27.6	131.5	$20.8_{0.8}$	$36.9_{0.9}$	$43.5_{0.8}$	$20.5_{0.5}$
M	WLASL	9K	23.5	1.1	2.9	4.2	1175.5	$11.3_{0.5}$	$23.0_{0.5}$	$29.5_{0.6}$	$67.3_{7.2}$
M+$D_{1,\text{WLASL},ft(H_M\text{WM})}$	WLASL	60K	72.7	9.3	19.9	24.5	208.8	$17.1_{0.6}$	$31.5_{0.6}$	$38.3_{0.4}$	$32.8_{2.5}$
M	random	9K	6.5	0.0	0.0	0.0	1174.5	$0.6_{0.1}$	$2.0_{0.2}$	$3.3_{0.5}$	$530.7_{16.3}$
M+$D_{1,\text{random},ft(H_M\text{WM})}$	random	136K	28.0	0.0	0.5	0.8	1175.0	$2.4_{0.2}$	$7.2_{0.2}$	$10.2_{0.0}$	$221.3_{6.4}$

We note that the performances of BOBSL versus BSL-1K pretraining are similar in Table 5.2. Our preliminary results also suggest that similar trends from Table 5.1 hold when pretraining all rows on BSL-1K instead of BOBSL.

We therefore report all our models ([a, c, e, g]) with BOBSL pretraining since this dataset [78] has recently become available (unlike the BSL-1K source data [8, 79], which is not public). However, we clarify that the Dictionary spottings were obtained with models ([b, d, f]) pretrained on BSL-1K.

Furthermore, we investigate whether the domain alignment between WLASL+MSASL exemplars and How2Sign is beneficial by comparing M+$D_{1,\text{BSL1K},ft(H_M\text{WM})}$ and M+$D_{1,\text{BSL1K}}$. The latter consists of 112K spottings (as opposed to 38K); however, the top-5 recognition accuracy drops to 60.0% (from 77.7%) suggesting the poor quality of feature alignment between the two domains in the absence of joint finetuning.

(iii) Video embedding aggregation. Next, we compare the use of different temporal pooling strategies on the sequence of sign video embeddings for a given sign language video. While more sophisticated temporal aggregations are possible, in this work, we opt for a simple and efficient average pooling mechanism, which has widely been shown to be effective for text-video retrieval tasks [158, 165]. In Table 5.3, we compare average pooling with maximum pooling over the temporal axis for each feature dimension. We observe that average pooling performs best.

Table 5.3: **Influence of sign video embedding aggregation strategy:** We compare temporal pooling strategies on the How2Sign retrieval benchmark. Performance metrics are reported as means and standard deviations over three randomly seeded runs.

Aggregation method	R@1↑	R@5↑	R@10↑	MedR↓
Max pooling	$23.3_{0.3}$	$39.7_{0.5}$	$46.3_{0.6}$	$15.3_{0.6}$
Avg. pooling	$24.5_{0.2}$	$40.7_{1.1}$	$46.7_{0.7}$	$15.7_{1.5}$

(iv) Text embedding. We then compare several choices of text embedding for the training of cross modal embeddings. We report the results in Table 5.4. We observe that word2vec [183] and GrOVLE [184] obtain competitive performance, outperforming higher capacity alternatives [177, 179, 181]. This phenomenon is also observed in [165], where the authors show that for a number of source text distributions, simpler word embeddings can outperform their "heavyweight" counterparts. We leave the end-to-end fine-tuning of the language models with our sign language translations to future work, which can potentially provide further improvements, and use GrOVLE embeddings for the rest of the experiments.

Table 5.4: **Influence of the text embedding:** We compare a variety of text embeddings on the How2Sign retrieval benchmark. Performance metrics are reported as means and standard deviations over three randomly seeded runs.

Text Embedding	R@1↑	R@5↑	R@10↑	MedR↓
GPT [177]	$15.4_{0.4}$	$30.5_{0.4}$	$37.6_{0.4}$	$30.2_{1.3}$
GPT-2-xl [179]	$17.0_{0.3}$	$32.5_{0.4}$	$39.6_{0.4}$	$25.7_{1.2}$
Albert-XL [181]	$19.7_{0.3}$	$36.7_{0.3}$	$43.8_{0.4}$	$19.2_{0.8}$
W2V [183]	$24.2_{0.4}$	$40.0_{0.4}$	$46.7_{0.2}$	$14.8_{0.3}$
GrOVLE [184]	$24.5_{0.2}$	$40.7_{1.1}$	$46.7_{0.7}$	$15.7_{1.5}$

(v) Sign recognition probabilities. Here, we ablate the text-based retrieval approach which employs a sign recognition classifier. Since the sliding window is applied at each frame densely, we obtain one sign prediction per frame (which can be very noisy). Consequently, an important hyperparameter for this method is the selection of which classification outputs to consider in our set of predicted words (which will in turn guides the text-based retrieval). Concretely, the hyperparameter we vary is the confidence threshold at which predictions are included as text tokens. We explore several threshold values in Table 5.5 and report retrieval performance. We observe that 0.5 performs best—we adopt this value for our remaining experiments.

Table 5.5: **Thresholding sign recognition probabilities:** We investigate the influence of the confidence threshold hyperparameter on How2Sign retrieval performance, and observe that 0.5 works best.

Threshold	R@1↑	R@5↑	R@10↑	MedR↓
0.00	13.1	26.5	32.0	75.5
0.10	13.4	26.4	32.4	74.0
0.25	17.5	30.9	35.4	56.5
0.50	**18.4**	**32.2**	**36.6**	**68.0**
0.75	15.0	27.9	32.4	91.0

(vi) Effect of alignment. As presented in Section 3.3.2, the How2Sign has two sets of English translation alignments. The original version that comes from the How2 dataset, which is aligned with the speech of the How2 videos, and a signing-aligned version where we manually re-aligned the English translation to match the sentences in the signing videos, described in Section 3.3.2. With this experiment, we highlight the importance of having aligned video-sentence pairs. We retrain cross modal embeddings on speech-aligned training data, and evaluate both the recognition and cross modal embedding models on the speech-aligned test data of How2Sign to compare with signing-aligned version. The results are reported in Table 5.6, where we observe that speech-aligned subtitles significantly damage retrieval performance.

Table 5.6: **Effect of subtitle alignment:** We report retrieval performance on How2Sign for models trained and evaluated on subtitles aligned to speech and to signing. We observe a significant drop in performance when using speech-aligned subtitles.

	Text-based retrieval				Cross-Modal retrieval			
Alignment	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
Speech	9.5	16.1	19.0	418.0	$5.9_{0.6}$	$13.6_{0.6}$	$18.0_{0.2}$	$483.5_{17.9}$
Signing	**18.4**	**32.2**	**36.6**	**68.0**	$\mathbf{24.5}_{0.2}$	$\mathbf{40.7}_{1.1}$	$\mathbf{46.7}_{0.7}$	$\mathbf{15.7}_{1.5}$

(vii) Combining several cues. Finally, in Table 5.7, we combine our two types of models based on sign recognition (SR) and cross-modal embeddings (CM). We perform *late fusion* (averaging the similarities, with equal weights) computed with individual models. Table 5.7 presents both T2V and V2T performances establishing a new benchmark on the task of retrieval for the recent How2Sign dataset. We conclude that sign recognition provides complementary cues to our cross-modal embedding training, significantly boosting the final performance.

Table 5.7: **Combination of models:** We report our final benchmark performance on How2Sign for the retrieval models based on text (from sign recognition) (SR) and cross modal (CM) embeddings. We observe that the two approaches are highly complementary.

	T2V				V2T			
Models	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
SR	18.4	32.2	36.5	68.0	11.5	27.9	33.3	66.0
CM	24.7	39.6	46.0	17.0	17.9	40.8	46.6	15.0
SR + CM	**32.8**	**47.7**	**52.9**	**7.0**	**23.3**	**48.5**	**53.7**	**7.0**

5.4.4 Qualitative results

Some qualitative examples of videos retrieved by our system are provided in Figure 5.4. We also qualitatively illustrate, with a project webpage [3], our retrieval results using the best model on the How2Sign dataset (SR+CM combination from Tab 5.7). For each query, we show the top three ranked videos as well as their corresponding topic category (see Section 3.3.2 for more details of video categories), signer ID and sentences (note that these are not used during retrieval, and are provided for visualisation purposes).

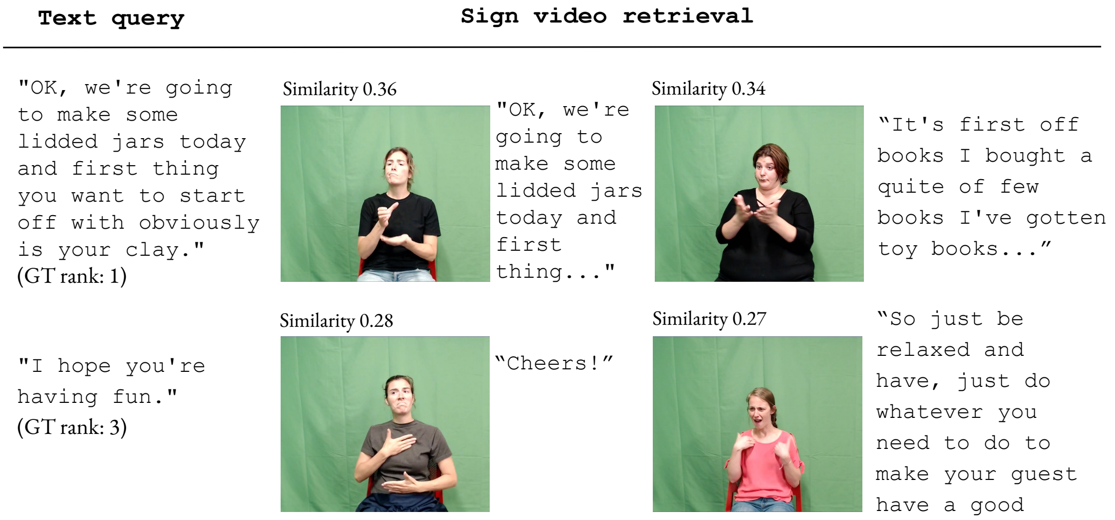
The top ten rows of the webpage show cases in which our model is able to correctly retrieve the video corresponding to the textual query. The middle five rows of the webpage show cases where the correct video is not retrieved successfully. For these failures, we nevertheless observe that the retrieval model makes reasonable mistakes (for instance, in the majority of cases, at least one of the top three ranked videos share the same topic category of the GT video). In the bottom five rows, we show examples of failure cases of our model.

In Figure 5.5, we illustrate two example queries for which the use of the sign recognition model substantially improves the performance of the cross modal embeddings.

5.4.5 Retrieval results on Phoenix2014T

In addition to the How2Sign ASL dataset that formed the primary basis of our study, we also provide retrieval baselines on the Phoenix2014T dataset [85, 132]. For cross modal embedding training, we employ a text embedding model trained on German language corpora, GPT-2 [179] released by Chan et al. [185]. For text-based retrieval, here

[3] `https://imatge-upc.github.io/sl_retrieval/app-qualitative/index.html`



"OK, we're going to make some lidded jars today and first thing you want to start off with obviously is your clay."
(GT rank: 1)

"OK, we're going to make some lidded jars today and first thing..."

"It's first off books I bought a quite of few books I've gotten toy books..."

"I hope you're having fun."
(GT rank: 3)

"Cheers!"

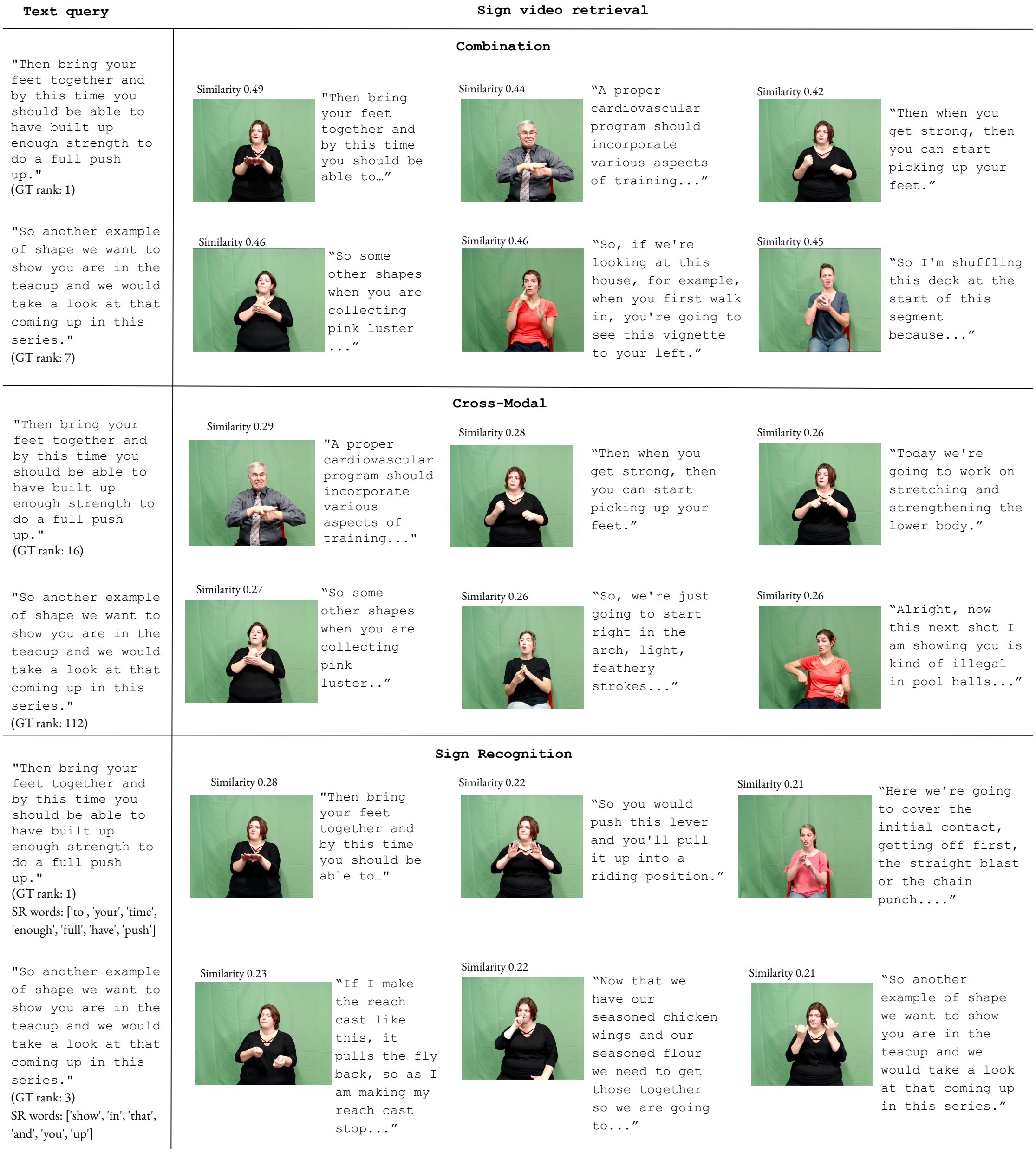
"So just be relaxed and have, just do whatever you need to do to make your guest have a good

Figure 5.4: **Qualitative results on text to sign language retrieval:** For each query, we show frames from the top two ranked videos as well as their corresponding sentences (these are not used during retrieval, and are provided for visualisation purposes). The top row shows a success case. The bottom row shows a failure case in which the retrieval system struggles with a less detailed query.

we incorporate a state-of-the-art sign language translation model [106], with which we compute an IoU similarity measure. Note that sign language translation performance is high on this dataset due to its restricted domain of discourse, which is the reason why we opt for a translation-based approach instead of the sign recognition-based retrieval as in Subsection. 5.3.3. The results are reported in Table 5.8. We observe that our cross-modal embeddings strongly outperform the translation-based retrieval. Their combination performs best (as in Table 5.7).

Table 5.8: **Retrieval performance on the PHOENIX2014T dataset:** We report baseline performances for cross modal embeddings, as well as text-based retrieval by sign language translation on the 642 sign-sentence pairs of the test set.

	T2V				V2T			
Text Embedding	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
Translation [106]	30.2	53.1	63.4	4.5	28.8	52.0	60.8	56.1
Cross-modal	48.6	76.5	84.6	2.0	50.3	78.4	84.4	**1.0**
Combination	**55.8**	**79.6**	**87.2**	**1.0**	**53.1**	**79.4**	**86.1**	**1.0**



Text query | **Sign video retrieval**

Combination

"Then bring your feet together and by this time you should be able to have built up enough strength to do a full push up."
(GT rank: 1)

Similarity 0.49 "Then bring your feet together and by this time you should be able to…"

Similarity 0.44 "A proper cardiovascular program should incorporate various aspects of training..."

Similarity 0.42 "Then when you get strong, then you can start picking up your feet."

"So another example of shape we want to show you are in the teacup and we would take a look at that coming up in this series."
(GT rank: 7)

Similarity 0.46 "So some other shapes when you are collecting pink luster ..."

Similarity 0.46 "So, if we're looking at this house, for example, when you first walk in, you're going to see this vignette to your left."

Similarity 0.45 "So I'm shuffling this deck at the start of this segment because..."

Cross-Modal

"Then bring your feet together and by this time you should be able to have built up enough strength to do a full push up."
(GT rank: 16)

Similarity 0.29 "A proper cardiovascular program should incorporate various aspects of training..."

Similarity 0.28 "Then when you get strong, then you can start picking up your feet."

Similarity 0.26 "Today we're going to work on stretching and strengthening the lower body."

"So another example of shape we want to show you are in the teacup and we would take a look at that coming up in this series."
(GT rank: 112)

Similarity 0.27 "So some other shapes when you are collecting pink luster.."

Similarity 0.26 "So, we're just going to start right in the arch, light, feathery strokes..."

Similarity 0.26 "Alright, now this next shot I am showing you is kind of illegal in pool halls..."

Sign Recognition

"Then bring your feet together and by this time you should be able to have built up enough strength to do a full push up."
(GT rank: 1)
SR words: ['to', 'your', 'time', 'enough', 'full', 'have', 'push']

Similarity 0.28 "Then bring your feet together and by this time you should be able to…"

Similarity 0.22 "So you would push this lever and you'll pull it up into a riding position."

Similarity 0.21 "Here we're going to cover the initial contact, getting off first, the straight blast or the chain punch...."

"So another example of shape we want to show you are in the teacup and we would take a look at that coming up in this series."
(GT rank: 3)
SR words: ['show', 'in', 'that', 'and', 'you', 'up']

Similarity 0.23 "If I make the reach cast like this, it pulls the fly back, so as I am making my reach cast stop..."

Similarity 0.22 "Now that we have our seasoned chicken wings and our seasoned flour we need to get those together so we are going to..."

Similarity 0.21 "So another example of shape we want to show you are in the teacup and we would take a look at that coming up in this series."

Figure 5.5: **Qualitative results:** We show two samples where text-based retrieval using sign recognition (SR) helps retrieval when combined with cross-modal embeddings (CM). Top, middle and bottom rows show the retrieval results for the same query using the average of the similarities from SR and CM (Combination), Cross-Modal and Sign Recognition models, respectively.

5.4.6 Limitations and Societal Impact

Limitations Qualitatively, we observe failure cases of our cross-modal retrieval model (illustrated in Figure 5.4), when using more generic queries that lack precise detail. It is also worth noting that all datasets used in our experiments are interpreted sign language rather than conversational (e.g. conversations between native signers – see [14] for a broader discussion on how this can limit models trained on such data).

Societal Impact. The ability to efficiently search sign language videos has a number of useful applications for content creators and researchers in the deaf community. However, by providing this technical capability, it also potentially brings increased risk of surveillance of signers, since large volumes of signing content can be searched automatically.

5.5 Final Remarks

In this work, we introduced the task of sign language video retrieval with free-form textual queries. The proposed task can be considered more challenging than retrieving isolated signs or "generic" cross-modal retrieval given the complexity of signed languages. In this case, the trained model must be able to capture the *linguistic* content of the video, while the large cross-modal retrieval literature would simply capture its *semantic* content (*e.g.* "an interpreter signing").

We believe this task can bring challenging and beneficial research problems to the computer vision community in addition to contributing to solve real-world problems. We envision research on this task to: (i) massively improve the quality of sign language video indexing and search on online video collections, enabling sign language users to search the web using their native and preferable language; (ii) facilitate and speed up learning sign language via example retrieval, and (iii) provide a highly challenging task motivating more researchers to work on the very challenging and impactful area of sign language understanding.

We further provided results for this task on the How2Sign and Phoenix2014T datasets establishing strong baselines on both datasets which we hope will serve to stimulate research in the area of sign language understanding.

6
Sign Language Video Generation

6.1 Introduction

Automatic Sign Language Video Generation refers to the task of automatically creating sign language video animations or avatars [1] given a body pose information. Those videos can be generated in different ways, such as via video rendering softwares or *synthesized* by generative models, like Generative Adversarial Networks (GANs) [188].

The generation of a human-like character can help improving the interpretation of signed languages in virtual scenarios turning it easier to comprehend and becoming more realistic for sign language users. But it is important to note that the Sign Language Video Generation task differs from Sign Language Production (SLP), where the translation from spoken to sign language is also required before generating a human-like signer on top of the predicted pose information– this process can also be done in an end-to-end approach, but has not been explored yet.

Sign Language Video Generation is an alternative to make the recent present methods that are working towards sign language production [94–97, 189] more accessible to sign language users. Such methods currently predict a temporal sequence of keypoints that form a skeleton representation from spoken language. In these systems, skeletons are represented by 2D/3D coordinates of human joints also known as *keypoints*.

Keypoints can carry detailed human pose information and can be an alternative for reducing the computational bottleneck that is introduced when working with the actual

[1] An avatar is a cartoon-style computer-animated character

video frames. However, no studies have been made so far on whether they are indeed useful when it comes to understanding sign language by its users.

In this chapter we present a study where we aim to understand *if and how well sign language users understand automatically generated sign language videos* that use 2D keypoints from How2Sign as sign language representation. In addition to skeleton visualizations, we go one step further and also generate realistic videos using a popular method for human motion transfer called Everybody Dance Now (EDN) [7]. We run this study with a set of ASL signers and record their understanding of the generated videos in terms of the category of the video, the translation into American English, and a final subjective rating about how understandable the videos were.

6.2 Related work

Human Motion Transfer. The task of human motion transfer, consists of transferring the motion of a person from a source to a target video. It has been used in different applications, such as in the synthesis of videos of a person imitating dancing movements from another person from a different video [7, 190, 191], to predict future frames an synthesize new videos [192, 193], and to generate videos with different backgrounds [194, 195] among others [196–200].

These models usually have as input a source pose, represented by 2D/3D keypoints or a body mask, and output the appearance of a person on top of those representations performing such pose/movement. In that sense, when applied to sign language research, these methods can be used, for example, as an addition block in the sign language production pipeline. In this case, the produced poses can be used to condition a human synthesis module and generate a human-like character in order to make the sign language output more realistic and comprehensible by sign language users. However, there has been limited research regarding the generation of novel poses not seen in source videos or conditioned on a given input. Most works have been restricted to conditioning pose generation on a given action [201] or audio [202, 203].

Sign Language Production the automatic translation from spoken to sign language, has just recently been explored with deep learning approaches [101, 107, 204–206]. Previously, this task was tackled in two steps. First, the motion of a native signer would be captured, or created by a expert in 3D animation in collaboration with a deaf person, and later transferred to an avatar [207–209]. This is very expensive and non-scalable approach that brings several challenges for the computer animation community as the motions generated must be realistic and have a precise semantic meaning.

Initial attempts to automatic SLP have focused on the production of concatenated isolated signs that disregard the grammatical syntax of sign language [95, 206]. Saunders et al. proposed the first SLP model to produce continuous sign language sequences direct from source spoken language [107]. A Progressive Transformer model was introduced that uses a counter decoding technique to predict continuous sequences of varying length represented as sequences of skeleton visualizations. Stoll et al. produced photo-realistic signers, but using low-resolution isolated signs that do not generalise to the continuous domain [95].

6.3 Generating Sign Language Videos

6.3.1 Video Synthesis

In order to generate the sign language videos used in this user study, we experiment with two different methods: 1) skeleton visualizations and 2) Generative Adversarial Network generated (GAN-generated) videos.

Skeleton visualizations. A common way of visualizing a set of estimated keypoints from videos is by connecting the modeled joints and create a wired skeleton following a given kinematic tree (detailed information about the kinematic tree used here can be found in Section 3.3.3). By visualizing the sign language poses in a sequence, one can have access to the human pose and motion information without requiring the full pixel-level video frames, avoiding the computational bottleneck that comes with it. Middle row of Figure 6.2 shows an example of a wired skeleton visualization.

GAN-generated videos. We use generative models to synthesize videos conditioned by the detected keypoints. To generate the animated video of a signer given a set of keypoints, we use an off-the-shelf motion transfer and synthesis approach called Everybody Dance Now (EDN) [7]. Given a video of a source person and another of a target person, the goal of this model is to generate a new video of the target enacting the same motions as the source. To accomplish this task, the overall pipeline is divided into three stages– pose detection, global pose normalization, and mapping from normalized skeleton visualizations to the target subject.

In the pose detection stage P, a pre-trained state-of-the-art pose estimation model is used to create skeleton visualizations, or also called pose stick figures in the original work, from the frames of the source video. The global pose normalization stage $Norm$ accounts for differences between the source and target body shapes and locations within the frame. And finally, a system was designed to learn the mapping from the skeleton visualizations to images of the target person using adversarial training.

The video synthesis method is based on Pix2PixHD [210], where the generator network G engages in a minimax game against a multi-scale discriminator D [188]. The generator is trained to synthesize images that would fool the discriminator, which must distinguish between "real" (ground truth) images and "fake" images produced by the generator. The two networks are trained simultaneously and drive each other to improve themselves–G learns to synthesize more detailed images to trick D, which in turn learns differences between generated outputs and ground truth data. In this case, G synthesizes images of a person given a skeleton visualization.

To be able to generate realistic sequences of images, EDN was further enhanced with a learned model of temporal coherence for better video and motion synthesis between adjacent frames by predicting two consecutive frames. Instead of generating individual frames, the model predicts two consecutive frames where the first output is conditioned on its corresponding skeleton visualization and a zero image (a placeholder since there is no previously generated frame at time $t-2$). The second output is conditioned on its corresponding skeleton visualization and the first output. Consequently, the discriminator must now determine both the difference in realism and temporal coherence between the "fake" sequence and "real" sequence. The temporal smoothing changes are also reflected in the updated GAN objective.

In addition to the temporal smoothing, EDN also includes a separate module for high resolution face generation, which is highly desirable in our case since facial landmarks are one of the critical features for sign language understanding. Figure 6.1 shows the overall pipeline of the EDN model.

Implementation Details. Below we specify the implementation details of each visualization method.

OP-skeletons. The keypoints used in this study were predicted from the videos of the How2Sign dataset using OpenPose [23]. We extracted the keypoints from the full-resolution videos and connect them use the original kinematic tree from OpenPose. We explain the keypoints extraction process in Section 3.3.3.

EDN-generated videos. The EDN model was trained using the original hyper-parameters detailed in [7]. Each model was trained from scratch using the videos of one signer from the How2Sign dataset. More specifically, keypoints extracted from videos of the first signer (top row in Figure 6.2) were used to learn the model that generates realistic videos of the second signer (bottom row) [2]. The subset used to train the model consisted

[2] A sample of a generated video can be seen at: https://youtu.be/w0xWUyXX6Ys

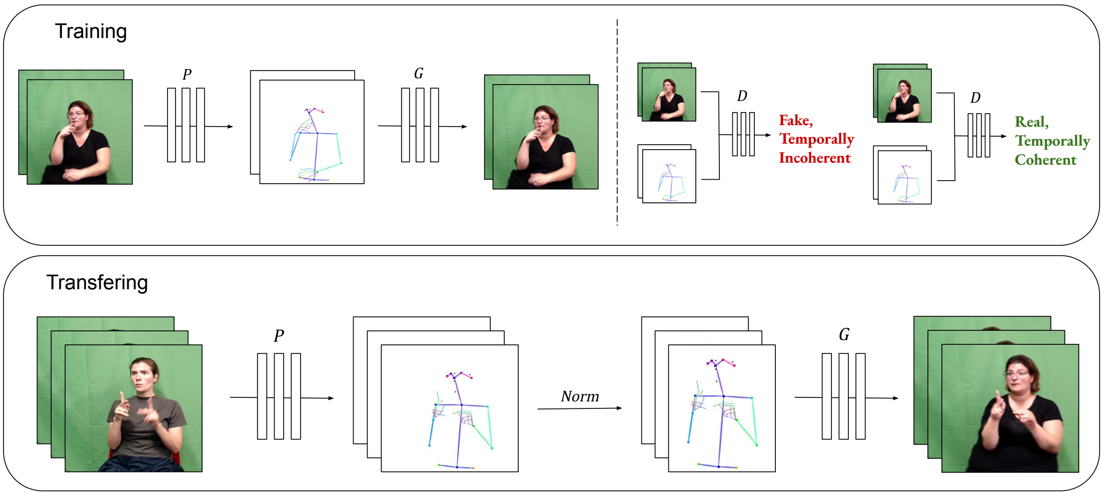


Figure 6.1: **Training** (Top): The EDN model [7] uses a pose detector P to create skeleton visualizations from video frames of the target subject. A mapping G is learned alongside with an adversarial discriminator D which attempts to distinguish between the "real" correspondences, and "fake" sequences. **Transfer** (Bottom): The pose detector P is used to obtain pose keypoints for the source person that are transformed by a normalization process $Norm$ into joints for the target person for which skeleton visualizations are created. A training mapping G is then applied.

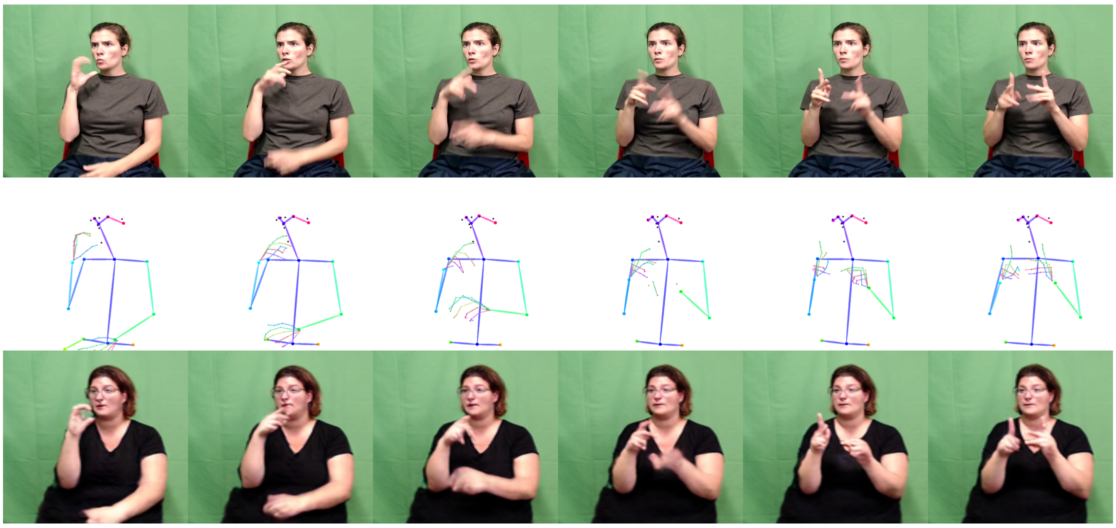


Figure 6.2: **Sample of generated Sign Language videos.** Source video (top row) was used to automatically extract 2D keypoints (middle row) and generate frames of a video with a different identity (bottom row).

of 28 hours of the training split– it consists of all the videos of "signer 8" and transfered to "signer 5".

6.3.2 Automatic evaluation

An approximate but automatic way of measuring the visual quality of the generated videos is by comparing the keypoints that can be reliably detected by OpenPose in the

Table 6.1: **Results on different OpenPose confidence scores.** Percentage of Detected Keypoints (PDK) and Percentage of Correct Keypoints (PCK) for all keypoints and just for the hands, when thresholding at different detection confidence scores of OpenPose (OP).

	PDK			PCK		
OP confidence scores	0	0.2	0.5	0	0.2	0.5
All keypoints	0.99	0.88	0.87	0.90	0.94	0.96
Hands	0.99	0.38	0.17	0.08	0.11	0.12

source and generated videos. We focus only on the 125 upper body keypoints which are visible in the How2Sign videos, and discard those from the legs. We use two metrics: a) the Percentage of Detected Keypoints (PDK), which corresponds to the fraction of keypoints from the source frame which were detected in the synthesized frame, and b) the Percentage of Correct Keypoints (PCK) [211], which labels each detected keypoint as "correct" if the distance to the keypoint in the original image is less than 20% of the torso diameter in all keypoints and 10% of the torso diameter for the hands.

In Table 6.1 we present these metrics for different minimum confidence thresholds of the OpenPose (OP keypoint detectors). We report results for all keypoints, as well as when restricting the evaluation only on the hand keypoints. We see that although the repeatability of keypoints is high in general, the model fails to predict reliable keypoints for the hands. This limitation is especially relevant in sign language processing.

6.3.3 Evaluation with the user in the loop

We evaluate the degree of understanding for both skeleton visualizations and the GAN-generated videos by showing 3-minute-long videos to four ASL signers. Two signers watched the skeletons visualizations, while the other two watched the GAN-generated videos. During the evaluation, each subject was asked to: a) classify six videos between the ten video categories (see subsection 3.2 for more information about the dataset categories); b) answer the question *"How well could you understand the video?"* on the five-level scale ((1) Bad, (2) Poor, (3) Fair, (4) Good, (5) Excellent); and c) watch two clips from the previously seen video and translate them into American English. Results averaged over all subjects are presented in Table 6.2. We report accuracy for the classification task, the Mean Opinion Score (MOS) for the five-scale question answers and BLEU [9] scores for the American English translations. Qualitative results are shown in Table 6.3.

Our results in Table 6.2 show a preference towards the generated videos rather than the skeleton ones, as the former result in higher scores across all metrics. In terms of general

Table 6.2: **Quantitative results.** Comparison between generated skeletons and GAN videos in terms of classification (Accuracy) between the ten video categories of How2Sign, mean opinion score (MOS) and translation (BLEU) [9].

	Acc.	MOS	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Skeleton	83.3 %	2.50	10.90	3.02	1.87	1.25
GAN-generated	**91.6 %**	**2.58**	**12.38**	**6.71**	**3.32**	**1.89**

Table 6.3: **Qualitative results.** Ground-truth (GT) and collected translations for two clips of the “Food and Drink" category. All subjects were able to correctly classify the category.

GT	*I'm not going to use a lot, I'm going to use very very little.*
Skeleton	That is not too much don't use much, use a little bit
EDN	Don't use a lot, use a little dont use lot use little bit
GT	*I'm going to dice a little bit of peppers here.*
Skeleton	cooking chop yellow peppers
EDN	cook with a little pepper chop it little bit and sprinkle

understanding of the topic, the subjects were able to mostly classify the videos correctly with both types of visualizations.

When it comes to finer grained understanding measured via the English translations, however, we can see from both Table 6.2 and Table 6.3 that neither skeletons nor GAN-generated videos are sufficient to convey important information needed from ASL signers to completely understand the sign language sentences. We hypothesize that current human pose estimation methods such as [23] are still not mature enough when it comes to estimate fast movements of the hands. We observed that due to the nature of sign language and the fast movements of the signers' hands, OpenPose lacks precision in those cases which can make the visualizations incomplete, harming the understanding of some important parts of sign language.

6.4 Final Remarks

In this chapter we have presented a study in which sign language videos generated from the automatically extracted annotations of our dataset were presented to ASL signers. To our knowledge, this is the first study that investigates how well keypoint-based synthetic videos, a commonly used representation of sign language production and translation, can be understood by sign language users.

In order to access this information, we proposed a evaluation protocol that include sign language users in the loop, and evaluate their understanding regarding two different tasks, video topic detection (easier) and English translation (harder). Both tasks were formulated using the available data in the How2Sign (video topic and English translation). Through this evaluation process, we show that subjects prefer synthesized realistic videos over skeleton visualizations; we also show that current video synthesis methods can generate videos that allows the understanding of sign language videos to a certain extent *i.e.*, the classification of the video category, but lack in fidelity to allow for a fine-grained understanding of the complete sign language sentence. We partially attribute poor understanding on the bad synthesis of the hands, and believe that future research towards that direction is highly important.

How can computer vision do better? Our results show that the EDN model used as an out-of-the-box approach is not enough for sign language video generation. Specifically, we show that the model struggles with generating the hands and detailed facial expressions, which play a central role in sign language understanding. We argue that human pose estimation plays an important key in this aspect and needs to be more robust to blurry images, especially in the hands and to fast movements in order to be suitable to sign language research. We also argue that it is worth pursuing generative models that focus on generating hand details, particularly on the movements of the fingers, as well as clear facial expressions on full-body synthesis.

7

Conclusion

In this chapter, we conclude this thesis by providing a summary of its contributions (Section 7.1) followed by a discussion and outlining some lines of future work (Section 7.2).

7.1 Summary of contributions

This thesis has addressed two main challenges of sign language understanding, the lack of a large-scale dataset and the development of vision-based machine learning models for sign language.

To tackle the lack of large-scale and reliable sign language datasets, our contributions are following:

- In Chapter 2 (Section 2.2), we present an **extensive survey on the existing sign language datasets** including information about the modalities they are composed of, as well as the vocabulary size, total duration of the videos, number of samples, domain of discourse, number of signers, video/image resolution, link to download (when available) and the dataset publication (if any). This survey can be used as a first step for future efforts in the development of methods that involve sign language as well as future data collection, providing the community with the information of all datasets up to date comprised in a single document. We also provide an online and collaborative version of this survey where we intend to keep updated for future use in the area.

- In Chapter 3, we presented the **How2Sign dataset**, describing the process of video recordings in both multiview studios, as well as the collection of the manual and non-manual annotations. We also provide information about costs and time spent for the construction of the dataset, as well as the pipeline used during our recordings. We conclude by describing our experience when working directly with subjects and dealing with large-scale amounts of data and what we have learned through the whole process. We believe the How2Sign have a large potential on supporting the advantage of the sign language research field given its quality and features, but we also believe that the provided information can support and facilitate future data collection and minimizing the challenges that were described in Chapter 1. In addition to that, How2Sign extends the How2 [99] dataset, an existing multimodal dataset with a new sign language modality, and therefore enables connecting with research performed in the vision, speech and language communities.

To address the development of vision-based machine learning models for sign language, our contributions are the following:

- In Chapter 4 we tackle the annotation scarcity challenge by presenting a framework based on sign spotting techniques that uses a re-train and re-query **methodology to automatically annotate** continuous sign language data with sparse sign-level labels. Using the presented framework, we annotated a large portion of the How2Sign dataset with sign-level annotations which allowed us to establish a **strong baselines for the Sign Language Recognition** task. We believe this baseline is an important starting point that can instigate the community to train and develop more complex and accurate models for this task.
- In Chapter 5, we introduce **a novel task– sign language video retrieval with free-form textual queries** and establish strong baselines for both the How2Sign and Phoenix2014T datasets. We believe that the proposed task brings challenging problems for the research community that can motivate further work on the area of content-based sign language video indexing and retrieval.
- In Chapter 6, we explore the **Sign Language Video Generation** task by applying a motion transfer technique to synthesize sign language videos. This pipeline is generally used by the computer vision community as a final step of sign language production models. However, no studies with sign language users were ever done before to access **if and how well they understand automatically generated sign language videos** and validate the used pipeline. To do so, we present a user-based evaluation protocol where we present two types of generated videos and ask

a group of ASL signer to evaluate the visualizations by responding which topic the video was about and by translating the ASL content back to English, both based on the How2Sign data. We believe such evaluation is beneficial to understand and design future approaches to sign language production and generation as well as understand how well off-the-shelf models work on sign language data.

7.2 Discussion and Future Work

Here we discuss some of the important points learned during this thesis and discuss future directions in the area of Sign Language understanding.

Inclusion of the Deaf community. First and foremost, we believe that future works should include the Deaf community from the beginning, in the design phase, until the end in the development and test phase. As pointed out by different members of different Deaf communities and recently in the study presented by Bragg et al. [14], sign language users (deaf people, sign language interpreters, sign language teachers and students, and hearing people who interact with deaf people professionally or socially) are the principal stakeholders of any future application of sign language research and should be integrated into the design and execution of it. This collaboration would avoid most of the common mistakes and inappropriate directions of research that can be taken when the research group lacks individuals with lived experience of the problems the technology could or should solve. It is also common to develop approaches that do not take into account the linguistic complexities of signed languages for which the algorithms must account. As a result, such single-disciplinary approaches to sign language understanding will often end up having limited real-world value [212]. In addition to that, sign languages are an important part of Deaf communities' identity and culture and should be respected and treated as such.

Data. Although various efforts, including this thesis, have been developed towards collecting large-scale and reliable sign language data that is suitable for the development of new technologies for sign language, as pointed out by Bragg et al [14], the lack of data is still one of the biggest challenges to be addressed in the field. Datasets that reflect real-world use cases, *e.g.* conversation, healthcare, public services, education, emergency situations, etc., are still needed. Up to date, the public sign language datasets still have shortcomings that limit the power and the ability to generalize systems trained on them. As discussed in Chapter 2, such shortcomings can fall into one or more of the following categories: limited size/duration, a limited domain of discourse/content, presence of non-native or fluent signers, signer variety, lack of proper annotations, and inclusion of linguist complexities of SL (*e.g.* use of classifiers). In addition to that, we believe that the development of tools to assist the data collection and annotations are also an essential and needed resource.

FATE in AI for Sign Language. With the increasing interest in the development of models and datasets to train artificial intelligence and machine learning systems to account for sign language, researchers should also be thoughtful and take into account the Fairness, Accountability, Transparency, and Ethics (FATE) of such models and data. Sign language datasets typically contain recordings of people signing, which is highly personal not just because it contains their face, but also their particular way of communicating and expressing themselves. The rights and responsibilities of the parties involved in data collection and storage are complex and involve different data sponsors, data collectors or owners, and data users. Deaf community members (and signers, more generally) are also central stakeholders in any end applications of sign language data. It is important to note that the centrality of sign language to deaf culture identity, coupled with a history of oppression, makes the usage of data and development of technology particularly sensitive [16]. While preliminary work has presented some approaches to preserve the identity of signers in sign language videos by using facial filters [18], it is not clear how to move forward in this context and preserve the privacy of signers while collecting or using sign language data.

Generalization to unseen situations and individuals is a major difficulty of machine learning, and tasks that involve sign language understanding are no exception. Sign Language recognition, translation and production models considered nowadays the state-of-the-art are usually trained in a small (11 hours) and restricted domain (weather forecast) dataset. Models trained in such data struggle to generalize to other domains as well as to deal with different types of data. Larger and more diverse datasets are essential for training generalizable models. However, generating such datasets can be extremely time-consuming and expensive. An alternative would be to develop models that are able to generalize better using limited resources, *e.g.* data, annotations, compute.

Synthetic data. One approach to address the scarcity, diversity and privacy of sign language data would be to create and use synthetic data. This approach has recently become popular in related computer vision domains to address low-level tasks such as body shape estimation and more recently human action recognition. Such tasks can usually be tackled with a more generic representation of the human body. However, sign language understating tasks requires fine-grained data that include movements of fingers and detailed facial expression. Such detailed data has not been largely explored in models that create body representation used to generate synthetic data. In addition to that, a recipe for how to create a good synthetic dataset for training is still unknown. To the best of our knowledge, there is no synthetically generated data that accounts for sign language.

We believe this is due to the scarcity of labeled data and the complexity of generating detailed human pose information. However, we believe the 3D pose information generated by the Panoptic studio presented in Chapter 3 has the potential to serve as the starting point potentially be used to generate synthetic 3D human pose representations with different backgrounds, clothing, skin color, illumination, etc.

Standard annotation system. As noted in Section 6.1 there is still no widespread manner of annotating or collecting sign language data. Label or annotation format immediately affects who can label the data, which can also additionally insert biases into labels or translations. Labels are required for documenting the contents of sign language data and to enable supervised learning of statistical models to learn mapping functions between the data and labels. A standard annotation system would expedite the development of sign language understanding tasks, where datasets annotated with a standard system could easily be combined and shared. A standard system would also reduce annotation costs and errors. Using complicated notation systems (*e.g.* linguistic notation systems vs. glossing) requires sophisticated software (*e.g.* ELAN [102]), which means that only trained annotators can contribute. This may exclude many willing deaf annotators from contributing and limit dataset size due to high costs [16].

Learn from past mistakes. New AI technologies that include sign languages in their pipeline are drawing the attention of computer science and linguistic researchers in the past years. Such tools can be powerful and offer significant benefits to deaf people. At the same time, as with any powerful tool, AI-enabled sign language technologies can pose risks of (unintended) harmful consequences. Here, we would like to conclude this thesis with this reminder and draw the attention of researchers that intend to pursue this path to such possible risks. Bragg et al [16] compile and presents a wonderful brief history of *deaf-related technology* (Section 2.3 of the referred work) and point out some common benefits and pitfalls so future work can learn from history and be better equipped to proceed thoughtfully.

A

Appendix A: Cross-modal video and audio retrieval

In this Appendix, we present our approach that explores a cross-modal retrieval technique that was later adapted to be used in the sign language context with the availability of the How2Sign dataset, presented in Chapter 5.

The increasing amount of online videos brings several opportunities for training self-supervised neural networks. Here we explore the large-scale video dataset YouTube-8M by taking advantage of the multi-modal information available [1]. By means of a neural network, we are able to create links between audio and visual files, by projecting them into a common region of the feature space, obtaining joint audio-visual embeddings. These links are then used to retrieve audio samples that fit well to a given silent video, and also to retrieve images that match a given a query audio. The results in terms of Recall@K obtained over a subset of YouTube-8M videos show the potential of this unsupervised approach for cross-modal feature learning. We train embeddings for both scales and assess their quality in a retrieval problem, formulated as using the feature extracted from one modality to retrieve the most similar videos based on the features computed in the other modality.

[1] The YouTube-8M dataset had been recently released by the time of this publication

A.1 Introduction

Videos have become the next frontier in artificial intelligence. Their rich semantics make them a challenging data type posing several challenges in both perceptual, reasoning or even computational level. Mimicking the learning process and knowledge extraction that humans develop from our visual and audio perception remains an open research question, and video contain both information in a format manageable for science and research.

Videos are used in this work for two main reasons. Firstly, they naturally integrate both visual and audio data, providing a weak labeling of one modality with respect to the other. Secondly, the high volume of both visual and audio data allows training machine learning algorithms whose models are governed by a high amount of parameters.

The popularization of deep neural networks among the computer vision and natural language processing communities has defined a common framework boosting multimodal research. Tasks like video sonorization, speaker impersonation or self-supervised feature learning have exploited the opportunities offered by artificial neurons to project images, text and audio in a feature space where bridges across modalities can be built.

This work exploits the relation between the visual and audio contents in a video clip to learn a joint embedding space with deep neural networks. Two multilayer perceptrons (MLPs), one for visual features and a second one for audio features, are trained to be mapped into the same cross-modal representation. We adopt a self-supervised approach, as we exploit the unsupervised correspondence between the audio and visual tracks in any video clip.

We propose a joint audiovisual space to address a retrieval task formulating a query from any of the two modalities. A video or an audio clip can be used as a query to search its matching pair in a large collection of videos. For example, an animated GIF could be sonorized by finding an adequate audio track, or an audio recording illustrated with a related video.

In this Appendix, we present a simple yet effective model for retrieving videos or audio files with a fast and light search. We do not address an exact alignment between the two modalities which would require a much higher computation effort. We make the code and trained model publicly available at https://github.com/surisdi/youtube-8m.

A.2 Related Work

In the past years, the relationship between the audio and the visual content in videos has been researched in several contexts. Overall, conventional approaches can be divided into four categories according to the task: generation, classification, matching and retrieval.

As online music streaming and video sharing websites have become increasingly popular, some research has been done on the relationship between music and album covers [213–216] and also on music and videos (instead of just images) as the visual modality to explore the multimodal information present in both types of data [217–220].

A recent study also explored the cross-modal relations between the two modalities but using images with people talking and speech [221]. It is done through Canonical Correlation Analysis (CCA) and cross-modal factor analysis. Also applying CCA, Zhang et al. [222] uses visual and sound features and common subspace features for aiding clustering in image-audio datasets. In a work presented by Ngiam el at. [223], the key idea was to use greedy layer-wise training with Restricted Boltzmann Machines (RBMs) between vision and sound.

This work is focused on using the information present in each modality to create a joint embedding space to perform cross-modal retrieval. This idea has been exploited especially using text and image joint embeddings [224–226], but also between other kinds of data, for example for creating a visual-semantic embedding [155] or using synchronous data to learn discriminative representations shared across vision, sound and text [227].

However, joint representations between the images (frames) of a video and its audio have yet to be fully exploited, being [228] the work that most has explored this option to the best of our knowledge by the time of the publication of this work. In Hong et al, they seek for a joint embedding space but only using music videos to obtain the closest and farthest video given a query video, only based on either image or audio.

The main idea of the this work is borrowed from [226], which is the baseline to understand our approach. In Salvador's et al. work, the authors create a joint embedding space for recipes and their images. Where they later use to retrieve recipes from any food image, looking to the recipe that has the closest embedding. Apart from the retrieval results, they also perform other experiments, such as studying the localized in the activation unit, or doing arithmetics with the images.

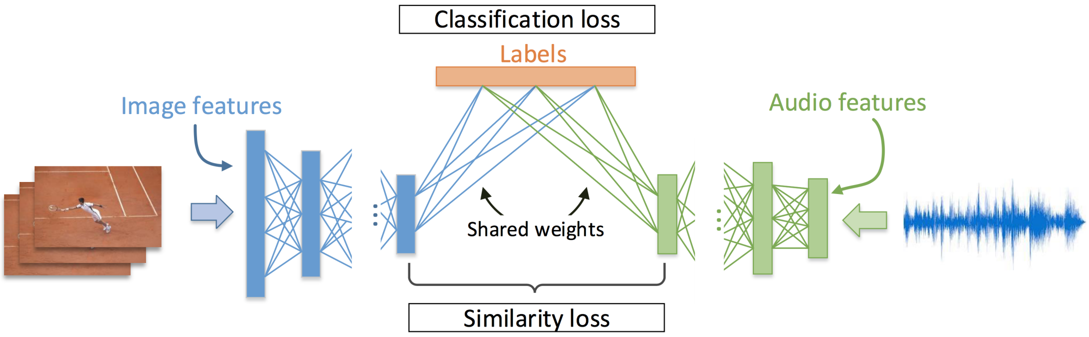


Figure A.1: **Model Architecture.** Here we illustrate the architecture used in the cross-modal retrieval task.

A.3 Cross-modal Video and Audio Retrieval

In this section we present the architecture for our joint embedding model, shown in Figure A.1. Our architecture have as input an image features vector and an audio features vector both pre-computed and provided together with the YouTube-8M dataset [229]. In particular, we use the *video-level* features, that represents the whole video clip with two vectors: one for the audio and another one for the video. These feature representations are the result of an average pooling of the local audio features computed over windows of one second, and local visual features computed over frames sampled at 1 Hz.

The main objective of the system is to transform the two different features vectors (image and audio, separately) to other features laying in a *joint space*. This means that for the same video, ideally the video and audio features will be transformed to the same joint features, in the same space. We will call these new features *embeddings*, and will represent them with Φ^i, for the image embeddings, and Φ^a, for the audio embeddings.

The idea of the joint space is to represent the *concept* of the video, not just the image or the audio, but a generalization of it. As a consequence, videos with similar concepts should be close in the embedding space and videos with different concepts should be further apart in the joint space. For example, the representation of a tennis match video will be close to the one of a football match, but not to the one of a maths lesson.

We use a set of fully connected layers of different sizes, stacked one after the other, going from the original features to the embeddings. They perform a non-linear transformation on the input features, mapping them to the embeddings, being the parameters of this non-linear mapping learned in the optimization process. It is important to note that both, the audio and the image network are completely separated. After that, a classification step is done, also using a fully connected layer using a sigmoid as activation function.

Each hidden layer uses ReLu as activation function, and all the weights in each layer are regularized using L2 norm.

A.3.1 Similarity Loss

Given the objective of getting the two embeddings of the same video to be as close as possible, while keeping embeddings from different videos as far as possible, we can formulate our problem as the following: given a video v_k, represented by the audio and visual features $v_k = \{i_k, a_k\}$ (i_k represents the image features and a_k the audio features of v_k), the objective is to maximize the similarity between Φ^i_k (the embedding obtained by transformations on i_k), and Φ^a_k (the embedding obtained by transformations on a_k).

At the same time, however, we have to prevent embeddings from different videos to be "close" in the joint space. In other words, we want them to have low similarity. However, the objective is not to force them to be opposite to each other. Instead of forcing them to have similarity equal to zero, we allow a margin of similarity small enough to force the embeddings to be clearly not in the same place in in the joint space, called α.

During the training, both positive and negative pairs are used, being the positive pairs the ones for which i_k and a_k correspond to the same video v_k, and the negative pairs the ones for which i_{k1} and a_{k2} do not correspond to the same video, this is, $k1 \neq k2$. The proportion of negative samples is p_{negative}.

For the negative pairs, we selected random pairs that did not have any common label, in order to help the network to learn how to distinguish different videos in the embedding space. The notion of "similarity" or "closeness" is mathematically translated into a cosine similarity between the embeddings, being the cosine similarity defined as:

$$similarity = \cos(x, z) = \frac{\sum\limits_{k=1}^{N} x_k z_k}{\sqrt{\sum\limits_{k}^{N} x_k^2}\sqrt{\sum\limits_{i}^{N} z_k^2}} \tag{A.1}$$

for any pair of real-valued vectors x and z resulting in the following loss:

$$\begin{aligned} L_{cos}((\Phi^a, \Phi^i), y) &= \\ &= \begin{cases} 1 - \cos(\Phi^a, \Phi^i), & \text{if} \quad y = 1 \\ \max(0, \cos(\Phi^a, \Phi^i) - \alpha), & \text{if} \quad y = -1 \end{cases} \end{aligned} \tag{A.2}$$

where $y = 1$ denotes positive sampling, and $y = -1$ denotes negative sampling.

A.3.2 Regularization

Inspired by the work presented by [226], we allow additional information to our system by incorporating the video labels (classes) provided by the YouTube-8M dataset. This information is added as a regularization term that seeks to solve the high-level classification problem, both from the audio and from the video embeddings, sharing the weights between the two branches. The key idea here is to have the classification weights from the embeddings to the labels shared by the two modalities.

This loss is optimized together with the previously explained similarity loss (Eqn A.3.1), serving as a regularization term. Basically, the system learns to classify the audio and the images of a video (separately) into different classes or labels provided by the dataset. We limit its effect by using a regularization parameter λ.

To incorporate this regularization to the joint embedding, we use a single fully connected layer, as shown in Figure A.1. Formally, we can obtain the label probabilities as $p^i = \text{softmax}(W\Phi^i)$ and $p^a = \text{softmax}(W\Phi^a)$, where W represents the learned weights, which are shared between the two branches. We use a softmax in order to obtain probabilities at the output. The objective is to make p^i as similar as possible to c^i, and p^a as similar as possible to c^a, where c^i and c^a are the category labels for the video represented by the image features and the audio features, respectively. For positive pairs, c^i and c^a are the same.

For the classification task, we used the cross entropy loss, as following:

$$L(x,z) = -\sum_k x_k \log(z_k) \tag{A.3}$$

Thus, the classification loss is:

$$L_{class}(p^i, p^a, c^i, c^a) = -\sum_k (p^i_k \log(c^i_k) + (p^a_k \log(c^a_k)) \tag{A.4}$$

Finally, the loss function to be optimized is:

$$L = L_{cos} + \lambda L_{class} \tag{A.5}$$

A.4 Experiments

A.4.1 Dataset

The experiments presented in this section were developed over a subset of 6,000 video clips from the YouTube-8M dataset [229]. This dataset does *not* contain the raw video files, but their representations as precomputed features, both from audio and video. Audio features were computed using the method explained in [230] over audio windows of 1 second, while visual features were computed over frames sampled at 1 Hz with the Inception model provided in TensorFlow [231].

The dataset provides *video-level* features, which represent all the video using a single vector (one for audio and another for visual information), and thus does not maintain temporal information; and also provides *frame-level* features, which consist on a single vector representing each second of audio, and a single vector representing each frame of the video, sampled at 1 frame per second.

The main goal of this dataset is to provide enough data to reach state of the art results in video classification. Nevertheless, such a huge dataset also permits approaching other tasks related to videos and cross-modal tasks, such as the one we approach in this paper. For this work, and as a baseline, we only use the *video-level* features.

A.4.2 Implementation details

We observe that starting with λ different than zero led to a bad embedding similarity because the classification accuracy was preferred. Thus, we began the training with $\lambda = 0$ and set it to 0.02 at step number 10,000, followed by a margin α of 0.2. The percentage of negative samples p_{negative} used was 0.6.

We used 4 hidden layers in each network branch, being the number of neurons per layer from features to embedding equal to 2000, 2000, 700, 700 in the image branch and 450, 450, 200, 200 in the audio branch. We trained with a batch size of 1024. We used the Tensorflow [231] code base provided by the authors of the *YouTube-8M* challenge [2].

A.4.3 Quantitative Evaluation

Evaluation metric. We obtain the quantitative results by applying the Recall@k metric. We define Recall@k as the recall rate at top K for all the retrieval experiments, this

[2] https://www.kaggle.com/c/youtube8m

Table A.1: **Audio to video retrieval.** We present the Recall@ 1, 5 and 10 when retrieving video given an audio query.

Size of the feature vector	Recall@1	Recall@5	Recall@10
256	21.5%	52.0%	63.1%
512	15.2%	39.5%	52.0%
1024	9.8%	30.4%	39.6%

Table A.2: **Video to audio retrieval.** We present the Recall@ 1, 5 and 10 when retrieving audio given a video query.

Size of the feature vector	Recall@1	Recall@5	Recall@10
256	22.3%	51.7%	64.4%
512	14.7%	38.0%	51.5%
1024	10.2%	29.1%	40.3%

is, the percentage of all the queries where the corresponding video is retrieved in the top K, hence higher is better.

The experiments are performed with different dimension of the feature vector. The Table A.1 shows the results of recall from audio to video. In other words, from the audio embedding of a video, how many times we retrieve the embedding corresponding to the images of that same video. Table A.2 shows the recall from video to audio.

To have a reference, the random guess result would be $k/\text{Number of elements}$. The obtained results show a very clear correspondence between the embeddings coming from the audio features and the ones coming from the video features. It is also interesting to notice that the results from audio to video and from video to audio are very similar, because the system has been trained bidirectionally.

A.4.4 Qualitative Evaluation

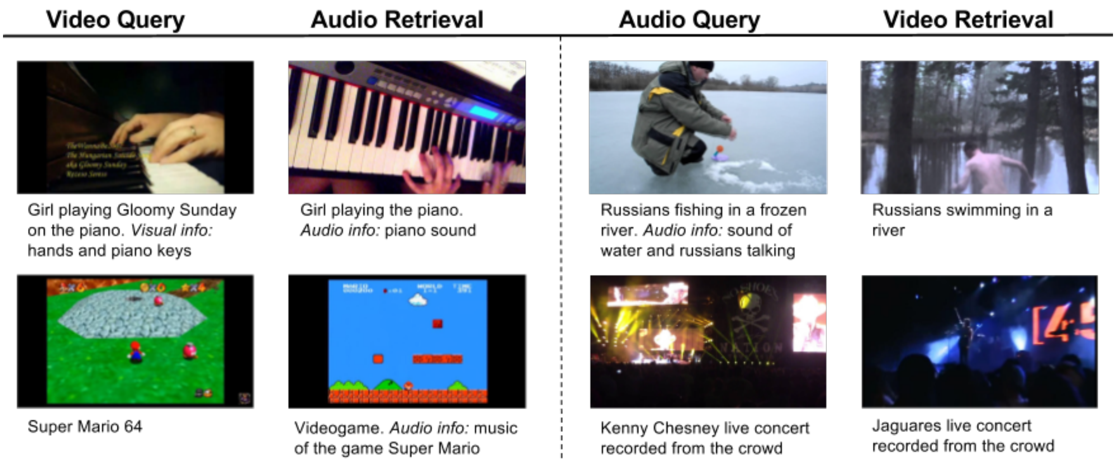
In addition to the objective results, we performed some insightful qualitative experiments. They consisted on generating the embeddings of both the audio and the video for a list of 6,000 different videos. Then, we randomly chose a video, and from its image embedding, we retrieved the video with the closest audio embedding, and the other way around (from one video's audio we retrieved the video with the closest image embedding). If the closest embedding corresponded to the same video, we took the second one in the ordered list.

The Figure A.2 shows some experiments. On the left, we can see the results given a video query and getting the closest audio; and on the right the input query is an audio. Examples depicting the real videos and audio are available online [3]. It shows both the results when going from image to audio, and when going from audio to image. Four

[3] https://goo.gl/NAcJah

different random examples are shown in each case. For each result and each query, we also show their YouTube-8M labels, for completeness.

The results show that when starting from the image features of a video, the retrieved audio represents a very accurate fit for those images. Subjectively, there are non negligible cases where the retrieved audio actually fits better the video than the original one, for example when the original video has some artificially introduced music, or in cases where there is some background commentator explaining the video in a foreign (unknown) language. This analysis can also be done similarly the other way around, this is, with the *audio colorization* approach, providing images for a given audio.



Girl playing Gloomy Sunday on the piano. *Visual info:* hands and piano keys

Girl playing the piano. *Audio info:* piano sound

Russians fishing in a frozen river. *Audio info:* sound of water and russians talking

Russians swimming in a river

Super Mario 64

Videogame. *Audio info:* music of the game Super Mario

Kenny Chesney live concert recorded from the crowd

Jaguares live concert recorded from the crowd

Figure A.2: **Qualitative results.** On the left we show the results obtained when we gave a video as a query. On the right, the results are based on an audio as a query.

A.5 Final Remarks

We presented an simple but effective method to retrieve audio samples that fit correctly to a given (muted) video. The qualitative results show that the already existing online videos, due to its variety, represent a very good source of audio for new videos, even in the case of only retrieving from a small subset of this large amount of videos. Due to the existing difficulty to create new audio from scratch, we believe that a retrieval approach is the path to follow in order to give audio to videos.

The range of possibilities to extend the presented work is excitingly broad. The first idea would be to make use of the YouTube-8M dataset variety and information. The temporal information provided by the individual image and audio features is not used in the current work. The most promising future work implies using this temporal information to match audio and images, making use of the implicit synchronization the audio and the images of a video have, without needing any supervised control. Thus, the next step in our research is introducing a recurrent neural network, which will allow us to create more accurate representations of the video, and also retrieve different audio samples for each image, creating a fully synchronized system.

Also, it would be very interesting to study the behavior of the system depending on the class of the input. Observing the dataset, it is clear that not all the classes have the same degree of correspondence between audio and image, as for example some videos have artificially (posterior) added music, which is not related at all to the images.

In short, we believe the YouTube-8M dataset allows for promising research in the future in the field of video sonorization and audio retrieval, for it having a huge amount of samples, and for it capturing multi-modal information in a highly compact way.

B

Appendix B: Image Generation from Audio

Recent works are making steps towards sign language production by automatically generating human pose keypoints from spoken language. However, such representation is not realistic and compromise the understanding of the language by its users (See Chapter 6 for more details). An ideal Sign Language Production pipeline would be and end-to-end approach where a spoken language speech input would be automatically translated into sign language representations and later conveyed by a synthesized human-like character performing the sign language translation.

However, such approach is yet not feasible with the available resources. Towards that end, we start by exploring the generation of facial features and expressions, which play an important role by carrying an large part of the grammar in sign languages [50]. Thus, in this in Appendix, we explore the potential of the speech signal to generate face images of a speaker by conditioning a Generative Adversarial Network with raw speech input. Our model is trained in a self-supervised manner, by exploiting the audio and visual signals naturally aligned in videos. With the purpose of training from video data, we present a novel dataset collected for this work, with high-quality videos of youtubers with notable expressiveness in both the speech and visual signals.

It is worth noting that this work does not use the How2Sign dataset and was developed as our first attempt on developing a pipeline for sign language video generation directly from speech. We leave the adaptation of this approach to an end-to-end approach of sign language production and video generation as future work.

B.1 Introduction

Audio and visual signals are the most common modalities used by humans to identify other humans and sense their emotional state. Features extracted from these two signals are often highly correlated, allowing us to imagine the visual appearance of a person just by listening to their voice, or build some expectations about the tone or pitch of their voice just by looking at a picture of the speaker. When it comes to image generation, however, this multimodal correlation is still under-explored.

In this Appendix, we focus on cross-modal visual generation, more specifically, the generation of facial images given a speech signal. Two recent approaches have recently popularized this research venue [232, 233]. Chung *et al.* [232] present a method for generating a video of a talking face starting from audio features and an image of the person's identity. Suwajanakorn *et al.* focus on animating a point-based lip model to later synthesize high quality videos of President Barack Obama [233]. Unlike the aforementioned works, however, we aim to generate the whole face image at pixel level, conditioning only on the raw speech signal (*i.e.* without the use of any handcrafted features) and without requiring any previous knowledge (e.g speaker image or face model).

To this end, we propose a conditional generative adversarial model (shown in Figure B.2) that is trained using the aligned audio and video channels in a self-supervised way. For learning such a model, high quality, aligned samples are required. This makes the most commonly used datasets such as *Lip Reading in the wild* [234], or *VoxCeleb* [235, 236] unsuitable for our approach, as the position of the speaker, the background, and the quality of the videos and the acoustic signal can vary significantly across different samples. We therefore built a new video dataset from YouTube, composed of videos uploaded to the platform by well-established users (commonly known as *youtubers*), who recorded themselves speaking in front of the camera in their personal home studios. Such videos are usually of high quality, with the faces of the subject featured in a prominent way and with notable expressiveness in both the speech and face. Our model, software and dataset are publicly released at: `https://imatge-upc.github.io/wav2pix/`.

B.2 Related Work

Generative Adversarial Networks (GANs) [188] are a state of the art deep generative model that consist of two networks, a Generator G and a Discriminator D, playing a min-max game against each other. This means both networks are optimized to fulfill their own objective: G has to generate realistic samples and D has to be good at rejecting G

samples and accepting real ones. This joint learning adversarial process lasts for as long as G begins generating samples which are as good enough as to fool D into making as many mistakes as possible. The way Generator can create novel data mimicking real one is by mapping samples $z \in \mathbb{R}^n$ of arbitrary dimensions coming from some simple prior distribution $\mathcal{Z}$ to samples x from the real data distribution $\mathcal{X}$ (in this case we work with images, so $\mathbf{x} \in \mathbb{R}^{w \times h \times c}$ where $w \times h$ are spatial dimensions width and height and c is the amount of channels).This means each $\mathbf{z}$ forward is like sampling from $\mathcal{X}$. On the other hand the discriminator is typically a binary classifier as it distinguishes *real* samples from *fake* ones generated by G. One can further condition G and D on a variable $e \in \mathbb{R}^k$ of arbitrary dimensions to derive the the conditional GANs [237] formulation, with the conditioning variable being of any type, *e.g.* a class label or text captions [238]. In our work, we generate images conditioned on raw speech waveforms.

Numerous improvements to the GANs methodology have been presented lately. Many focusing on stabilizing the training process and enhance the quality of the generated samples [239, 240]. Others aim to tackle the vanishing gradients problem due to the sigmoid activation and the log-loss in the end of the classifier [241–243]. To solve this, the least-squares GAN (LSGAN) approach [243] proposed to use a least-squares function with binary coding (1 for real, 0 for fake). We thus use this conditional GAN variant with the objective function is given by:

$$\begin{aligned} \min_D V_{\text{LSGAN}}(D) = &\frac{1}{2} e_{\mathbf{x},\mathbf{e} \sim p_{\text{data}}(\mathbf{x},\mathbf{e})}[(D(\mathbf{x},\mathbf{e}) - 1)^2] \\ &+ \frac{1}{2} e_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z}), \mathbf{e} \sim p_{\text{data}}(\mathbf{e})}[D(G(\mathbf{z},\mathbf{e}),\mathbf{e})^2]. \end{aligned} \tag{B.1}$$

$$\min_G V_{\text{LSGAN}}(G) = \frac{1}{2} e_{\mathbf{z} \sim p_{\mathbf{e}}(\mathbf{e}), \mathbf{y} \sim p_{\text{data}}(\mathbf{y})}[(D(G(\mathbf{z},\mathbf{e}),\mathbf{e}) - 1)^2], \tag{B.2}$$

Multi-modal generation. Data generation across modalities is becoming increasingly popular [238, 244–246]. Several works [238, 246] present different approaches for synthesizing realistic images given a text description. Recently, a number of approaches combining audio and vision have appeared, with tasks such as generating speech from a video [247] or generating images from audio/speech [248]. In this paper we will focus on the latter.

Most works on audio conditioned image generation adopt non end-to-end approaches and exploit previous knowledge about the data. Typically, speech has been encoded with handcrafted features, such us the MEL spectrum or Mel-frequency Cepstral Coefficients (MFCC), which have been very well engineered to represent human speech. At the visual part, point-based models of the face [249] or the lips [233] have been adopted. In contrast to that, our network is trained entirely end-to-end solely from raw speech to generate image pixels.

A direct synthesis of facial pixels was obtained in [232] with a discriminative model whose input were a pair of audio features and a visual example of the face to predict. In that case, the model had the help of a additional identity information (image of the speaker) to help in the prediction, so the network learned how to modify this input to match with the speech utterance. Following a similar architecture, a generative model trained with adversarial training was proposed in [250]. In this case, they introduced a temporal regularization to improve the smoothness of the output video sequence. Our work differs from theirs in that we use raw speech instead of hand-crafted features, and we do not need any image of the speaker as all identity information is extracted from the speech only.

B.3 The *Youtubers* Dataset

In this section we describe the multi-stage pipeline adopted to collect the new audio-visual dataset of human speech used in this work. We collected videos uploaded to YouTube by well-established users (so-called *youtubers*), who tend to record themselves speaking in front of the camera in a well controlled environment. Such videos are usually of high quality, with the faces of the subject featured in a prominent way and with notable expressiveness in both the speech and face. The Youtubers dataset is composed of two sets: the complete noisy dataset automatically generated, and a clean subset which was manually curated to obtain high quality data.

B.3.1 Data collection

In total we collected 168,796 seconds of speech with the corresponding video frames, and cropped faces from a list of 62 youtubers active during the past few years. The dataset is gender balanced and manually cleaned keeping 42,199 faces, each with an associated 1-second speech chunk. The pipeline used for downloading and pre-processing the full dataset is summarized in Figure B.1, and the key stages are discussed in the following paragraphs:

YouTubers selection and video download. A list of 62 different Spanish speaker *youtubers* was built, consisting on 29 males and 33 females from different ethnicity and accents. This list was chosen accordantly to their popularity and expressions in front of the camera. After having the final list, the last 15 videos uploaded to their channel were downloaded.

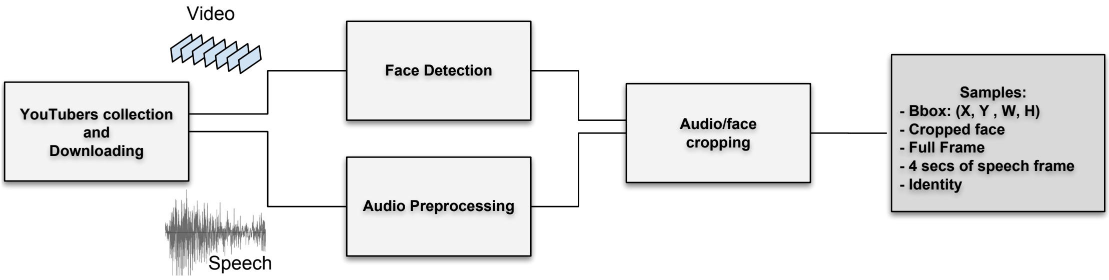


Figure B.1: **High level representation of the data collection pipeline.** Each detected face is associated with a 4 seconds length audio and the corresponding identity. Besides that we also kept the bounding box coordinates and the original image frame.

B.3.2 Data Pre-processing

Audio preprocessing. The audio was originally downloaded in Advance Audio Coding (AAC) format at 44100 Hz and stereo and converted to WAV, as well as re-sampled to 16 kHz with 16 bits per sample and converted to mono.

Face Detection. The faces were detected using a Haar Feature-based Classifier [251] trained with frontal face features. We prevent the method from having false positives by taking only the most confident detection for each frame.

Audio/faces cropping. From each detection it is saved the bounding box coordinates, an image of the cropped face in BGR format, the full frame and a 4 seconds length speech frame, which encompasses 2 seconds ahead and behind the given frame. Moreover, we keep an identity (name) for each sample. We apply a pre-emphasis step to each speech frame and normalize it between $[-1, 1]$.

As stated in section B.5 our model demonstrate a loss of performance when trained with noisy data. Thus, a part of the dataset was manually filtered to obtain the high-quality data required to improve the performance of our network. We took a subset of 10 identities, five female and five male, from our dataset and manually filtered them making sure that all faces were visually clear and all audios contain speech, so that all the silence and music parts were removed. As a result, the cleaned dataset contains a total of 4,860 images and audios (4 seconds length).

B.4 Method

Since our goal is to train a GAN conditioned on raw speech waveforms, our model is divided in three modules trained altogether end-to-end: a speech encoder, a generator network and a discriminator network described in the following paragraphs respectively.

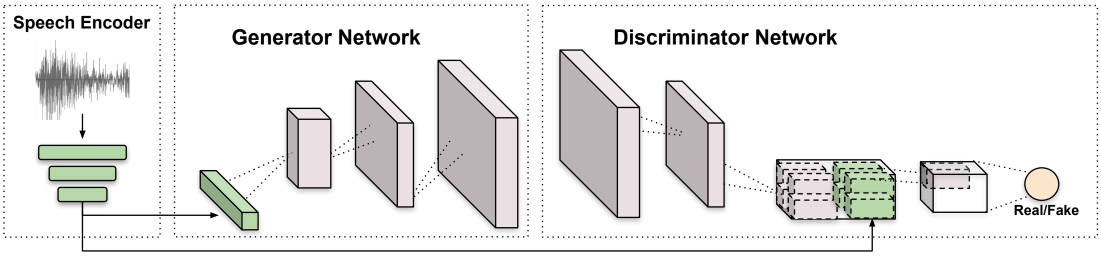


Figure B.2: **Overall diagram of our speech-conditioned face generation GAN architecture.** The network consists of a speech encoder, a generator and a discriminator network. An audio embedding (green) is used by both the generator and discriminator, but its error is just back-propagated at the generator. It is encoded and projected to a lower dimension (vector of size 128). Pink blocks represent convolutional/deconvolutional stages.

The speech encoder was adopted from the discriminator in [252], while both the image generator and discriminator architectures were inspired by [238]. The whole system was trained following a Least Squares GAN [243] scheme. Figure B.2 depicts the overall architecture.

B.4.1 Speech Encoder

As mentioned in Section B.2, many existing methods [232, 233, 250] require the extraction of handcrafted audio features before feeding the data into the neural network. This could limit the representation learning, as the audio information is extracted manually and not optimized for our generative task. In contrast, SEGAN [252] proposed a method for speech enhancement in which they do not work on the spectral domain, but at the waveform level. We coupled a modified version of the SEGAN discriminator Φ as input to an image generator G. Our speech encoder was modified to have 6 strided one-dimensional convolutional layers of kernel size 15, each one with stride 4 followed by LeakyReLU activations. Moreover we only require one input channel, so our input signal is $\mathbf{s} \in \mathbf{R}^{T\times 1}$, being $T = 16,384$ the amount of waveform samples we inject into the model (roughly one second of speech at $16\,kHz$). The aforementioned convolutional stack decimates this signal by a factor $4^6 = 4096$ while increasing the feature channels up to 1024. Thus, obtaining a tensor $f(\mathbf{s}) \in \mathbb{R}^{4\times 1024}$ in the output of the convolutional stack f. This is flattened and injected into three fully connected layers that reduce the final speech embedding dimensions from $1024 \times 4 = 4096$ to 128, obtaining the vector $\mathbf{e} = \Phi(\mathbf{s}) \in \mathbb{R}^{128}$.

B.4.2 Image Generator Network

We take the speech embedding $\mathbf{e}$ as input to generate images such that $\hat{\mathbf{x}} = G(\mathbf{e}) = G(\Phi(\mathbf{s}))$. The inference proceeds with two-dimensional transposed convolutions, where the input is a tensor $\mathbf{e} \in \mathbb{R}^{1\times1\times128}$ (an image of size 1×1 and 128 channels), based on [244]. The final interpolation can either be $64 \times 64 \times 3$ or $128 \times 128 \times 3$ just by playing with the amount of transposed convolutions (4 or 5). It is important to mention that we have no latent variable $\mathbf{z}$ in G inference as it did not give much variance in predictions in preliminary experiments. To enforce the generative capacity of G we followed a dropout strategy at inference time inspired by [253].

In preliminary experiments, we found it convenient to add a secondary component to the loss of G: a *softmax* classifier trained over the given speech embedding. This classifier helped the whole network into preserving the identity of the speaker. The magnitude of the classification component is controlled by a new hyper-parameter λ. Therefore, the G loss, follows the LSGAN loss presented in Equation B.2 with the addition of this weighted auxiliary loss for identity classification.

B.4.3 Image Discriminator Network

The Discriminator D is designed to process several layers of stride 2 convolution with a kernel size of 4 followed by a spectral normalization [254] and leakyReLU (apart from the last layer). When the spatial dimension of the discriminator is 4×4, we replicate the speech embedding $\mathbf{e}$ spatially and perform a depth concatenation. The last convolution is performed with stride 1 to obtain a D score as the output.

B.5 Experiments

B.5.1 Implementation details

The *Wav2Pix* model was trained on the cleaned dataset described in Section B.3 combined with a data augmentation strategy. In particular, we copied each image five times, pairing it with 5 different audio chunks of 1 second randomly sampled from the 4 seconds segment. Thus, we obtained $\approx$ 24k images and paired audio chunks of 1 second used for training our model. Our implementation is based on the PyTorch library [255] and trained on a GeForce Titan X GPU with 12GB memory. We kept the hyper-parameters as suggested in [238], changing the learning rate to 0.0001 in G and 0.0004 in D as suggested in [256]. We use ADAM solver [257] with momentum 0.1.

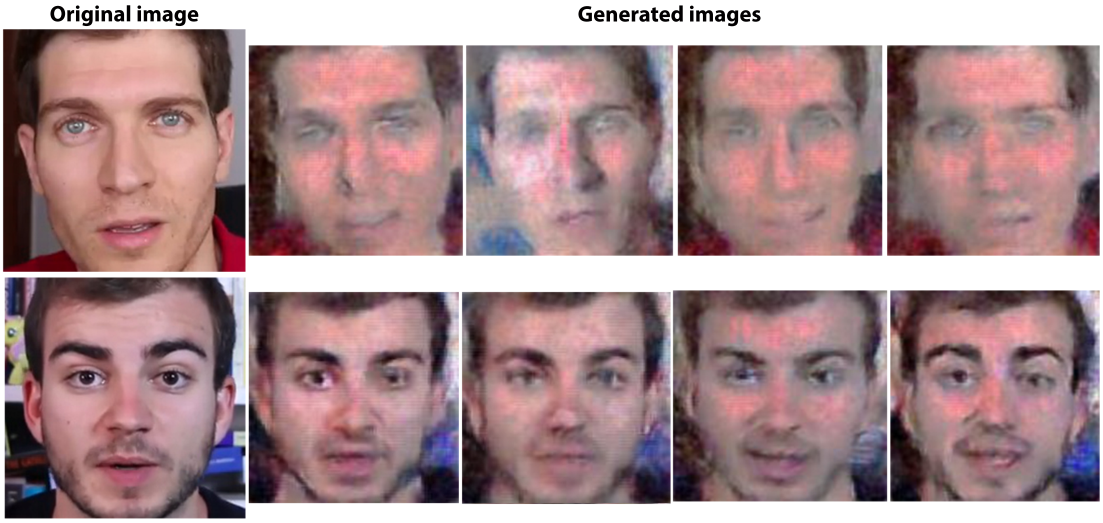


Figure B.3: **Qualitative results.** Examples of generated faces compared to the original image of the person who the voice belongs to. In the generated images, we can observe that our model is able to preserve the physical characteristics and produce different face expressions. We thank the Youtubers Javier Muñiz (top) and Jaime Altozano (bottom) for their authorization on using their image in this work.

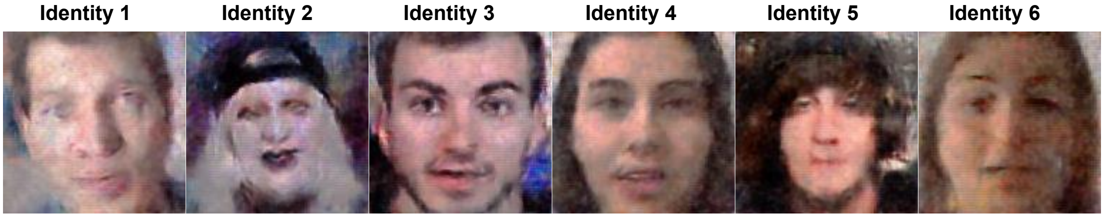


Figure B.4: **Samples of different identities generated by our model.** We condition our image generation to raw speech in order to generate the different faces associated to the given identities.

B.5.2 Qualitative results

Figure B.3 shows a set of generated images given a raw speech chunk, compared to the original image of the person who the voice belongs to. Different speech waveform produced by the same speaker were fed into the network to produce such images. Although the generated images are blurry, it is possible to observe that the model learns the person's physical characteristics, preserving the identity, and present different face expressions depending on the input speech [1]. Other examples from six different identities are presented in Figure B.4.

[1] Some examples of images and it correspondent speech as well as more generated images are available at: https://imatge-upc.github.io/wav2pix/

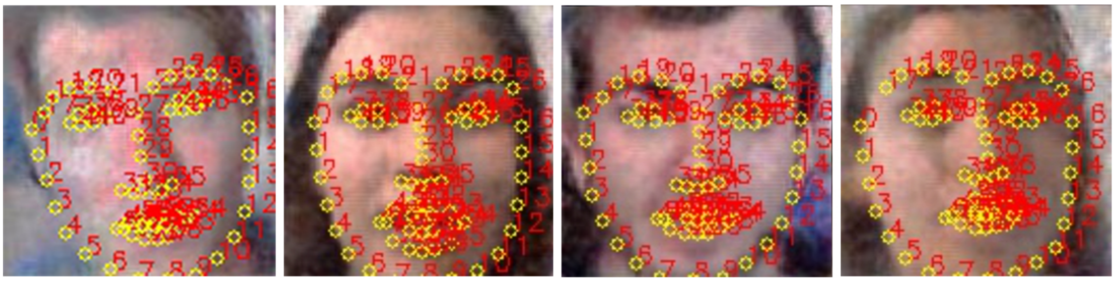


Figure B.5: **Examples of the 68 key-points detected on images generated by our model.** Yellow circles indicate facial landmarks fitted to the generated faces, numbered in red fonts.

B.5.3 Quantitative results

To quantify the model's accuracy regarding the identity preservation, we fine-tuned a pre-trained VGG-Face Descriptor network [258, 259] with our dataset. We predicted the speaker identity from the generated images of both the speech train and test partitions, obtaining an identification accuracy of 76.81% and 50.08%, respectively.

We also assessed the ability of the model to generate realistic faces, regardless of the true speaker identity. To have a more rigorous test than a simple Viola & Jones face detector [251], we measured the ability of an automatic algorithm [260] to correctly identify facial landmarks on images generated by our model. We define detection accuracy as the percentage of images where the algorithm is able to identify *all* 68 key-points. For the proposed model and all images generated for our test set, the detection accuracy is 90.25%, showing that in most cases the generated images retain the basic visual characteristics of a face. This detection rate is much higher than the identification accuracy of 50.08%, as in some cases the model confuses identities, or mixes some of them in a single face. Examples of detected faces together with their numbered facial landmarks can be seen in Figure B.5.

B.5.4 Ablation study

We also tried to generate faces with noisy speech, experiments that resulted in failures. Firstly, we used audio snippets from the same yotubers videos that presented background noise, silences or other people's voice. Secondly, using the well known VoxCeleb 1 dataset [235], which contains a larger amount of images and identities but present a lower audio quality. In both cases, the quality of results was very poor, making it almost impossible to recognize faces. These results show the importance of having clean speech samples to train the proposed model.

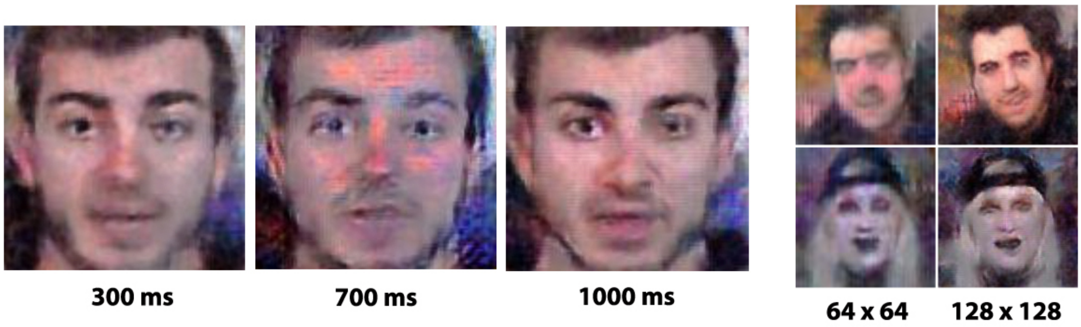


Figure B.6: **Effect of image resolution.** (Left) Images generated for three speech chunk lengths. (Right) Images generated at two spatial resolutions.

We also observed a drop in performance when working with smaller speech chunks and lower image definitions. We observed a visual degradation when using audio chunks of 300 and 700 milliseconds, which was reflected in a decrease of the face detection rate. Detection accuracy when using 300 and 700 ms chunks was 81.16% and 89.12%, respectively, in both cases worse than the 90.25% accuracy achieved when using 1000 ms chunks. Figure B.6 (left) shows examples of generated images for the three speech chunk lengths. Figure B.6 (right) shows how using a lower definition of 64x64 pixels instead of 128x128 results into blurrier images.

B.6 Ethical considerations

Although this work is a purely academic investigation, in this section we would like to explicitly discuss a set of ethical considerations that are important to be addressed due to the potential sensitivity of facial information and usage of this research.

Privacy. Although our method is trained to generate image of faces of a particular set of individuals, it cannot recover the true identity of a person from their voice (*i.e.* an exact image of their face). We designed our method with the intention of investigating how much of a physical appearance of a person can be recovered by given just their raw speech as input. We trained it to capture visual features (related to physical appearances and facial expressions) that are common to different individuals.

Voice-face correlations and dataset bias. As mentioned above, our model is designed to generate correlations that exist between facial features and voices of speakers in the training data. The training data used is a collection of personal video blogs from YouTube, and does not represent equally the entire world population. Therefore, the model– as is the case with any machine learning model– is affected by this uneven distribution of data. More specifically, if a set of speakers might have vocal-visual traits that are relatively uncommon in the data, then the quality of our reconstructions for such cases may degrade.

For the collection of our dataset, a set of video channels of popular spanish *Youtubers* were selected and videos that show a variety of facial expressions were used. Therefore, the image generation will perform better when similar data is used (*e.g.* Spanish speakers).

We recommend that any further investigation or practical use of this technology should be carefully tested to ensure that the training data is representative of the intended user population. If that is not the case, more representative data should be broadly collected. It is important to note that the current training data was used in order to test a hypothesis and conduct a academic investigation.

B.7 Final Remarks

In this Appendix we introduced a simple yet effective cross-modal approach for generating images of faces given only a short segment of speech, and proposed a novel generative adversarial network variant that is conditioned on the raw speech signal.

As high-quality training data are required for this task, we further collected and curated a new dataset, the Youtubers dataset, that contains high quality visual and speech signals. Our experimental validation demonstrates that the proposed approach is able to synthesize plausible facial images with an accuracy of 90.25%, while also being able to preserve the identity of the speaker about 50% of the times. Our ablation experiments further showed the sensitivity of the model to the spatial dimensions of the images, the duration of the speech chunks and, more importantly, on the quality of the training data. Further steps may address the generation of a sequence of video frames aligned with the conditioning speech, as well exploring the behaviour of the *Wav2Pix* when conditioned on unseen identities.

Bibliography

[1] Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. Including signed languages in natural language processing. In *Proceedings of the 11th International Joint Conference on Natural Language Processing*, pages 7347–7360, 2021.

[2] Lindsay Ferrara and Vibeke Bo. A pilot corpus of Norwegian Sign Language [video dataset]. *Norwegian University of Science and Technology, NTNU*, 2015.

[3] Lindsay Ferrara and Gabrielle Hodge. Language as description, indication, and depiction. *Frontiers in Psychology*, 9:716, 2018.

[4] Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. Aligning subtitles in sign language videos. 2021.

[5] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020.

[6] Hamid Reza Vaezi Joze and Oscar Koller. MS-ASL: A large-scale data set and benchmark for understanding American Sign Language. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019.

[7] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei Efros. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[8] Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Read and attend: Temporal localisation in sign language videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[10] World Health Organization 2021. Deafness and hearing loss. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, 2021. Accessed: 2021-03-10.

[11] James C Woodward. Implications for sociolinguistic research among the deaf. *Sign Language Studies*, pages 1–7, 1972.

[12] United Nations. We sign for human rights. https://www.un.org/en/observances/sign-languages-day, 2021. Accessed: 2022-02-16.

[13] John A Albertini, Marc Marschark, and Pamela J Kincheloe. Deaf students' reading and writing in college: Fluency, coherence, and comprehension. *Journal of Deaf Studies and Deaf Education*, 21(3):303–309, 2016.

[14] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, and Tessa Verhoef. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31, 2019.

[15] Amit Moryossef and Yoav Goldberg. Sign Language Processing. https://sign-language-processing.github.io/, 2021. Accessed: 2022-01-20.

[16] Danielle Bragg, Naomi Caselli, Julie A Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E Ladner. The FATE landscape of sign language AI datasets: An interdisciplinary perspective. In *Proceedings of the ACM Transactions on Accessible Computing (TACCESS)*, volume 14, pages 1–45. ACM New York, NY, USA, 2021.

[17] Raychelle Harris, Heidi M Holmes, and Donna M Mertens. Research ethics in sign language communities. *Sign Language Studies*, 9(2):104–131, 2009.

[18] Danielle Bragg, Oscar Koller, Naomi Caselli, and William Thies. Exploring collection of sign language datasets: Privacy, participation, and model performance. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–14, 2020.

[19] Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. Extending the public dgs corpus in size and depth. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 75–82, 2020.

[20] Amy Isard. Approaches to the anonymisation of sign language corpora. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 95–100, 2020.

[21] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. Articulated people detection and pose estimation: Reshaping the future. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3178–3185. IEEE, 2012.

[22] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1212–1221, 2017.

[23] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):172–186, 2019.

[24] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018.

[25] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[26] Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Muller, and Sarah Ebling. Evaluating the immediate applicability of pose estimation for sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3434–3440, 2021.

[27] Valerie Sutton. *Lessons in sign writing: Textbook*. SignWriting, 2014.

[28] Siegmund Prillwitz and Heiko Zienert. Hamburg notation system for sign language: Development of a sign writing with computer application. In *Proceedings of the 3rd European Congress on Sign Language Research - Current trends in European Sign Language Research*, pages 355–379, 1990.

[29] William C Stokoe Jr. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of Deaf Studies and Deaf Education*, 10(1):3–37, 2005.

[30] Johanna Mesch and Lars Wallin. Gloss annotations in the Swedish Sign Language corpus. *International Journal of Corpus Linguistics*, 20(1):102–120, 2015.

[31] Trevor Johnston and Louise De Beuzeville. Auslan corpus annotation guidelines. *Auslan Corpus*, 2016.

[32] Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. Public DGS corpus: Annotation conventions. Technical report, Project Note AP03-2018-01, DGS-Korpus project, IDGS, Hamburg University, 2018.

[33] Kayo Yin and Jesse Read. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, 2020.

[34] William C Stokoe Jr. Sign language structure: An outline of the visual communication systems of the American deaf. *Journal of Deaf Studies and Deaf Education*, 10(1):3–37, 2005.

[35] Paul Gary Dudis. *Depiction of events in ASL: Conceptual integration of temporal components*. PhD thesis, University of California, Berkeley, 2004.

[36] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2Sign: A large-scale multimodal dataset for continuous American Sign Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[37] Amanda Duarte. Cross-modal neural sign language translation. In *Proceedings of the 27th ACM International Conference on Multimedia (ACMMM) - Doctoral Symposium*, 2019.

[38] Amanda Duarte, Samuel Albanie, Xavier Giro-i Nieto, and Gül Varol. Sign language video retrieval with free-form textual queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[39] Didac Surís, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i Nieto. Cross-modal embeddings for video and audio retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[40] Amanda Duarte, Francisco Roldan, Miquel Tubau, Janna Escur, Santiago Pascual, Amaia Salvador, Eva Mohedano, Kevin McGuinness, Jordi Torres, and Xavier

Giro-i Nieto. WAV2PIX: Speech-conditioned face generation using generative adversarial networks. In *IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, 2019.

[41] Lucia Specia, Loic Barrault, Ozan Caglayan, Amanda Duarte, Desmond Elliott, Spandana Gella, Nils Holzenberger, Chiraag Lala, Sun Jae Lee, Jindrich Libovicky, et al. Grounded sequence to sequence transduction. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):577–591, 2020.

[42] Peter Muschick. Learn2Sign: Sign language recognition and translation using human keypoint estimation and transformer model. Master's thesis, Universitat Politecnica de Catalunya, 2020.

[43] Pol Pérez Granero. 2d to 3d body pose estimation for sign language with deep learning. Bachelor's thesis, Universitat Politecnica de Catalunya, 2020.

[44] Miquel Tubau. WAV2PIX: Enhancement and evaluation of a speech-conditioned image generator. Master's thesis, Universitat Politecnica de Catalunya, 2019.

[45] Sandra Roca. Block-based speech-to-speech translation. Bachelor's thesis, Universitat Politecnica de Catalunya, 2018.

[46] Janna Escur i Gelabert. Exploring automatic speech recognition with tensorflow. Bachelor's thesis, Universitat Politecnica de Catalunya, 2018.

[47] David F Armstrong, William C Stokoe, and Sherman E Wilcox. *Gesture and the nature of language*. Cambridge University Press, 1995.

[48] Rachel Sutton-Spence and Bencie Woll. *The linguistics of British Sign Language: an introduction*. Cambridge University Press, 1999.

[49] Roland Pfau, Josep Quer, et al. *Nonmanuals: their grammatical and prosodic roles*. Self-published, 2010. URL https://web.archive.org/web/20200709222141id_/https://www.cnlse.es/sites/default/files/Nonmanuals.their%20grammatical%20and%20prosodic%20roles.pdf.

[50] Susanne Mohr. *Mouth actions in sign languages: An empirical study of Irish Sign Language (Vol. 3)*. Walter de Gruyter, 2014.

[51] Richard A Tennant, Marianne Gluszak, and Marianne Gluszak Brown. *The American sign language handshape dictionary*. Gallaudet University Press, 1998.

[52] William C Stokoe, Dorothy C Casterline, and Carl G Croneberg. *A dictionary of American Sign Language on linguistic principles*. Linstok Press, 1976.

[53] Eeva Anita Elliott and Arthur M Jacobs. Facial expressions, emotions, and sign languages. *Frontiers in Psychology*, 4:115, 2013.

[54] Michele P. An intro to asl grammar rules. https://takelessons.com/live/american-sign-language/asl-grammar-rules, 2021. Accessed: 2022-03-20.

[55] Pamela Perniss. Use of sign space. *The Routledge Handbook of Theoretical and Experimental Sign Language Research*, 2021.

[56] Myriam Vermeerbergen, Jan Nijen Twilhaar, and Mieke Van Herreweghe. Variation between and within sign language of the Netherlands and Flemish sign language. *De Gruyter*, 3:680–699, 2013.

[57] Inge Zwitserlood. Classifiers. In *Sign language: An international handbook*. De Gruyter, 2012.

[58] Myriam Vermeerbergen. Past and current trends in sign language research. *Language & Communication*, 26(2):168–192, 2006.

[59] Ross E Mitchell, Travas A Young, Bellamie Bachelda, and Michael A Karchmer. How many people use ASL in the United States? Why estimates need updating. *Sign Language Studies*, 6(3):306–335, 2006.

[60] Benjamin J Bahan. *Non-manual realization of agreement in American Sign Language*. PhD thesis, Boston University, 1996.

[61] Nina Timmermans et al. *The status of sign languages in Europe*. Council of Europe, 2005.

[62] Carolyn McCaskill, Ceil Lucas, Robert Bayley, and Joseph Christopher Hill. *The hidden treasure of Black ASL: Its history and structure*. Gallaudet University Press Washington, DC, 2011.

[63] Carol Padden. The Deaf community and the culture of deaf people. *Sign language and the deaf community*, pages 89–103, 1980.

[64] Neil Stephen Glickman. *Deaf identity development: Construction and validation of a theoretical model*. PhD thesis, University of Massachusetts Amherst, 1993.

[65] H-Dirksen L Bauman and Joseph J Murray. *Deaf gain: Raising the stakes for human diversity*. U of Minnesota Press, 2014.

[66] Jon Henner and Octavian Robinson. Signs of oppression in the academy: The case of signed languages. In *Linguistic Discrimination in US Higher Education: Power, Prejudice, Impacts, and Remedies*, pages 92–109. Routledge, 2021.

[67] H-Dirksen L Bauman. Audism: Exploring the metaphysics of oppression. *Journal of deaf studies and deaf education*, 9(2):239–246, 2004.

[68] Richard Clark Eckert and Amy June Rowley. Audism: A theory and practice of audiocentric privilege. *Humanity & Society*, 37(2):101–130, 2013.

[69] Harlan Lane. A chronology of the oppression of sign language in France and the United States. *Recent perspectives on American Sign Language*, pages 119–161, 2017.

[70] Ann E Geers, Christine M Mitchell, Andrea Warner-Czyz, Nae-Yuh Wang, Laurie S Eisenberg, CDaCI Investigative Team, et al. Early sign language exposure and cochlear implantation benefits. *Pediatrics*, 140(1), 2017.

[71] Karolina Kozik. Without sign language, deaf people are not equal. https://www.hrw.org/news/2019/09/23/without-sign-language-deaf-people-are-not-equal, 2019. Accessed: 2022-02-22.

[72] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020.

[73] Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. The ASL-LEX 2.0 project: A database of lexical and phonological properties for 2,723 signs in american sign language. *The Journal of Deaf Studies and Deaf Education*, 26(2):263–277, 2021.

[74] Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. The American Sign Language lexicon video dataset. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop.*, pages 1–8. IEEE, 2008.

[75] Morteza Zahedi, Philippe Dreuw, David Rybach, Thomas Deselaers, and Hermann Ney. Continuous sign language recognition-approaches from speech recognition and available data resources. In *Workshop on Representation and Processing of Sign Languages*, 2006.

[76] Carol Neidle and Christian Vogler. A new web interface to facilitate access to corpora: Development of the ASLLRP data access interface (DAI). In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, 2012.

[77] Adam Schembri, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. Building the british sign language corpus. *Language Documentation & Conservation*, 7:136–154, 2013.

[78] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*, 2021.

[79] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

[80] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *Proceedings of the thirty-second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[81] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325, 2021.

[82] Ville Viitaniemi, Tommi Jantunen, Leena Savolainen, Matti Karppa, and Jorma Laaksonen. S-pot: A benchmark in spotting signs within continuous signing. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, 2014.

[83] Necati Cihan Camgoz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. Content4All Open research sign language translation datasets. *arXiv preprint arXiv:2105.02351*, 2021.

[84] U. Von Agris and K.-F. Kraiss. Signum database: Video corpus for signer-independent continuous sign language recognition. In *Workshop on Representation and Processing of Sign Languages*, pages 243–246, 2010.

[85] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 7784–7793, 2018.

[86] Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. A comprehensive study on sign language recognition methods. *arXiv preprint arXiv:2007.12530*, 2020.

[87] Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. Include: A large scale dataset for Indian Sign Language recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1366–1375, 2020.

[88] Laura Cristina Galea and Alan F Smeaton. Recognising irish sign language using electromyography. In *Proceedings of the International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–4. IEEE, 2019.

[89] Alfarabi Imashev, Medet Mukushev, Vadim Kimmelman, and Anara Sandygulova. K-RSL: a corpus for linguistic understanding, visual evaluation, and recognition of sign languages. In *Proceedings of the 24th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2020.

[90] Johanna Mesch and Lars Wallin. From meaning to signs and back: Lexicography and the Swedish Sign Language corpus. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 123–126, 2012.

[91] Ozge Mercanoglu Sincan and Hacer Yalim Keles. AUTSL: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 2020.

[92] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015.

[93] International Standards Organization (ISO). Interpreting services — general requirements and recommendations, 2018. URL https://www.iso.org/obp/ui/iso:std:iso:18841:ed-1:v1:en.

[94] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Progressive transformers for end-to-end sign language production. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[95] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908, 2020.

[96] Jan Zelinka, Jakub Kanis, and Petr Salajka. Nn-based czech sign language synthesis. In *International Conference on Speech and Computer*, pages 559–568. Springer, 2019.

[97] Jan Zelinka and Jakub Kanis. Neural sign language synthesis: Words are our glosses. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3395–3403, 2020.

[98] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13), 2019.

[99] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of 32nd Conference on Neural Information Processing Systems Wokshops*, 2018.

[100] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015.

[101] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Adversarial training for multi-channel sign language production. In *Proceeding of the British Machine Vision Virtual Conference (BMVC)*, 2020.

[102] Onno Crasborn and Han Sloetjes. Enhanced ELAN functionality for sign language corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 39–43, 2008.

[103] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[104] Liliane Momeni, Gül Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Watch, read and lookup: learning to spot signs from multiple supervisors. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020.

[105] Judy Reilly and Marina L McIntire. American Sign Language and Pidgin Sign English: What's the difference? *Sign Language Studies*, pages 151–192, 1980.

[106] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10023–10033, 2020.

[107] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Progressive transformers for end-to-end sign language production. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 687–705. Springer, 2020.

[108] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020.

[109] Gary J Grimes. Digital data entry glove interface device, 1983. US Patent 4,414,537.

[110] Thad Starner and Alex Pentland. Real-time American Sign Language recognition from video using Hidden Markov Models. In *Motion-based recognition*, pages 227–243. Springer, 1997.

[111] Thad E Starner. Visual recognition of american sign language using hidden markov models. Technical report, Massachusetts Inst Of Tech Cambridge Dept Of Brain And Cognitive Sciences, 1995.

[112] Holger Fillbrandt, Suat Akyol, and K-F Kraiss. Extraction of 3d hand shape and posture from image sequences for sign language recognition. In *Proceedings of the IEEE International SOI Conference (Cat. No.03CH37443)*, pages 181–186. IEEE, 2003.

[113] Ali Farhadi, David Forsyth, and Ryan White. Transfer learning in sign language. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[114] Patrick Buehler, Andrew Zisserman, and Mark Everingham. Learning sign language by watching TV (using weakly aligned subtitles). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2961–2968, 2009.

[115] Helen Cooper, Nicolas Pugeault, and Richard Bowden. Reading the signs: A video based sign dictionary. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 914–919. IEEE, 2011.

[116] Eng-Jon Ong, Helen Cooper, Nicolas Pugeault, and Richard Bowden. Sign language recognition using sequential pattern trees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2200–2207. IEEE, 2012.

[117] Tan Dat Nguyen and Surendra Ranganath. Tracking facial features under occlusions and recognizing facial expressions in sign language. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–7. IEEE, 2008.

[118] Epameinondas Antonakos, Anastasios Roussos, and Stefanos Zafeiriou. A survey on mouth modeling and analysis for sign language recognition. In *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–7. IEEE, 2015.

[119] Oscar Koller, Hermann Ney, and Richard Bowden. Read my lips: Continuous signer independent weakly supervised viseme recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 281–296. Springer, 2014.

[120] Oscar Koller, Hermann Ney, and Richard Bowden. Deep learning of mouth shapes for sign language. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 85–91, 2015.

[121] Ulrich Von Agris, Jörg Zieren, Ulrich Canzler, Britta Bauer, and Karl-Friedrich Kraiss. Recent developments in visual sign language recognition. *Universal Access in the Information Society*, 6(4):323–362, 2008.

[122] Ali Farhadi and David Forsyth. Aligning asl for statistical translation using a discriminative word model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1471–1476. IEEE, 2006.

[123] Jens Forster, Christian Oberdörfer, Oscar Koller, and Hermann Ney. Modality combination techniques for continuous sign language recognition. In *Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis*, pages 89–99. Springer, 2013.

[124] Yuancheng Ye, Yingli Tian, Matt Huenerfauth, and Jingya Liu. Recognizing american sign language gestures from within continuous videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2064–2073, 2018.

[125] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. Sign language recognition using 3d convolutional neural networks. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2015.

[126] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13009–13016, 2020.

[127] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3056–3065, 2017.

[128] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.

[129] Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Read and attend: Temporal localisation in sign language videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16857–16866, 2021.

[130] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Andrew Brown, Chuhan Zhang, Ernesto Coto, Necati Cihan Camgöz, Ben Saunders, Abhishek Dutta, et al. Seehear: Signer diarisation and a new dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2280–2284. IEEE, 2021.

[131] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. Transferring cross-domain knowledge for video sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6205–6214, 2020.

[132] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015.

[133] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021.

[134] Eng-Jon Ong, Oscar Koller, Nicolas Pugeault, and Richard Bowden. Sign spotting using hierarchical sequential patterns with temporal intervals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1923–1930, 2014.

[135] Helen Cooper and Richard Bowden. Learning signs from subtitles: A weakly supervised approach to sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2568–2574. IEEE, 2009.

[136] Daniel Kelly, John Mc Donald, and Charles Markham. Weakly supervised training of a sign language recognition system using multiple instance learning density matrices. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(2):526–541, 2010.

[137] Tomas Pfister, James Charles, and Andrew Zisserman. Large-scale learning of sign language by watching TV (using co-occurrences). In *Proceedings of the British Machine Vision Conference (BMVC)*, 2013.

[138] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. Transferring cross-domain knowledge for video sign language recognition. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6204–6213, 2020.

[139] Themos Stafylakis and Georgios Tzimiropoulos. Zero-shot keyword spotting for visual speech recognition in-the-wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[140] Liliane Momeni, Triantafyllos Afouras, Themos Stafylakis, Samuel Albanie, and Andrew Zisserman. Seeing wake words: Audio-visual keyword spotting. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

[141] Emma Flint, Elliot Ford, Olivia Thomas, Andrew Caines, and Paula Buttery. A text normalisation system for non-standard English words. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 107–115, 2017.

[142] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.

[143] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 577–585, 2015.

[144] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964, 2016.

[145] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*, 2020.

[146] Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. Self-training and pre-training are complementary for speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3030–3034, 2021.

[147] Lin-shan Lee, James Glass, Hung-yi Lee, and Chun-an Chan. Spoken content retrieval—beyond cascading speech recognition with text retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9), 2015.

[148] Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, 33:12034–12045, 2020.

[149] Tao Jiang, Necati Cihan Camgoz, and Richard Bowden. Looking for the signs: Identifying isolated sign instances in continuous video footage. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2021.

[150] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.

[151] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 2630–2640, 2019.

[152] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2021.

[153] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[154] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12046–12055, 2019.

[155] Andrea Frome, Greg S Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeViSE: a deep visual-semantic embedding model. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pages 2121–2129, 2013.

[156] Niluthpol Chowdhury Mithun, Juncheng Billy Li, Florian Metze, and Amit K. Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018.

[157] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018.

[158] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.

[159] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019.

[160] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 450–459, 2019.

[161] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11592–11601, 2019.

[162] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[163] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer, 2020.

[164] Hazel Doughty, Ivan Laptev, Walterio Mayol-Cuevas, and Dima Damen. Action modifiers: Learning from adverbs in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 868–878, 2020.

[165] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021.

[166] Vassilis Athitsos, C. Neidle, Stan Sclaroff, Joan P. Nash, Alexandra Stefan, Ashwin Thangali, Haijing Wang, and Quan Yuan. Large lexicon project: American Sign Language video corpus and sign language indexing and retrieval algorithms. In *Proceeding of the Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT)*, 2010.

[167] Hee-Deok Yang, Stan Sclaroff, and Seong-Whan Lee. Sign language spotting with a threshold model based on conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31(7):1264–1277, 2008.

[168] Shilin Zhang and Bo Zhang. Using revised string edit distance to sign language video retrieval. In *Proceedings of the 2nd International Conference on Computational Intelligence and Natural Computing*, volume 1, pages 45–49, 2010.

[169] François Lefebvre-Albaret and Patrice Dalle. Video retrieval in sign language videos: How to model and compare signs? In *Proceeding of the 7th International Conference on Language Resources and Evaluation (LREC)*, 2010.

[170] Shilin Zhang and Bo Zhang. Using HMM to sign language video retrieval. In *Proceedings of the 2nd International Conference on Computational Intelligence and Natural Computing*, volume 1, pages 55–59, 2010.

[171] Nazif Can Tamer and Murat Saraçlar. Keyword search for sign language. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8184–8188, 2020.

[172] Nazif Can Tamer and Murat Saraçlar. Cross-lingual keyword search for sign language. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, 2020.

[173] Nazif Can Tamer and Murat Saraçlar. Improving keyword search performance in sign language with hand shape features. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2020.

[174] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Word2visualvec: Image and video to sentence matching by visual feature prediction. *arXiv preprint arXiv:1604.06838*, 2016.

[175] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and A. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, pages 207–218, 2014.

[176] Relja Arandjelović, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40:1437–1451, 2018.

[177] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *pre-print*, 2018. URL `https:`

//s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

[178] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 19–27, 2015.

[179] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[180] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[181] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *ArXiv*, abs/1909.11942, 2020.

[182] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[183] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

[184] Andrea Burns, Reuben Tan, Kate Saenko, Stan Sclaroff, and Bryan A. Plummer. Language Features Matter: Effective language representations for vision-language tasks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[185] Branden Chan, Stefan Schweter, and Timo Möller. German's next language model. *ArXiv*, abs/2010.10906, 2020.

[186] Pedro Ortiz Suarez, Benoît Sagot, and Laurent Romary. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, 2019.

[187] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[188] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[189] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*, 2020.

[190] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara Berg. Dance dance generation: Motion transfer for internet videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR) Workshops*, pages 0–0, 2019.

[191] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

[192] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8639–8648, 2018.

[193] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3560–3569. PMLR, 2017.

[194] Jian Ren, Menglei Chai, Sergey Tulyakov, Chen Fang, Xiaohui Shen, and Jianchao Yang. Human motion transfer from poses in the wild. In *European Conference on Computer Vision*, pages 262–279. Springer, 2020.

[195] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8340–8348, 2018.

[196] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 99–108, 2018.

[197] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in Neural Information Processing Systems*, 30, 2017.

[198] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3408–3416, 2018.

[199] Zhichao Huang, Xintong Han, Jia Xu, and Tong Zhang. Few-shot human motion transfer by personalized geometry and texture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2297–2306, 2021.

[200] Dongxu Wei, Xiaowei Xu, Haibin Shen, and Kejie Huang. Gac-GAN: A general method for appearance-controllable human video motion transfer. *IEEE Transactions on Multimedia*, 23:2457–2470, 2020.

[201] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[202] Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. Self-supervised dance video synthesis conditioned on music. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 46–54, 2020.

[203] Joao P Ferreira, Thiago M Coutinho, Thiago L Gomes, José F Neto, Rafael Azevedo, Renato Martins, and Erickson R Nascimento. Learning to dance: A graph convolutional adversarial network to generate realistic dance motions from audio. *Computers & Graphics*, 94:11–21, 2021.

[204] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[205] Qinkun Xiao, Minying Qin, and Yuting Yin. Skeleton-based chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural Networks*, 125:41–55, 2020.

[206] Jan Zelinka and Jakub Kanis. Neural sign language synthesis: Words are our glosses. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3395–3403, 2020.

[207] Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. Tessa, a system to aid communication with deaf people. In *Proceedings of the 5th International ACM Conference on Assistive Technologies*, pages 205–212, 2002.

[208] Kostasand George Caridakis Karpouzis, S-E. Fotinea, and Eleni Efthimiou. Educational resources and implementation of a Greek Sign Language synthesis architecture. In *Proceedings of the Computers and Education 49*, pages 54–74, 2007.

[209] John McDonald, Rosalee Wolfe, Jerry Schnepp, Julie Hochgesang, Diana Gorman Jamrozik, Marie Stumbo, Larwan Berke, Melissa Bialek, and Farah Thomas. An automated technique for real-time production of lifelike animations of American Sign Language. In *Proceedings of the Universal Access in the Information Society 15*, pages 551–566, 2016.

[210] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[211] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35:2878–90, 12 2013.

[212] Michael Erard. Why sign-language gloves don't help deaf people. https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/, 2017. Accessed: 2022-02-28.

[213] Eric Brochu, Nando De Freitas, and Kejie Bao. The sound of an album cover: Probabilistic multimedia and information retrieval. In *Proceedings of the Artificial Intelligence and Statistics (AISTATS)*, 2003.

[214] Rudolf Mayer. Analysing the similarity of album art with self-organising maps. In *Proceedings of the International Workshop on Self-Organizing Maps*, pages 357–366. Springer, 2011.

[215] Janis Libeks and Douglas Turnbull. You can judge an artist by an album cover: Using images for music annotation. *IEEE Multimedia*, 18:30–37, 2011.

[216] Jiansong Chao, Haofen Wang, Wenlei Zhou, Weinan Zhang, and Yong Yu. Tunesensor: A semantic-driven music recommendation service for digital photo albums. In *Proceedings of the 10th International Semantic Web Conference*, 2011.

[217] Alexander Schindler and Andreas Rauber. An audio-visual approach to music genre classification through affective color features. In *Proceedings of the European Conference on Information Retrieval*, pages 61–67. Springer, 2015.

[218] Xixuan Wu, Yu Qiao, Xiaogang Wang, and Xiaoou Tang. Bridging music and image via cross-modal ranking analysis. *IEEE Transactions on Multimedia*, 18(7): 1305–1318, 2016.

[219] Esra Acar, Frank Hopfgartner, and Sahin Albayrak. Understanding affective content of music videos through learned representations. In *Proceedings of the International Conference on Multimedia Modeling*, pages 303–314. Springer, 2014.

[220] Olivier Gillet, Slim Essid, and Gal Richard. On the correlation of automatic audio and visual segmentations of music videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):347–355, 2007.

[221] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K Sethi. Multimedia content processing through cross-modal association. In *Proceedings of the 11th ACM International Conference on Multimedia*, pages 604–611. ACM, 2003.

[222] Hong Zhang, Yueting Zhuang, and Fei Wu. Cross-modal correlation learning for clustering on image-audio dataset. In *Proceedings of the 15th ACM International Conference on Multimedia*, pages 273–276. ACM, 2007.

[223] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 689–696, 2011.

[224] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. *CoRR*, abs/1511.06078, 2015. URL http://arxiv.org/abs/1511.06078.

[225] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

[226] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[227] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

[228] Sungeun Hong, Woobin Im, and Hyun S Yang. Deep learning for content-based, cross-modal retrieval of videos and music. *arXiv preprint arXiv:1704.06761*, 2017.

[229] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. *CoRR*, abs/1609.08675, 2016. URL http://arxiv.org/abs/1609.08675.

[230] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN architectures for large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017.

[231] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[232] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[233] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95, 2017.

[234] Joon Son Chung, Andrew W Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (CVPR)*, pages 3444–3453, 2017.

[235] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: a large-scale speaker identification dataset. In *Interspeech*, 2017.

[236] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In *Interspeech*, 2018.

[237] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[238] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

[239] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2018.

[240] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

[241] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[242] David Berthelot, Thomas Schumm, and Luke Metz. BEGAN: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[243] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2813–2821. IEEE, 2017.

[244] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016.

[245] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *International Conference on Machine Learning (ICML)*, 2017.

[246] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

[247] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Improved speech reconstruction from silent video. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

[248] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357. ACM, 2017.

[249] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):94, 2017.

[250] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[251] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, volume 1, 2001.

[252] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. Segan: Speech enhancement generative adversarial network. *Interspeech*, pages 3642–3646, 2017.

[253] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (CVPR)*, pages 1125–1134, 2017.

[254] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[255] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

[256] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

[257] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[258] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[259] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74, 2018.

[260] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1867–1874, 2014.