Human Attention Target Estimation and Application on Images

Nora Horanyi

University of Birmingham

The School of Computer Science

This dissertation is submitted for the degree of
Doctor of Philosophy

Birmingham **January, 2023**

This thesis is dedicated to my loving father, who encouraged me to continue my studies and pursue a PhD degree, as well as to my mother, sister and fiancé, who have accompanied me on this journey and provided me with their constant love and support.

Ezt a dolgozatot szerető édesapámnak ajánlom, aki ösztönzött engem, hogy folytassam a tanulmányaimat és elérjem a PhD fokozatomat. Valamint anyukámnak, húgomnak és vőlegényemnek, akik végig mellettem álltak ezen az úton és állandó szeretetükkel és támogatásukkal segítettek.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others except as specified in the text and Acknowledgements. This thesis contains fewer than 65,000 words, including appendices, a bibliography, footnotes, tables and equations, and has fewer than 150 figures.

May 2023

Acknowledgements

Pursuing this PhD degree has been a remarkable experience that has greatly impacted my life, and I am deeply grateful for the support and guidance provided to me by many people who made this accomplishment possible.

I express my gratitude for the "School of Computer Science Studentship" funding that I received towards my PhD from the University of Birmingham’s School of Computer Science. This research was supported by the Institute of Information and communications Technology Planning, and evaluation (IITP) grants funded by the Korean government (MSIT) (2021-0-00537, Visual common sense through self-supervised learning for restoration of invisible parts in images; 2021-0-00034, Clustering technologies of fragmented data for time-based data analysis). This work was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant “Deep Visual Geometry Machines” (RGPIN-2018-03788) and by systems supplied by Compute Canada. I also acknowledge MoD/Dstl and EPSRC for providing the grant to support the UK academics involvement in a Department of Defense funded MURI project through EPSRC grant EP/N019415/1, and IITP grant funded by the Korean government (MSIT) (IITP-2020-0-01789, ITRC support program). We thank NVIDIA for sponsoring one GeForce Titan Xp GPU.

I extend my sincerest gratitude to my supervisors, Dr Hyung Jin Chang and Professor Ales Leonardis, for their unwavering support throughout my research journey. Their guidance has been invaluable in helping me to complete this thesis. I would also like to thank the members of my thesis committee Professor Iain Styles and Professor Per Kristian Lehre, for their valuable comments and suggestions on my research.

I am particularly grateful to my colleague, Linfang Zheng, for her constant support and for being a source of strength throughout my PhD journey. I also want to thank my colleagues in the Intelligent Robotics Laboratory and my office mates from 144 for the stimulating discussions and enjoyable times we shared during the past years.

Words cannot express my gratitude towards my family for their endless love and support throughout my PhD journey. My mother, father, and little sister have been my constant source of motivation; their belief in me and my dreams has been the driving force behind my success. I can’t thank them enough for their constant love, patience, and always being there for me. I am blessed to have them in my life; this achievement would not have been possible without them.

I am eternally grateful to my fiancé for his constant support and encouragement throughout this journey. His patience and understanding have been truly invaluable, and I cannot thank him enough for being there for me every step of the way. His belief in me has been my compass, and his encouragement has been my motivation. This achievement is as much his as it is mine, and I will always be grateful for his role in making it possible.

I want to express my sincere appreciation to my best friend for being by my side and keeping our friendship alive through the years despite the distance. Thank you for always making time for our conversations and visits.

With the most special and sincere thanks to my only angel who protects our family and always directs my steps from above.

"It is only with the heart that one can see rightly.
What is essential is invisible to the eye."
Antoin de St Exupery

Köszönetnyilvánítás

Szeretném kifejezni őszinte hálámat Dr. Hyung Jin Chang-nak es Prof. Ales Leonardis-nak, a folyamatos támogatásukért a kutatásom során. Az ő iránymutatásuk elengedhetetlen volt ahhoz, hogy ez a tézis létrejöjjön. Továbbá szeretnék köszönetet mondani a tézis csoportom bizottsági tagjainak, Professor Iain Styles- nak és Professor Per Kristian Lehre-nek, hogy értékes javaslataikkal és megjegyzéseikkel gyarapították a kutatásomhoz.

Különösen hálás vagyok kollégámnak, Linfang Zheng-nek, hogy mindenben támogatott a PhD munkám során. Valamint a kollégáimnak az Intelligens Robotika Laboratóriumban, a 144-es irodámban lévő kollégáimnak a stimuláló beszélgetésekért.

Szavakkal nem tudom kifejezni hálámat a családomnak támogatásukért a PhD tanulmányom során. Anyukám, apukám és kis húgom voltak a motivációm forrásai; a belém és álmaimba vetett hitük volt a sikerem hajtóereje. Nem tudok elég hálás lenni nekik végtelen szeretetükért, türelmükért és azért, hogy mindig mellettem voltak. Szerencsésnek mondhatom magam, hogy ők a családom, mindez nem lehetett volna lehetséges nélkülük.

Örökké hálás leszek a vőlegényemnek a folyamatos támogatásáért és biztatásáért ezen a rögös úton. Türelme és megértése nélkülözhetetlen volt, és nem tudom eléggé megköszönni neki, hogy mindenben mellettem állt. Hite volt az én iránytűm, és biztatása a motivációm. Ez a siker épp annyira az ő érdeme, mint az enyém, és mindig hálás leszek azért, hogy támogatásával ezt lehetővé tette.

Szeretném kifejezni az őszinte hálámat a legjobb barátnőmnek, hogy mellettem volt és megőrizte a barátságunkat az évek alatt, a távolság ellenére is. Köszönöm, hogy mindig talált időt a beszélgetéseinkre és találkozásainkra.

Külön és őszinte köszönettel az egyetlen angyalomnak, aki fentről óvja a családunkat és mindig irányt mutat nekem.

"Jól csak a szívével lát az ember.
Ami igazán lényeges, az a szemnek láthatatlan."
Antoin de St Exupery

Abstract

Most computer vision applications, such as automatic image cropping and attention target estimation, aim to perform or solve a task as humans would. While recent works using Neural Networks showed promising results in numerous research areas, complex and subjective tasks are still challenging to solve by only deriving information from images and videos. Therefore to enhance the ability of the machine to localise a part of an image or to interpret complex social interactions between multiple people in the scene like humans would, explicit or implicit user input could be integrated into the algorithm. This thesis investigates the usefulness of explicit verbal and implicit non-verbal human social clues and their combination in frameworks designed for attention-based computer vision tasks. The proposed computational methods in this thesis aim to better understand the user's intention through different input modalities. Specifically, this work used natural language and its combination with eye-tracking user inputs for description-based image cropping and visual attention for joint attention target estimation.

This work studied how a natural language expression of the users could be directly used to automatically localise the described part of an image and output an aesthetically pleasing image crop. The proposed solution re-purposed existing deep learning models into a single optimisation framework to solve this complex, highly subjective problem. In addition to the explicit language expressions and a semi-direct social clue, the eye movements of the users were integrated into a novel multi-modal framework. Finally, motivated by the usefulness of the user's semi-direct attention input, a deep neural network was developed for estimating attention targets in images to detect and follow the joint attention target of the subjects within the scene.

The presented approaches have achieved state-of-the-art performances in quantitative and qualitative measures on different benchmark datasets in their respective research areas. Furthermore, the conducted studies confirmed that the users favoured the output of the proposed solutions. These findings prove that integrating explicit or implicit user input and their combination into computational methods can produce more human-like outputs.

Keywords: multi-modal framework, computer vision applications, description-based image cropping, eye-tracking, joint attention target estimation

Table of contents

List of figures

List of tables

List of Acronyms

AI Artificial intelligence

AP Average precision

AUC Area under the curve

CAGIC Caption and aesthetic-guided image cropping

CN Caption network

Conv-LSTM Convolutional Long Short-Term Memory network

FOV Field-of-view

G-DAIC Gaze initialised, description and aesthetics-based image cropping

GIS Gaze inside score

GRU Gated Recurrent Unit

IoU Intersection over Union

JAT Joint attention target

LAEO Looking At Each Other

L-BFGS Limited-memory Broyden-Fletcher-Goldfarb-Shann

MR Multiple restart

MS Multi-scale bilinear sampling

NL Non-local

NWI Normalized word index

ROC Receiver operating characteristic

SA Scale anneal

SED Structured Edge Detector

SAT Single attention target

SNS Social networking services

TCN Temporal Convolutional Network

VAT VideoAttentionTarget

VCA VideoCoAttention

Chapter 1

Introduction

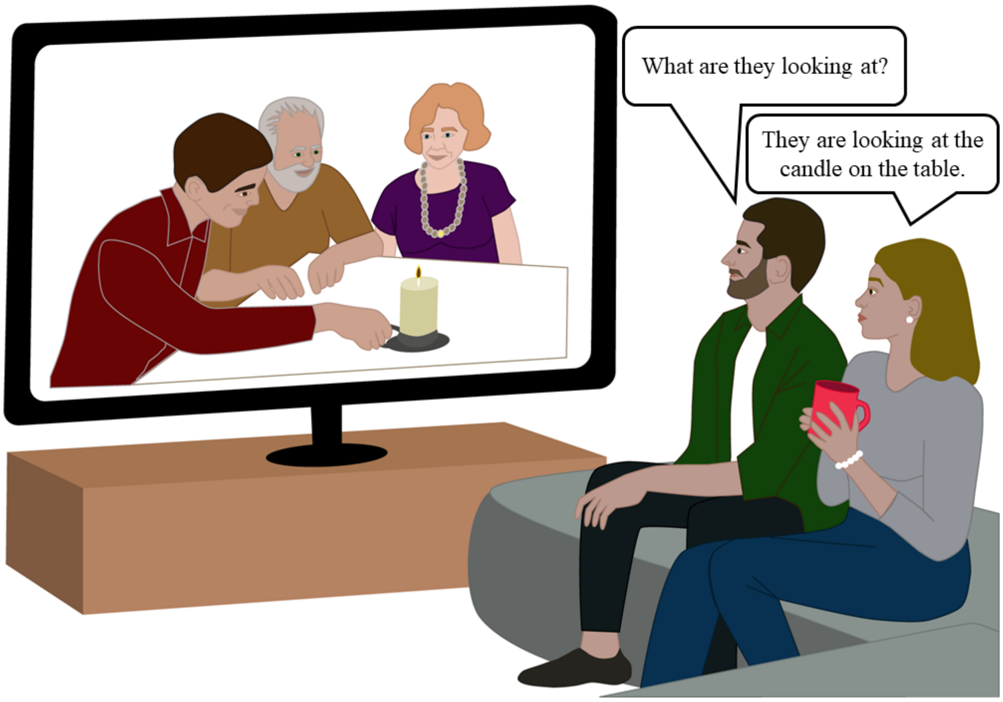

What are they looking at?
They are looking at the candle on the table.

Figure 1.1: **Illustration of the real-world applications associated with this thesis.** Example visualisation of the tasks of the description-based image part localisation and joint attention target estimation of multiple subjects from a third-person view.

1.1 Introduction

Human-centred artificial intelligence (AI) is an expanding field of computer science that leverages the insights of data-driven predictions while placing human users at the centre of the model design [1, 2]. The term AI encompasses the development of computational models that mimic human abilities and perform tasks that require human intelligence [3]. This research area focuses on complex, meaningful interactions and mutual understanding between humans and computers. It aims to advance the understanding of neural networks and machine learning that form the adaptive mechanisms of AI through mathematical and technological advancement [4].

An emerging interdisciplinary scientific field of AI is computer vision which aims to deduce useful information from digital images, videos, and other visual inputs [5]. It attempts to gain a high-level understanding of various visual inputs to automate tasks people can perform using their visual system. The recent advancement of biologically inspired deep learning models played a crucial role in the rapid development of this research field [6, 7] through its capability to extract high-level features from the input data. In recent years, new network designs with different learning strategies have been introduced to generate models that can perform similarly to humans or even better in different computer vision applications [8].

Although deep learning methods achieved outstanding results in various use cases, they have several limitations when analysing complex, high-dimensional, and noise-contaminated data sets [7]. In addition, in the case of complex and highly subjective tasks with more than one possible solution, where even humans would not fully agree on a solution, the models often fail to achieve the desired output. This is not surprising as many deep learning models are trained using the supervised learning paradigm through human annotations [9]. The complex nature of human intuition is hard to describe and learn, so the annotations used for training limit the model's output quality. Besides, humans rely on more than just their visual system during daily-life decision-making [10, 11]. To achieve human-like performance, relying solemnly on image or video data to perform a complex task might not be sufficient. Therefore, receiving additional input or using other modalities, such as human input, is potentially useful.

In human-centred computer vision, some of the most studied and utilised human inputs and behaviours are natural language expressions, facial expressions, eye movements, hand gestures, and body gestures [12–15]. Based on the source of the input information, user inputs can be categorised as implicit or explicit. We refer to user input as explicit or direct when there is no intermediary, meaning the user input is intended to be transferred to another person or machine. Typically generating this type of input requires effort from the user. For example, the user could generate explicit language expressions to describe a part of an image or something in their surroundings, and they could use hand gestures coupled with the speech to express their intention. Assistive technologies often use explicit inputs where the users can control the machine using direct user input. For example, voice-controlled smart home technology allows people to control different devices, such as hubs like Amazon Echo, which perform various actions directly or through connected devices following the user's voice commands. While this technology is primarily designed to bring comfort and improve productivity to its user, other voice-controlled hands-free assistant devices can improve the life quality of people with disabilities [16, 17]. The related research fields benefiting from explicit user inputs which enable these technologies include speech recognition, body and pose estimation and action recognition. Instead of direct input, implicit user input could be collected without the user's knowledge and producing this type of input should not take any extra effort from the user. A use case example for this is tracking applications aiming to identify the person's interest, where the user is observed through, *e.g.* a CCTV camera while performing everyday activities such as shopping [18]. Monitoring the shopping behaviour of the subject includes tracking their path and their interactions with products, including grasping and gazing. Based on this information, the proposed algorithms, such as [18–21], attempt to understand the customer's interest, intention and appreciation towards a given product. The desired outcome of such models is to assess the probability of the product being bought by a given customer and, if applicable, detect their reasons for not choosing the product. These algorithms require implicit user input to perform motion detection, attention target estimation and customer behaviour analysis.

One of our primary communication sources is speech. Using explicit language expressions to convey our thoughts to the people around us is the easiest and fastest way. Language expressions can be categorised into four sentence pattern categories: *statements*, *questions*, *commands*, and *exclamations*. The aforementioned smart home devices and products using speech rely on commands. These expressions are often pre-defined (*e.g.* Amazon Echo's wake word 'Alexa') and typically short sentences. These explicit voice commands require extra attention from the user to use the expression the device can understand. In our everyday life, it is more common to use statements in social interactions. For example, when answering the question "What do you like about this painting?" humans typically use full sentences to describe their interest, like "I like the reflection of the sun's rays on the water's surface." instead of "The reflection." or when describing where they placed something, they would say "I left the car key on the counter in the hall next to the vase." instead of "The hall.". These more complex statements are harder to understand and utilise by machines but require less effort from the users.

Besides natural language expressions, humans tend to rely heavily on their vision. Visual attention is a fundamental explicit social clue, and it provides us with much information regarding the observed person's social, affective and cognitive states. Beyond the usefulness of understanding the subject's momentary state, eye tracking and attention target estimation aimed to investigate further the user's personality, general interest and health condition. This non-verbal signal has been researched for decades through different intrusive (*e.g.* head-mounted eye-tracking devices), and non-intrusive systems (*e.g.* monitor-mounted tracking devices and CCTV cameras) due to its usefulness [22].

Eye-tracking systems allow specialists to record and monitor the movement and the position of the subject's eyes during an event or stimuli. The stimuli can originate from a display (*e.g.* advertisement) or in the wild (*e.g.* objects or people in the user's surroundings) [23]. The most widely used oculography method to track the eyes' location, motion and gaze direction is video oculography, in which systems obtain information from image data processed with computer vision techniques [24]. Two main research areas of video-based eye tracking are eye localisation and eye gaze direction estimation - approaches of the first category attempt to detect the eye in the image. The latter is focused on estimating the gaze direction from the detected eye region.

With the recent improvement of eye-tracking systems, head and monitor-mounted eye-tracker devices have become cheaper and improved in terms of resolution and recording frequency, enabling high-precision eye-movement tracking and analysis in the field of human-computer interaction [25]. Furthermore, eye-tracker devices are particularly useful and frequently used due to their accessibility and portability in the field of psychology [26], clinical research [27], and academic marketing and consumer research [28]. These human-focused research topics aim to analyse eye movements *w.r.t.*the provided stimuli and draw a conclusion regarding the subject's health or personal interest based on it. However, the mountable trackers restrict the user's head movements as they are designed to record the subject's gaze points within the display. The wearable devices are obtrusive and require the subject to wear them, which depending on the research field (such as subjects with Autism Spectrum Disorder or Schizophrenia), is not always possible [29]. Furthermore, these devices need to be more adequate to gather large-scale true-to-life gaze data, *e.g.*in contexts such as 3D object manipulation or in-person social interactions [30].

Therefore, researchers developed unobtrusive appearance-based gaze estimation techniques to tackle the drawbacks of eye-tracker devices. Despite the lower accuracy of these methods, it allows us to observe and study human behaviour surreptitiously, without the subject's knowledge [31]. The challenges of this field include head pose variations, subject differences, illumination conditions, *etc.* However, the advancements in deep learning and computer vision techniques improved the reliability of appearance-based methods. They enabled us to observe multi-user, unconstrained environments instead of single-user-constrained ones [14]. This research field aims to map the image data directly on the image content; therefore, it does not require camera calibration or geometry data. Due to its successes, many research areas formed based on the foundations of this research field, including single and joint attention target estimation, which aims to estimate the subject's gaze target from a third-person view. Attention target estimation techniques are widely used in social awareness tracking and neurophysiology studies [32].

This thesis investigates the previously introduced social clues, natural language expressions, and gaze, which humans rely on during everyday life. To demonstrate the usefulness of this work, in Figure 1.1, we show a simple overview example of the applications and use cases where the contributions of this work are to be considered. In the figure, the subjects observe the TV screen while discussing parts of the visual stimuli. During this conversation, they use natural language expressions to describe the part of the image to direct the other person's attention to something relevant. This everyday life scenario, description-based image part localisation, is investigated in Chapters 2 and 3. Furthermore, from the third-person point of view, the subjects attempt to understand better what is happening in the video they are watching. Here, we show a typical example where multiple actors are gazing at the same target during a scene in the movie. This action, the joint attention target estimation of multiple people, is investigated in Chapter 4.

1.2 Aim and Objectives

1.2.1 Aim

This thesis investigates the usefulness of integrating implicit and explicit social clues and their combination into frameworks to solve different computer vision tasks.

1.2.2 Objectives

The objectives of this thesis are:

- To propose a solution to the user description-based automatic image cropping problem where the algorithm explicitly considers the user's natural language input and produces a high-quality output crop that best represents the described part of the image.

- To investigate the usefulness of the semi-explicit eye movement input of the user collected during the image description as part of the description-based image cropping optimisation framework.

- To use implicit social clues such as gaze direction to estimate the joint attention target of the subjects within the scene from a third-person view.

1.3 Thesis Statement

Integrating verbal and non-verbal social clues, such as natural language description, gaze and their combination, in a multimodal framework can improve the algorithm's performance in computer vision applications such as automatic description-based image cropping or joint attention target estimation of subjects within the 3D scene from a third person view.

1.4 Thesis overview

My research is organised into three main chapters, and their overview is as follows:

1. **Caption and Aesthetic-Guided Image Cropping**
   In Chapter 2, I introduce a novel description and aesthetics-guided image cropping optimisation framework using existing pre-trained networks. This method is the first one to integrate explicit natural user description into an image cropping framework to automatically produce an aesthetically pleasing output which well preserves the user's intention. The proposed solution is extensively evaluated using a novel dataset through quantitative evaluation and multiple user studies. The dataset and the algorithm presented in this chapter were published in the paper by Horanyi *et al.* [33].

2. **Gaze Initialised, Description and Aesthetic-Based Image Cropping** In Chapter 3, I present a new multimodal framework which leverages information from the user's eye movements during caption generation. By using an eye-tracking device, the collected eye movements of the user can be used to improve further the localisation accuracy of the previously introduced solution for the description-based image cropping task. The eye-tracking data and the novel image-cropping framework were accepted to *The 2023 ACM Symposium of Eye Tracking Research & Applications (ETRA)* [34].

3. **Where Are *They* Looking in the 3D Space?** In Chapter 4, I present the findings on depth-guided joint attention target estimation on images. In this chapter, I worked with implicit social clues, such as gaze direction and head and body pose, to estimate the attention target of the subjects within the 3D scene from a third-person view. Experiments show that the method achieves state-of-the-art results on multiple benchmark datasets for both joint and single attention target estimation tasks. The proposed joint attention target estimation framework by Horanyi *et al.* [35] was accepted to The 5th International Workshop on Gaze Estimation and Prediction in the Wild (GAZE 2023) at CVPR 2023 and won the Best Paper Award.

The publications mentioned above [33–35] are based on the work presented in this thesis. The figures from the submitted and published papers have been used in this thesis.

Chapter 2

Repurposing Existing Deep Networks for Caption and Aesthetic-Guided Image Cropping

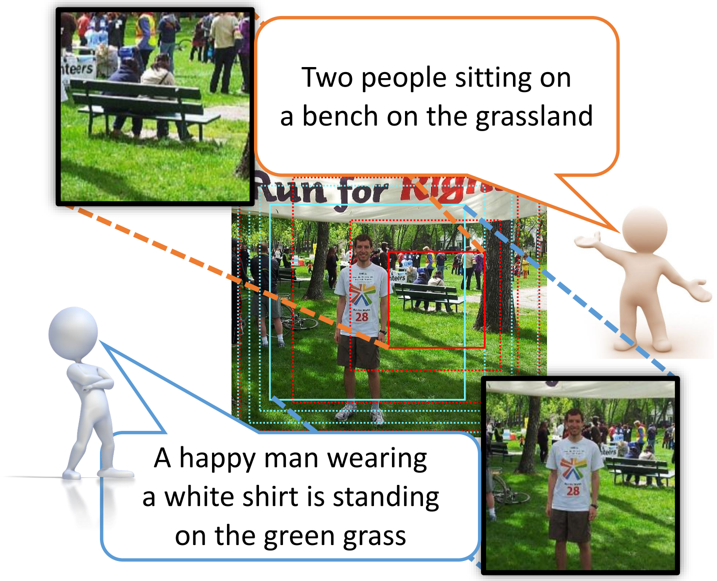

Two people sitting on a bench on the grassland
A happy man wearing a white shirt is standing on the green grass

Figure 2.1: **Illustration of different output image crops of the proposed method.** Our proposed framework reliably crops visually pleasing images by following the natural descriptions of users, and this can produce distinct crops on the same image based on different users' descriptions.

2.1 Introduction

With the advent of social networks, it is now common that images are provided with captions and tags – for example, via Instagram or Twitter – where captions are highly tied in with the user's intentions regarding these images. Therefore, an automated process for enhancing images, for example, providing artistic crops or making thumbnail images that respect user intentions, would be useful for these social networking services (SNS). Besides SNS applications, more connected in with computer vision applications such as semi-automated image dataset generation [36], tracking target object initialisation [37], and story-based automatic image transition generation [38] can also benefit from text-based image cropping.

In this Chapter, as illustrated in Figure 2.1, we focus on a novel image cropping task, which aims to crop an image automatically based on user intent – expressed through a natural text-based description – and the aesthetics of the cropping outcome. Because of the usefulness of an automated image cropping system, various methods have been suggested. However, existing image cropping methods [39–42] are typically designed to be purely automatic, leaving the user out of the loop. For example, [41] automatic cropping is based on maximising the saliency inside the cropping region. Attention-based methods like this try to preserve the most salient part of the image during cropping. One major downside of them is that the user cannot influence their behaviour. Recent works [43–46] have focused on making this process even easier by automatically cropping a photo based on aesthetics. Although fully automated, aesthetics-based methods leave no room for the user to intervene. Furthermore, the work based on aesthetics provides no guarantee that the images' initial content and intent are preserved.

Efforts have also been taken toward methods that take user intention as input. Description-based object detection [47] and localisation [48] have recently been proposed for this purpose. However, these description-based methods do not consider how natural images are created, and their behaviours are far from how humans would crop. Most distinctively, these methods provide very tight cropping around an object, which is different from how people crop images and is often not visually pleasing.

As in many other areas of computer vision, towards this goal, one could apply an end-to-end deep learning framework to find crops that fit the descriptions given an image [47]. However, training such a network in a typical supervised deep learning setup would require an immense amount of labelled data, with captions for multiple sub-regions of the images and their respective ground-truth crops, which may be subjective depending on the creator of the dataset. This is challenging as there is no guarantee that one crop is better than the other as long as the contents inside the cropped regions are identical. Therefore, in this Chapter, we take an alternative approach that exploits existing networks trained for related tasks – image captioning and aesthetics estimation – and repurpose them to automatically crop the images, thereby avoiding the hardships of training a separate network.

Our contributions are four-fold:

- We propose a new deep network repurposing framework to optimise crop parameters *directly* using a bilinear sampler [49], a pre-trained image captioning network [50], and a pre-trained aesthetic estimation network [46].
- We optimise to find the crop region that best fits the provided caption in terms of the image captioning network losses and maximises the aesthetics network scores.
- We generate a new dataset with multiple ground truth bounding box annotations for each caption.
- With the approaches above, we could outperform state-of-the-art methods and produce more visually pleasing image crops reflecting user intention well.

The outcome of the optimisation should be a crop of a photo whose content would fit the caption provided by the user and be aesthetically pleasing. To achieve these objectives, we optimise the parameters of the bilinear sampler, such that the cropped image minimises the losses related to image captioning and aesthetics. Utilising the two networks requires special attention. More specifically, since we are repurposing the two pre-trained networks for a different purpose, we keep them intact to minimise the change inflicted upon them. However, as we are not using the two networks with their original purpose - learning to generate image captions or compare two images to find the one that looks better, we propose a new loss term that we minimise instead of their original ones: learning to generate image captions and measuring how good the image looks. For the image captioning network, we propose to ignore the order of the words as we want the contents to be accurate. We aim to ensure that the correct objects are present and not concerned with the referring expressions where the order matters. For the aesthetic network, we directly maximise the aesthetic score.

Since this optimisation process is, in its basic form, highly unstable due to the nature of the image gradients, we further propose a new optimisation strategy based on *scale annealing* and *multiple restarts*. Instead of directly optimising for the position and the scale of the crop, we optimise only for the position and anneal the scale throughout the optimisation. We use a *multiple restart technique*, starting from a random location for each scale and using the average of the optimisation outcomes. Finally, we take advantage of the fact that image captions should not differ drastically as the image is blurred and use a multi-scale representation of the image for the captioning network.

2.2 Related Works

In the following two subsections, we will discuss the most relevant related works: image cropping and image captioning.

2.2.1 Image Cropping

The increasing amount of digital data produced daily by different mobile devices requires automatised processing and editing techniques to identify their meaningful and important parts. Since manual editing of images is time-consuming even for professionals, image editing methods have been extensively studied, yielding various solutions. Most approaches focus on image cropping, which removes unwanted or distracting elements of the original image while preserving the content and enhancing the visual quality based on aesthetic value or human attention focus.

Image cropping methods can be applied to a wide range of applications. Park *et al.* [51] proposed a photo re-arrangement application based on a learning-based photo composition model. Photo re-arrangement is a set of post-processing techniques for improving photo appearance through cropping or re-targeting. More recent papers used image cropping in new applications. Shan *et al.* [43] developed an automatic photo cropping system that determines the optimal cropping area for better aesthetics. This hybrid deep learning-based framework learned internal image representations using a convolutional auto-encoder and manually extracted features for automatic cropping of distracting people in photographs. The image cropping research can be categorised into *attention-based*, *aesthetics-based*, and *description-based* approaches.

Attention-based methods

This class of methods exploits visual saliency models or salient object detectors to find the most visually important regions in the original image [52, 53]. Most attention-based cropping methods rank candidates based on their attention score [54, 55]. Thus, these methods can identify the image's most important and attractive regions. Recent methods choose the important area based on certain attention scores [56–59]. One of the most recent methods [41] similarly to [39, 60–62] focuses on those image regions that attract the human gaze at first sight.

Aesthetics-based methods

Methods based on aesthetics crop images by relying on the attractiveness of the cropped image with the help of a quality classifier [39, 63, 64]. These methods aim to extract the optimal rectangular sub-region of a given image to produce an image with a high aesthetic score [42, 45]. Recently, Li *et al.* [44] formulated the automatic image cropping problem as a sequential decision-making process and proposed a weekly supervised approach which only uses aesthetic information as supervision. This method was the first to use reinforcement learning for automatic image cropping and overcame the disadvantages of the sliding window method. The existing aesthetics-based cropping methods can be further categorised into supervised [65, 66] or weekly supervised [40, 46, 55, 67] methods. In particular, the weekly supervised methods, which do not include bounding box supervision, have been researched actively as producing cropping box annotations for the training is expensive [68].

Description-based methods

These methods aim to localise a region described by a given referring expression. Most of these methods treat comprehension as bounding box localisation, similar to our cropping task. A recent method by Rohrbach *et al.* [48] uses joint embedding to find the object directly by selecting the best region based on an input expression. Yu *et al.* [47] proposed a modular network for referring expression comprehension. Unlike the previous works, this method does not treat expressions as a single unit but decomposes them into three-phrase embeddings. This module-based approach enabled it to outperform previous state-of-the-art methods in comprehension tasks, both bounding-box-level and pixel-level. However, the performance of this method is limited in our application because it was originally designed for object detection purposes. Similarly to our application, MAttNet [47] and Align2Ground [69] can locate the image region described by a general referring expression. Their downfall is that these methods rely strictly on input expressions and do not consider aesthetics. In addition, MAttNet heavily relies on the specific decomposition of the expression into subject appearance, location, and relationship to other objects, which may not exist in natural descriptions of images.

2.2.2 Image Captioning

Image caption generation aims to compress a large amount of salient visual information of images into descriptive language, meeting the grammar rules. Following the successes of deep neural networks in machine translation, neural networks became the main tools for solving image caption problems [70]. Following methods [71–74] had considerable interest in LSTM-based [75] image captioning, which depends on the pre-specified visual attribute quality. Xu *et al.* [50] introduced attention-based image caption generators, where the framework learns latent alignments from scratch and does not explicitly use object detectors. The most recent LSTM-based method [76] generates high-quality guidance by considering object detection-dependent attention instead of searching at the whole noisy image. In this work, we utilise [50].

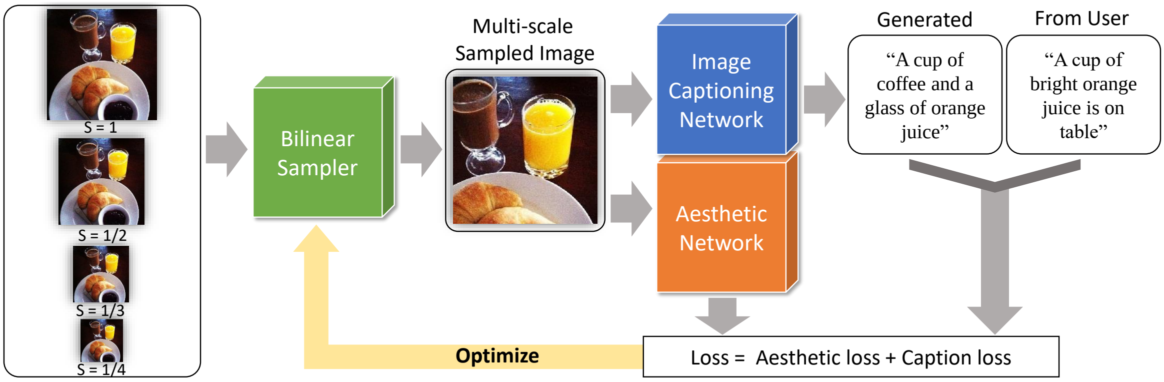

S = 1
S = 1/2
S = 1/3
S = 1/4
Bilinear Sampler
Multi-scale Sampled Image
Image Captioning Network
Aesthetic Network
Generated
"A cup of coffee and a glass of orange juice"
From User
"A cup of bright orange juice is on table"
Optimize
Loss = Aesthetic loss + Caption loss

Figure 2.2: **Overall framework of the proposed method: CAGIC.** The framework takes an image as input, which goes through multi-scale bilinear sampling to produce a cropped image. Note that the parameters for this sampling do not come from a network as in other existing works but are the parameters we will directly optimise for later. We then use this cropped region as input to both the image captioning and aesthetic networks.

2.3 Methodology

We currently have deep networks designed to do many traditional computer vision tasks, and they perform, in many cases, even better than the traditional ones. Using pre-trained networks as backbones for performing a certain task is also quite common. Here, we propose an alternative and use these existing networks as building blocks. We are proposing to go beyond the paradigm that deep networks should give a solution in a single shot and instead perform inference through optimisation, as was a common strategy before deep learning.

We first describe the overall architecture, including our *multi-scale sampling strategy*, and then explain how we perform inference by *optimising* the framework instead of training networks to obtain the desired crop region. We further detail how we can stabilise this optimisation process through *scale annealing* and *multiple restarts*.

2.3.1 Framework

Figure 2.2 shows the overall framework of the proposed Caption and Aesthetic-Guided Image Cropping (CAGIC) method. Our framework comprises three major components:

1. Bilinear sampler that operates on a multi-scale

2. Image captioning network

3. Aesthetic network

The image captioning network automatically generates a natural language expression describing the given image's content. There is a large body of work on this problem [77]. Among those for the image captioning network, we use the method from [50], with the models pre-trained with the MS-COCO [78] dataset. Visual aesthetic preference can be described as either a single score or a distribution of scores. We follow the definition in [46], where professional photographs provide the aesthetic score. For the aesthetic network, we use [46] with the pre-trained models[1]. Note that we take special care that none of the images used in training any of the pre-trained models is included in our evaluation later. Even though these two networks were trained on entirely different datasets, we found that the pre-trained models were good enough for our purpose.

Multi-scale bilinear sampling

As we are directly optimising for the crop region's location and scale, the bilinear sampling's gradients must be robust. To ensure this, we propose to use a multi-scale strategy inspired by the observation that even when an image becomes blurry, its content does not change. Therefore, if we denote the bilinear sampling process as $\mathrm{Sample}\left(\mathbf{I},\theta\right)$, where $\mathbf{I}$ is the image and θ is the crop parameters composed of the centre coordinates of the crop x and y, and its scale s, and $\mathrm{Resize}\left(\cdot\right)$ is the resizing operation, for the cropped image $\mathbf{I}_{crop}\left(\theta_s\right)$ we can write

$$\mathbf{I}_{crop}\left(\theta_s\right) = \frac{1}{|\mathbf{S}|}\sum_{s\in\mathbf{S}}\mathrm{Sample}\left(\mathrm{Resize}\left(\mathbf{I},s\right),\theta\right), \tag{2.1}$$

where $\mathbf{S} \in \left\{\frac{1}{4},\frac{1}{3},\frac{1}{2},1\right\}$ is the set of scales, and $|\mathbf{S}|$ is the cardinality of this set, which is four. We also omit $\mathbf{I}$ on the left-hand side for brevity. Here, we set the sampling process always to consider the source image coordinates to be between -1 and 1, thus removing the need to adjust the sampling parameters per image. After each resizing operation, we also apply Gaussian image blurring filters.

2.3.2 Inference

With the framework, we infer the parameters of the crop θ by optimising the network *w.r.t.*the image captioning networks loss and the aesthetic network output. However, as we are not using the two networks with their original purpose—learning to generate image captions or compare two images to find the one that looks better—, here, we formally introduce our optimisation objective.

[1] https://github.com/yiling-chen/view-finding-network

If we denote the two objectives as $\mathscr{L}_{caption}$ for the image captioning part and $\mathscr{L}_{aesthetic}$ for the aesthetic part, for a given pair of image $\mathbf{I}$ and caption $\mathbf{y}$, our objective is therefore to find $\hat{\theta}$ such that

$$\hat{\theta} = \arg\min_{\theta} \mathscr{L}_{total}\left(\mathbf{I}, \mathbf{y}, \theta\right), \tag{2.2}$$

where

$$\mathscr{L}_{total}\left(\mathbf{I}, \mathbf{y}, \theta\right) = \mathscr{L}_{caption}\left(\mathbf{I}, \mathbf{y}, \theta\right) + \lambda \mathscr{L}_{aesthetic}\left(\mathbf{I}, \theta\right), \tag{2.3}$$

and λ is the hyper-parameter that balances the two loss terms. In all our experiments, we empirically set $\lambda = 0.01$ (see Figure 2.12). The two losses we choose to optimise are closely related to the networks' original formulation but modified to our needs.

Image Caption Loss $\mathscr{L}_{caption}\left(\mathbf{I}, \mathbf{y}, \theta\right)$

When training a network to output an appropriate caption, the order of the words is important. However, this is not necessarily so in our task, as we only want the described objects to be present. We need not regenerate the sentence that the user inputs. In our earlier experiments, we found that when the order of the words was considered, depending on the user, the network focused too much on the order of the words, not on the contents, and the performance could have been better. Thus, for the image caption loss, we ignore the order of the words coming out of the captioning network.

Specifically, if we denote the ground truth one-hot encoded vector representation for the t-th word of the user caption as $\mathbf{y}_t$, the captioning network as $f\left(\cdot\right)$, and cross-entropy as H, we write

$$\mathscr{L}_{caption}\left(\mathbf{I}, \mathbf{y}, \theta\right) = H\left(\frac{1}{T_u}\sum_{t=1}^{T_u} \mathbf{y}_t, \frac{1}{T_c}\sum_{t=1}^{T_c} f\left(\mathbf{I}_{crop}\left(\theta\right)\right)_t\right), \tag{2.4}$$

where T_u and T_c represent the number of words in the user caption and the captioning network f generating caption, respectively. Note that we average the word vectors, effectively removing the order information.

Image Aesthetic Loss: $\mathscr{L}_{aesthetic}\left(\mathbf{I}, \mathbf{y}\right)$

For the aesthetic term, we aim to maximise the aesthetic score output from the network. If we denote the aesthetic network as $g\left(\cdot\right)$, we therefore write

$$\mathscr{L}_{aesthetic}\left(\mathbf{I}, \theta\right) = -g\left(\mathbf{I}_{crop}\left(\theta\right)\right). \tag{2.5}$$

where, as above, $\mathbf{I}_{crop}$ is from Eq. (2.1) and $\mathbf{y}$ is the user caption. This loss helps our crops to be realistic crops close to how humans would crop images as the aesthetic network learned any photographic rules implicitly encoded in professional photographs [46].

2.3.3 Stabilising the optimisation

We design a new cost function considering the two networks' outputs together, searching among various bilinear samples. In Figure 2.3 we visualised the total loss space for different cropping parameters. During the optimisation process, the cropping parameter shrinks by 2% in every iteration (See more details in Scale Annealing). The percentages in this figure correspond to the cropping sizes *w.r.t.*the original image size. Choosing the correct scale is important to ensure that the crop will include all the relevant parts of the image based on the user's description. However, the image cropping parameter space is too large and non-convex, as shown in Figure 2.3, so unstable convergence is inevitable. Our initial attempts of indirectly optimising Eq. (2.2) were not very successful, even with the help of the scale-space bilinear sampling in Eq. (2.1). We, therefore, propose two additional methods that stabilise the optimisation process, leading to better final results. We explain these methods below and summarise the entire optimisation algorithm in Alg. 1. Finally, in Figure 2.4 we show an example of how the cropping changes over the iterations to fit a given caption. We visualise the ground truth region as green and the cropping region of the current iteration as red. As we can see, the cropping results gradually converge near the ground truth over iterations.

Scale Annealing

One hardship when directly optimising for the crop parameters is that pixels outside the crop region have no means to affect the optimisation process once determined to be outside of the crop region. This leads to instability when also optimising for scale, as, for example, when the crop accidentally shrinks, it will be difficult for the system to recover from it. Therefore, we exclude the scale parameter from optimisation and anneal the scale to become smaller

Algorithm 1 Optimization with multiple restart.

```
Require: I : input image, y : user caption
function OPTIMIZE(I, y)
    for $i = 1$ to $S$ do                                  ▷ For each scale level
        for $k = 1$ to $K$ do                              ▷ Optimize $K$ times
            $x_0 \sim \mathcal{N}\left(x^i, \sigma\right)$          ▷ Restart initialization
            $y_0 \sim \mathcal{N}\left(y^i, \sigma\right)$
            $\hat{x}^i_k, \hat{y}^i_k \leftarrow$ Eq. (2.7)          ▷ Find optimum point
        end for
        $x^{i+1} \leftarrow \frac{1}{K}\sum_{k=1}^{K}\hat{x}^i_k$        ▷ Gather result
        $y^{i+1} \leftarrow \frac{1}{K}\sum_{k=1}^{K}\hat{y}^i_k$
    end for
end function
```

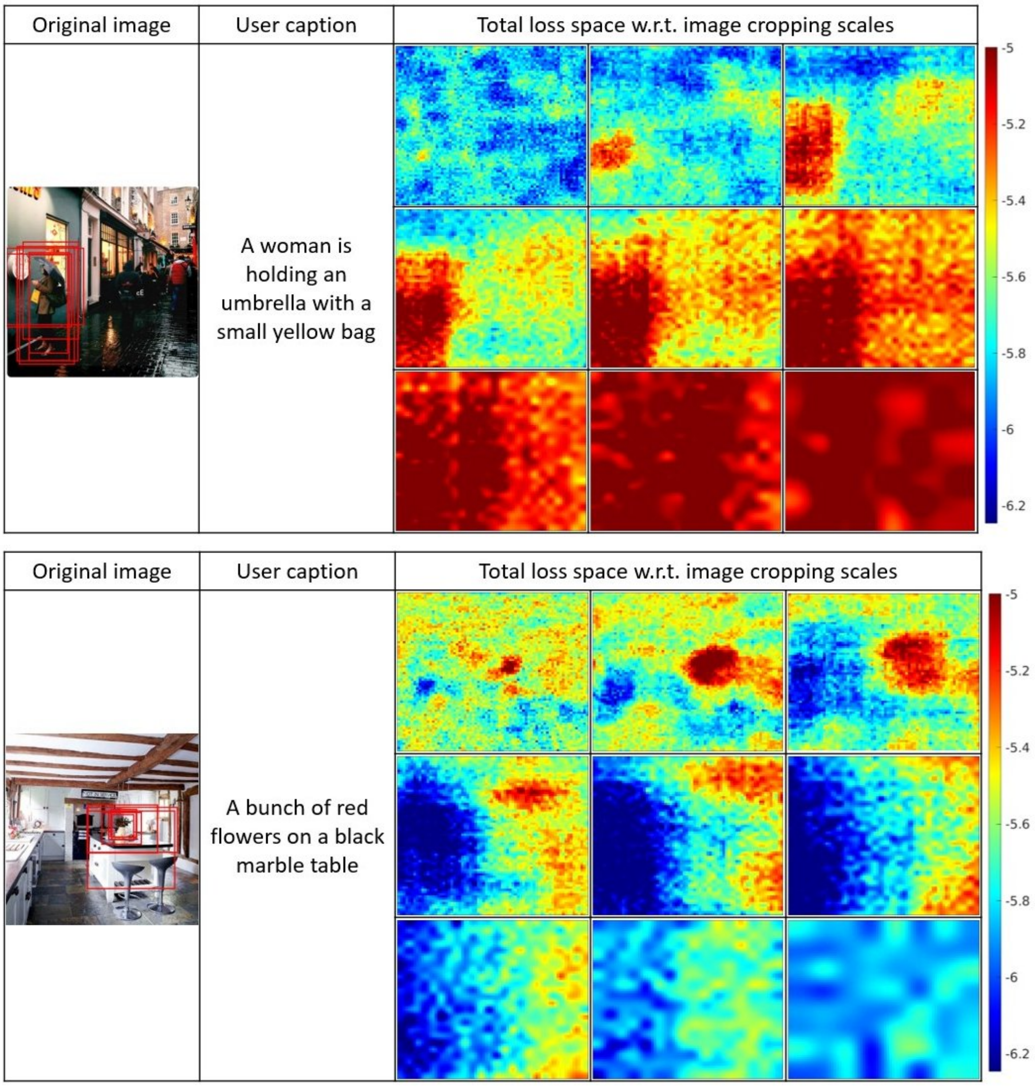


Figure 2.3: **Visualisation of the total loss space for different image cropping parameters.**. From the top left to the bottom right, we show the cropping scales 10-90% of the original image size.

throughout the optimisation process. Specifically, we set the scale to be s^i, where i is the optimisation iteration. We empirically set $s = 0.98$. In other words, the scale is reduced by 2% at each iteration. During optimisation, we track which crop parameter gave the lowest loss and return that crop region as our final result.

Optimising with Multiple Restarts

To further stabilise the optimisation process and escape local optima, we employ a multiple restart technique [79] as shown in Figure 2.5. Based on the research of [80], for each

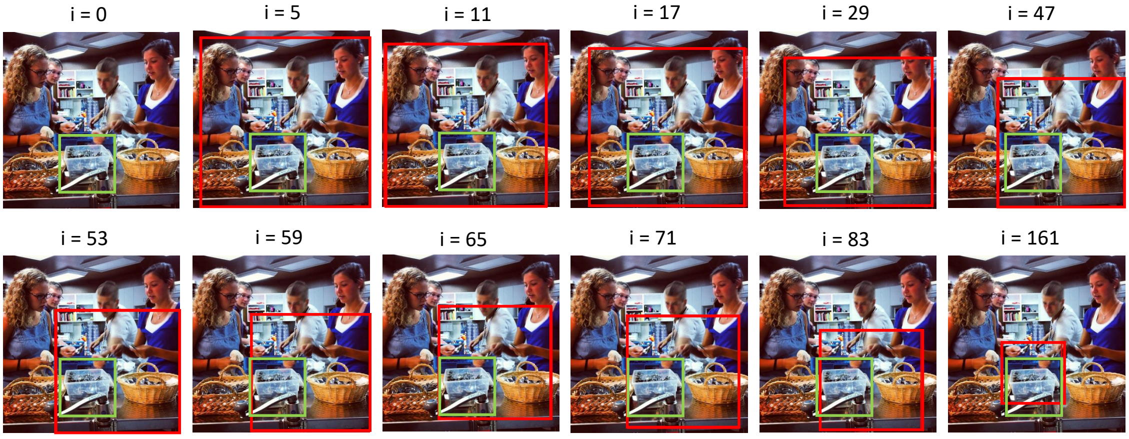


Figure 2.4: **Output image crop change over time.** Example of our method iteratively updating crop to fit caption: “A plastic box of many metal forks".

optimisation iteration, we apply random noise $\Delta x \sim \mathscr{U}(0,1)$ and $\Delta y \sim \mathscr{U}(0,1)$, where $\mathscr{U}(0,1)$ is the uniform distribution, to the outcome of the previous iteration. After adding the noise, we further clip them to prevent the cropped region from going out of the image border. We then run our optimisation to find the optimal crop for this given scale, repeat the process K times and average their results to obtain our final solution for this scale (in our experiments, we use the $K = 10$).

If we denote the crop centre estimates at iteration i as x^i and y^i, we write

$$\left(x^{i+1}, y^{i+1}\right) = \frac{1}{K} \sum_{k=1}^{K} \left(\hat{x}_k^i, \hat{y}_k^i\right), \tag{2.6}$$

where

$$\left(\hat{x}_k^i, \hat{y}_k^i\right) = \underset{x,y|x_0 \sim \mathscr{U}(0,1), y_0 \sim \mathscr{U}(0,1)}{\arg\min} \mathscr{L}_{total}\left(\mathbf{I}, \mathbf{y}, \left(x, y, s^i\right)\right), \tag{2.7}$$

and x_0 and y_0 are the starting points of optimisation for x and y respectively, and x_i and y_i are the final converged locations for this scale level. To find the $\hat{\theta}$, we apply limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [81], as we are searching for the optimal location and not simply seeking to perform gradient descent for each scale. The L-BFGS approach is one of the most popular quasi-Newton methods that construct positive definite Hessian approximations. It is a local search algorithm for convex optimisation problems with a single optimum. Thus, the L-BFGS is suitable for our method, as we apply scale annealing and prefer local optimisations. Although the multiple restart strategy does not theoretically guarantee that the average location is lower in terms of the final objective, it ensures that we always optimise towards the general improving direction.

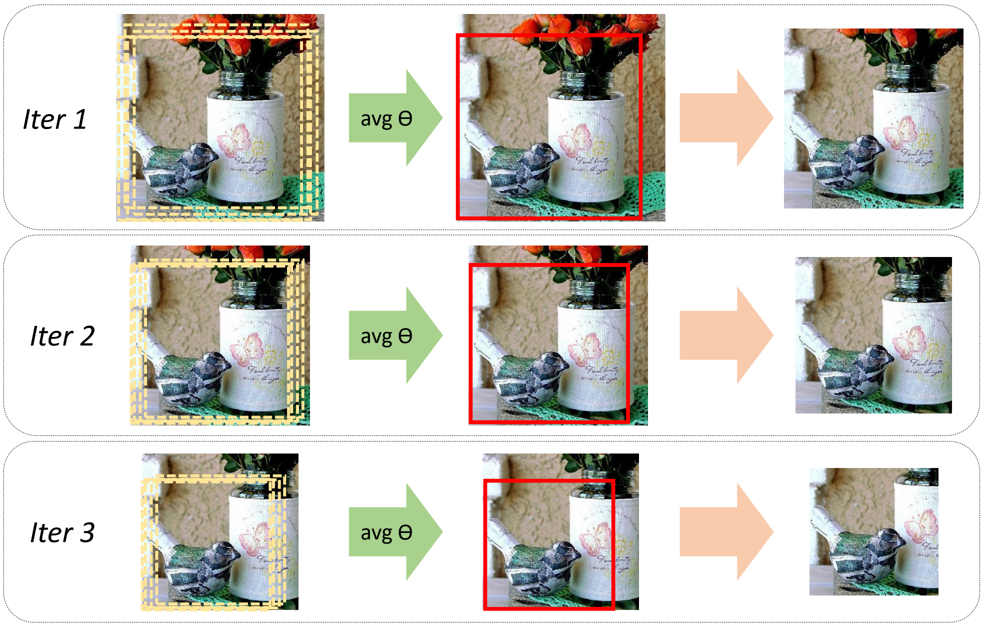


Figure 2.5: **Visualisation of the multiple restart strategy and scale annealing.** For each iteration, we employ the multiple restart strategy which starts random location and takes the average of the optimisation outcomes. In the next iteration, the scale is annealed as scheduled.

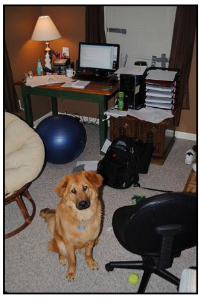
2.4 Experiments

We implement our method in TensorFlow [82]. All experiments are run on an Intel i7- CPU @ 3.40GHZ, 16 GB RAM, and two NVIDIA TITAN Xp GPU.

2.4.1 Dataset

We created a novel dataset as the task we aim to solve, namely natural language description-based image cropping, differs from what existing datasets offer. Before we discuss our dataset in detail, we first give an overview of existing datasets and their drawbacks.

A widely used dataset for natural images with captions is the MS-COCO dataset [78], which contains 123k images with predefined splits for training, validation, and testing. Each image is annotated with five captions by Amazon Mechanical Turkers. However, most captions describe the whole image and are unsuitable for caption-guided cropping. Therefore, based on MS-COCO, the ref-COCO [83] dataset was proposed, where parts of images are described with annotated bounding boxes. However, this dataset is heavily skewed towards the tight detection of certain objects and is unsuitable for creating natural crops. They are aimed more at text-guided object detection.



Dataset	Caption
MS-COCO	A puppy dog sitting in an office with a computer in it
refCOCO	Chair to the left
Proposed	A computer and a lamp on the top of the desk

Figure 2.6: **Caption comparison of different datasets.** We show the captions from the MS-COCO, ref-COCO, and our new dataset corresponding to ImgId=38353 of the MS-COCO train2014 set.

In Figure 2.6, we show example captions on the same image from different datasets to demonstrate that captions from ref-COCO [83] and MS-COCO [78] are inadequate for our tasks. As one can see, the caption from MS-COCO describes the entire image, while the captions from ref-COCO are very short and only related to a single object. However, in our dataset, we ask the users to provide captions describing a specific image region, resulting in a natural description. The caption also often describes the surroundings, providing information on what people think as context.

Visual Genome dataset

The Visual Genome dataset [84] is the closest to what we need for our task. This dataset was designed to connect language and vision through natural language expressions. This large dataset consists of 108k images from the YFCC100M [85] and MS-COCO datasets. Different from the original captions of the MS-COCO dataset, the Visual Genome provides

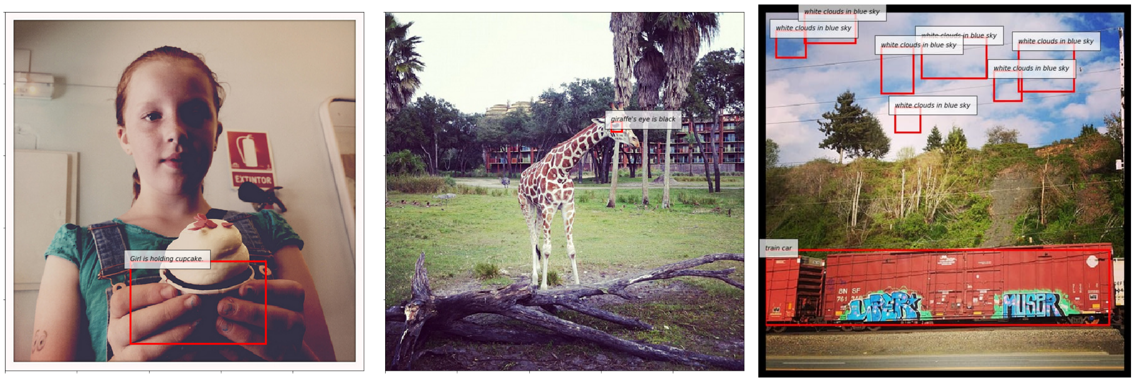


Figure 2.7: **Example captions and corresponding annotations of the Visual Genome dataset.**

about 50 descriptions of different image regions. The annotations for this dataset are also geared towards more traditional localisation and description tasks. However, two problems in this dataset make it not feasible to evaluate natural crops:

1. Annotations are tight bounding boxes focused on objects – that do not represent how people typically take photos.

2. Identical captions can denote multiple regions – for example, “red sky” could correspond to any bounding box within the sky – demonstrating that annotations are only part of the possible ground truth and not overlapping with any of these does not mean that a crop is wrong.

As shown in Figure 2.7, the captions and annotations would not look like natural crops if they were to be cut out – they are either too tight and sometimes not even capturing the entire object, as shown by the examples. Also, notice that the captions are very short, almost as if they are descriptions of a single object. There is also an issue of non-specific regions, as in the case of Figure 2.7 right, where the descriptions all talk about the sky and the cloud. In addition, none of the crops would be recognisable to a human being on what they are about. Repetitive captions that refer to different image parts are problematic for caption-based localisation tasks as they can result in misleading outputs.

Figure 2.8 presents the subjective nature of the bounding box annotations. While the annotations vary largely, when a user is requested to draw a bounding box around the described image region, the main object of the caption is fully included. In contrast, in Figure 2.7 left, we show an example annotation from the Visual Genome dataset with the caption "Girl is holding cupcake", where the corresponding ground truth bounding box does not include the entire cupcake or the girl itself. If we show this image’s ground truth output crop, it would not be possible to reconstruct the original caption; therefore, we argue that this annotation needs to be revised and corrected. In the middle of Figure 2.7, the caption is "Giraffe’s eye is black". The corresponding bounding box is so tight that by seeing only the content of that image part, a user would not be able to identify what is shown on the image. In the last example, we show one of the many images with multiple identical captions with different bounding box positions. For example, for Image id 2342467 (See in Figure2.7 right), we visualised 7 of the 16 different image parts assigned to the "White clouds in red sky" caption.

A novel dataset

Therefore, we create a novel dataset based on MS-COCO, similar to the ref-COCO and Visual Genome datasets. Similarly to the ref-COCO, we used a subset of the MS-COCO

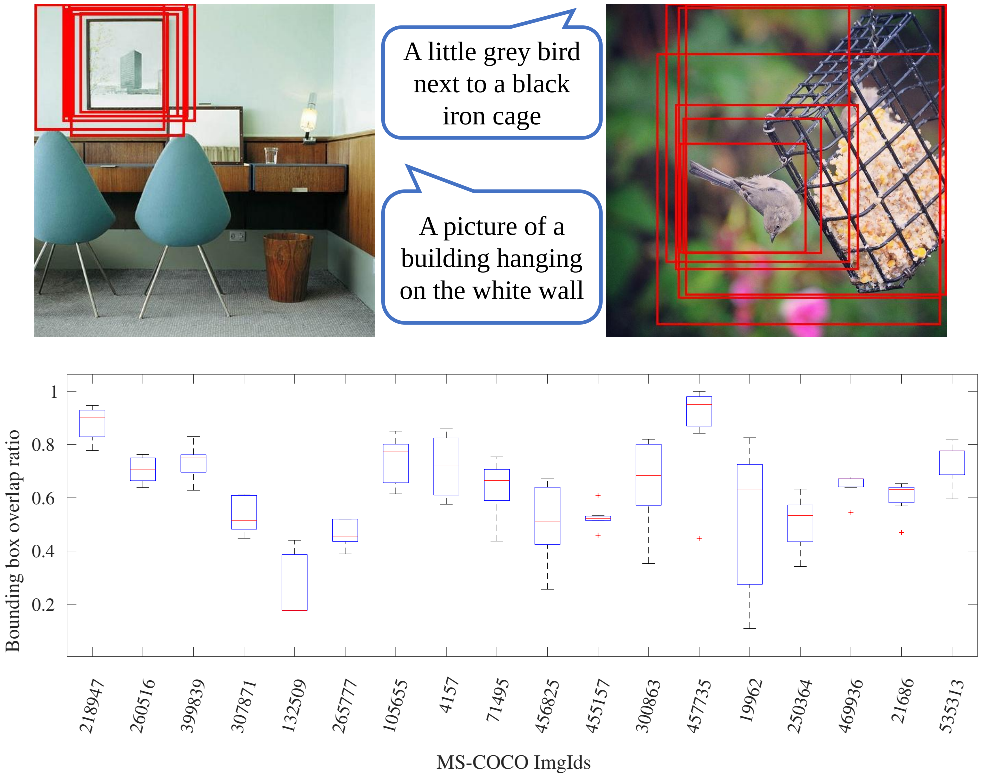


Figure 2.8: **Analysis of the ground truth bounding box annotations.** (Top) Illustration of the diversity of the bounding box annotations for different images. (Bottom) Box plots of the distribution of ground truth overlap for different MS-COCO images.

dataset. We select 100 images randomly from the MS-COCO test set to avoid the images being ever seen by any of our pre-trained networks and create our captions for each image. The documentation of the dataset is available in Section A.2.

Ground-truth captions. In our new dataset, we are interested in evaluating the ability to generate crops related to a given caption automatically. To remove the subjective nature of our ground truth annotations as much as possible, we carefully select regions of the image that are unique and distinctive for a human annotator to create a "natural" crop out of. We then manually generate descriptive expressions, focusing on the selected distinctive parts of the image. The annotators were asked to produce an image description freely. We did not instruct them to generate referring expressions, they were asked to describe the image part with their own words. During the caption generation, we confirmed that the captions described only one part of the image. Our captions are roughly ten words long on average to ensure that they are specific and clear.

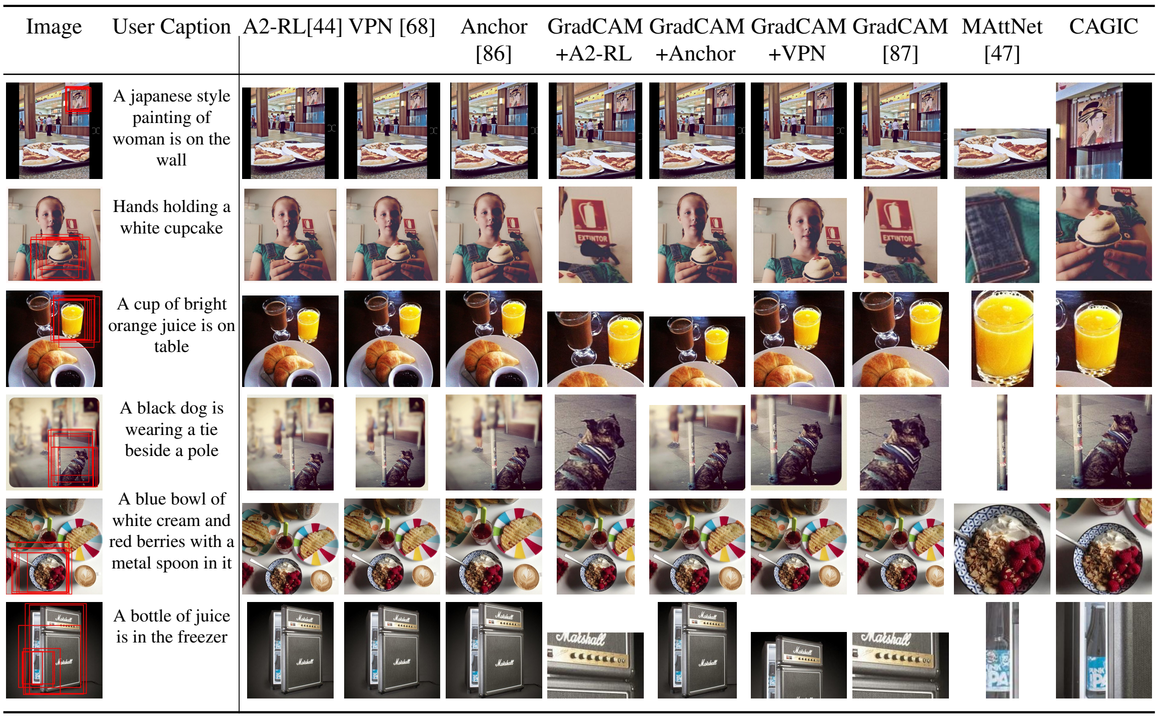


Image
User Caption
A2-RL[44]
VPN [68]
Anchor [86]
GradCAM +A2-RL
GradCAM +Anchor
GradCAM +VPN
GradCAM [87]
MAttNet [47]
CAGIC
A japanese style painting of woman is on the wall
Hands holding a white cupcake
A cup of bright orange juice is on table
A black dog is wearing a tie beside a pole
A blue bowl of white cream and red berries with a metal spoon in it
A bottle of juice is in the freezer

Figure 2.9: **Qualitative comparison with the baseline methods.** The cropped images obtained by the proposed method and the eight baseline methods. The user-defined ground truth bounding box annotations are shown on the original images in red. The proposed method well crops the images as the user described.

Ground-truth crops. One of the tricky parts in creating a dataset for caption-based image cropping is the definition of ground truth. The concept of "which image region fits the description well" is subjective and can wildly differ from person to person. Therefore, we asked seven participants to generate ground truth crops based on our captions individually. We further asked the participants to consider the aesthetics of these crops. We use these ground truth annotations to perform quantitative comparisons. As an evaluation metric, we use the Intersection over Union (IoU), a standard metric for evaluating bounding box-based tasks [46, 88].

Even with care, it is inevitable that the cropping task is subjective and dependent on the annotator. Note that some of the caption and crop annotators were non native English speakers and were from different countries. This is important to note as their country of origin could influence how they generate captions and what they find visually pleasing. However, despite the subjective nature the collected ground truth annotations for each caption are useful to understand better the flexible and non-unique nature of image cropping [86]. As we employ a multiple ground truth strategy to alleviate annotator bias, we investigated how much in agreement the annotations are. We show an example of the ground-truth annotations in Figure 2.8 (top), as well as the agreement between annotators in Figure 2.8 (bottom). Note that even when disagreeing on the exact crop, they are all overlapping – the main content is shared.

2.4.2 Baseline methods

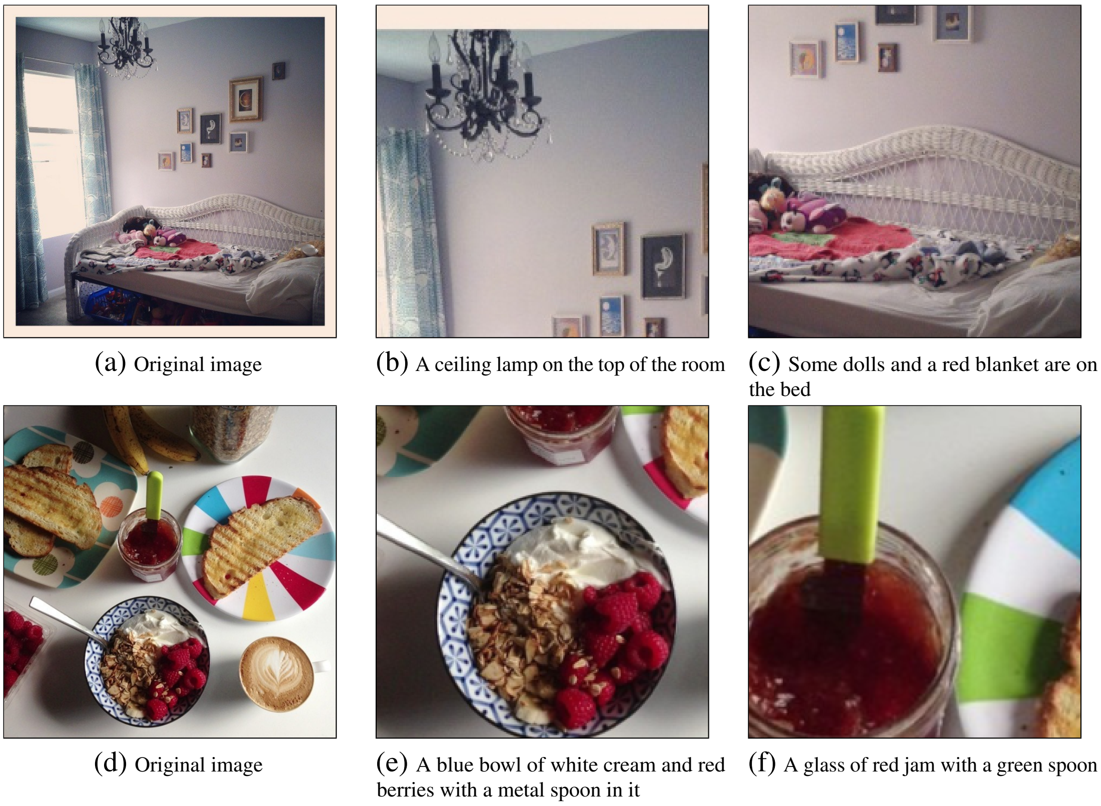
Due to the novel nature of our problem formulation, only a limited number of baselines exist. We compare our method against the following eight different methods:

1. **GradCAM [87]** a naive untrained baseline where we apply GradCAM with the captioning network to extract regions in the image corresponding to the user caption. GradCAM is an algorithm that can be used to visualise the class activation maps of a Convolutional Neural Network, highlighting where the network is "looking". We then threshold the activation map with a threshold of 0.2 of the maximum value, which we empirically set.
2. **A2-RL [44]** We also compare with a solely aesthetic-based method to demonstrate that these crops do not necessarily correspond to the users' intentions.
3. **GradCAM+A2-RL** We combine the two baselines by executing them sequentially.
4. **VPN [68]** VPN, similar to the A2-RL method, is an aesthetic-based method which was trained on a large-scale Comparative Photo Composition dataset.
5. **GradCAM+VPN [68]** We combined a state-of-the-art view proposing method with GradCAM. VPN generates a set of crops, and we selected the one closest to the GradCAM bounding box.
6. **Anchor [86]** We compared the performance of our algorithm with the grid anchor-based, data-driven image cropping method. This end-to-end trained model can not alter its output according to a given description of the ROI.
7. **GradCAM+Anchor [68]** We combined a state-of-the-art automatic image cropping method with GradCAM.
8. **MAttNet [47]** We compared our results with the state-of-the-art referring expression comprehension network. MAttNet was trained on the ref-COCO dataset [89] to localise the image region described by a natural language expression.

2.4.3 Performance comparison on the proposed dataset

Qualitative results.

We compare the proposed image caption and aesthetics-based image cropping approach with the baseline methods. Here we present the result of these methods in different cases. We



(a) Original image (b) A ceiling lamp on the top of the room (c) Some dolls and a red blanket are on the bed

(d) Original image (e) A blue bowl of white cream and red berries with a metal spoon in it (f) A glass of red jam with a green spoon

Figure 2.10: **Different captions can lead to entirely different crops even on a single image.**

first show qualitative highlights in Figure 2.9. As shown, our approach provides crops of the highest quality. For further qualitative highlights, see Section A.1.

The proposed method can generate much more accurate results than other compared methods. In some cases, the methods may try to exclude irrelevant objects according to the caption, which results in a very small crop region. In this case, although the contents are correct, it becomes difficult for a user to understand the output without the original context which contradicts with our aim to preserve the user's intention. This maybe an issue for potential applications which intend to use the output crop as input.

In Figure 2.10, we present the result of our automatic cropping method using the same image *w.r.t.* different image captions. The results show that our method can deliver entirely different results on the same image when different captions are provided. Notice how our results correspond well to the provided captions and are good-looking, demonstrating that our method works in a well-balanced manner between the two loss functions. In particular, as shown in sub-figures 2.10 (b) and (f), our method well crops regions when the description is about parts of the image that are small or are located near the edge. The cropping is well directed by the provided caption, even when the caption describes an extreme region of the image or a relatively small object from the middle.

Quantitative results – IoU.

To evaluate the performance of the baselines and our method, we computed the bounding box overlap ratio as the area of intersection between two boxes, divided by the area of the union of the two for each ground truth set and calculated the average IoU values for every method. We report the average IoU between produced crops and all seven of the ground-truth annotations in Table 2.1. As shown, our method provides the highest IoU value among the compared methods. Note that we have higher numbers than even those which were trained specifically for referencing from captions. This is mainly due to the fact that existing methods perform tight crops. However, when a human is asked to perform the same task, they tend to include context.

Furthermore, the obtained results of the baseline methods show similar characteristics on our dataset to the previously reported results in [86]. Namely, we observed that A2-RL 28%, VPN 61% and GradCAM+VPN 24% of the images returned the original image as it failed to crop the image. We show that when the aesthetics-based cropping methods (A2-RL, VPN and Anchor) were combined with GradCAM, their performance improved. The major disadvantage of the second-best method, MAttNet is that it tends to generate close-up views of objects. Meanwhile, our approach preserves as many useful parts of the image as possible to ensure that the output reflects the contextual information given by the caption. Furthermore, we show that every method produced image crops with higher IoU measures than the original image.

User study – is the crop really what we want?

Another measure for evaluating the quality of the crops is to measure how well we can preserve the contextual information. As the task is to crop the images using the user's description of the region of interest, we were interested in how similar is the input caption to the output crop's description. We asked four users who were not exposed previously to the original images or the input captions to re-annotate the crop outcomes of the top three methods from the IoU evaluation: GradCAM, MAttNet and the proposed algorithm. We then evaluate how similar these new descriptions are to the target descriptions of our dataset. Note that the input captions have a one-to-one relationship with the generated output descriptions. We report results in terms of metrics used for natural language processing[2] in Table 2.2. Our

[2] https://github.com/Maluuba/nlg-eval

Table 2.1: **Quantitative comparison of the different methods using IoU measure on the output bounding boxes.**

Method	Mean ± Std.
Original	0.287 ± 0.028
A2-RL[44]	0.298 ± 0.024
VPN[68]	0.315 ± 0.023
Anchor [86]	0.333 ± 0.024
GradCAM+A2-RL	0.347 ± 0.011
GradCAM+Anchor	0.355 ± 0.020
GradCAM+VPN	0.356 ± 0.013
GradCAM[87]	0.360 ± 0.202
MAttNet[47]	0.385 ± 0.261
CAGIC	**0.416 ± 0.013**

Table 2.2: **Quantitative comparison of the cross-crop similarity with the two best-performing baseline methods.** Comparison of user intention presence. We ask users to caption cropped images and compare with natural language metrics how similar they are with the original desired caption.

Method	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE_L	CIDEr
GradCAM[87]	0.273	0.141	0.087	0.053	0.118	0.279	0.697
MAttNet[47]	0.172	0.094	0.060	0.036	0.113	0.295	0.715
CAGIC	**0.342**	**0.188**	**0.102**	**0.063**	**0.170**	**0.297**	**0.905**

Table 2.3: **Quantitative evaluation of the user preference.** The top three methods of qualitative comparison along with the original image were compared through a human survey and evaluated by aggregation.

	Original Image	MAttNet[47]	GradCAM[87]	CAGIC
Aggregated percentage (%)	21.04	23.93	25.51	**29.52**

method provides the best results for all metrics, demonstrating that the user intention is best preserved.

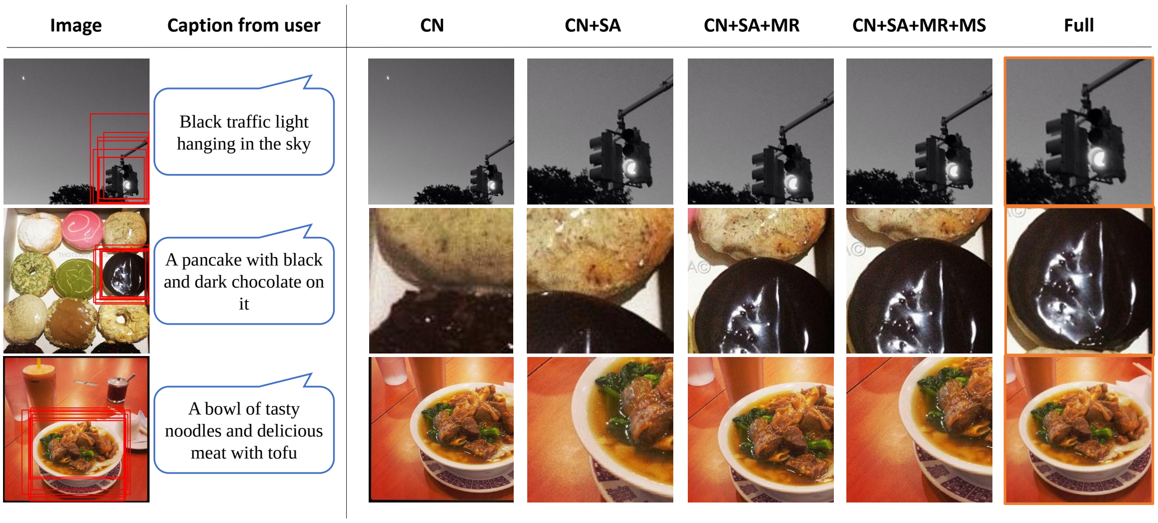


Image
Caption from user
CN
CN+SA
CN+SA+MR
CN+SA+MR+MS
Full
Black traffic light hanging in the sky
A pancake with black and dark chocolate on it
A bowl of tasty noodles and delicious meat with tofu

Figure 2.11: **Qualitative results of the ablation study of CAGIC.** Qualitative comparison among our methods which consist of the combination of Caption Network (CN), Scale Anneal (SA), Multiple Restart (MR) and Multi-Scale Bilinear Sampling (MS). The results show that the combination of all of these elements (proposed full method) together can provide the best output image.

User study – which crop is better?

Due to the subjective nature of our task, we further perform a user study, where we ask users to select which image is preferred over the Top-3 methods, as well as the original image as the baseline. Specifically, we ask users to *"Select the crop described by the caption which looks the best,"* given the four images in a graphical user interface. We ask a total of 13 users, resulting in more than 1000 decisions. In the user study, for each user, we show a randomly selected subset of our dataset. We report the probability of being selected for each method in Table 2.3. Our method is the most preferred among compared methods. The user study shows that the subjects preferred the cropped images over the original image. Users preferred the methods with wider output crops that were not mainly focused on the subject of the image caption; therefore, they were able to preserve the contextual information. While the output of GradCAM was contextually correct, users tended to select the proposed method when the aesthetics of the crop was not pleasing. This highlights that aesthetics need to be considered when emulating what humans would do.

2.4.4 Comparison with the state-of-the-art on the Visual Genome dataset

Despite the drawbacks discussed in Section 2.4.1, for the completeness of this study, we compared the eight baseline methods (See Section 2.4.2) with our performance on the Visual

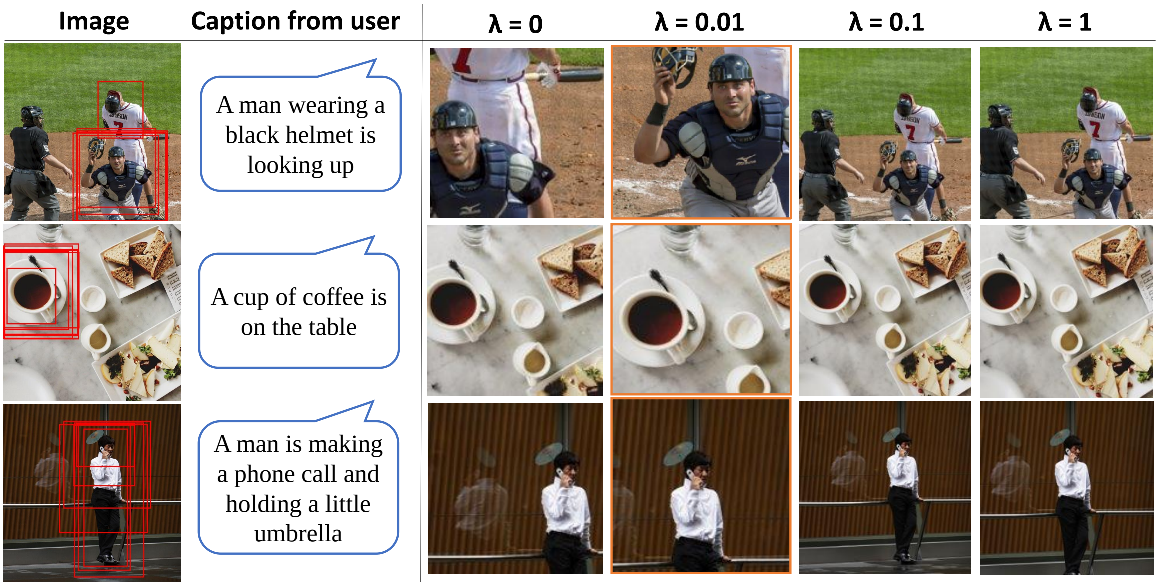


Image
Caption from user
λ = 0
λ = 0.01
λ = 0.1
λ = 1
A man wearing a black helmet is looking up
A cup of coffee is on the table
A man is making a phone call and holding a little umbrella

Figure 2.12: **Comparison of different aesthetics ratios.**

Genome benchmark dataset. Here, we show the qualitative highlights and the quantitative evaluation of the results.

As the dataset is not directly applicable, we randomly select 100 images and a caption from the dataset. We show qualitative results in Figure 2.13 and 2.14. Notice how our results successfully return regions corresponding to the provided captions more naturally than MAttNet. For example, in the first two examples in Figure 2.14, notice how the caption describes the context in which the object is placed, whereas the annotation nor the result of MAttNet have such context – *e.g.* the fire hydrant, the TV. Another example is the fourth example in Figure 2.13, the annotation is tightly bounding the car, whereas the caption also refers to the street. When cropped according to the ground truth, which is what MAttNet successfully did, the context "street" is gone. This further highlights the deficiency of this dataset – it is not suitable for evaluating natural caption-guided crops.

Table 2.4 presents the quantitative results of all the baselines. Note that these results should be understood with care, as they do not directly measure the performance of each method for our task, rather, they measure how well a method locates an object. For the caption-based image cropping task on the Visual Genome dataset [84] MAttNet [47] performed the best and our method was the second best among the baselines. This is because MAttNet is trained to produce tight bounding boxes around the described objects, similar to the ground truth annotations of the dataset. However, as we discussed before, those image crops are often not useful as they do not fully represent the user's caption, or the given user caption could belong to multiple bounding boxes.

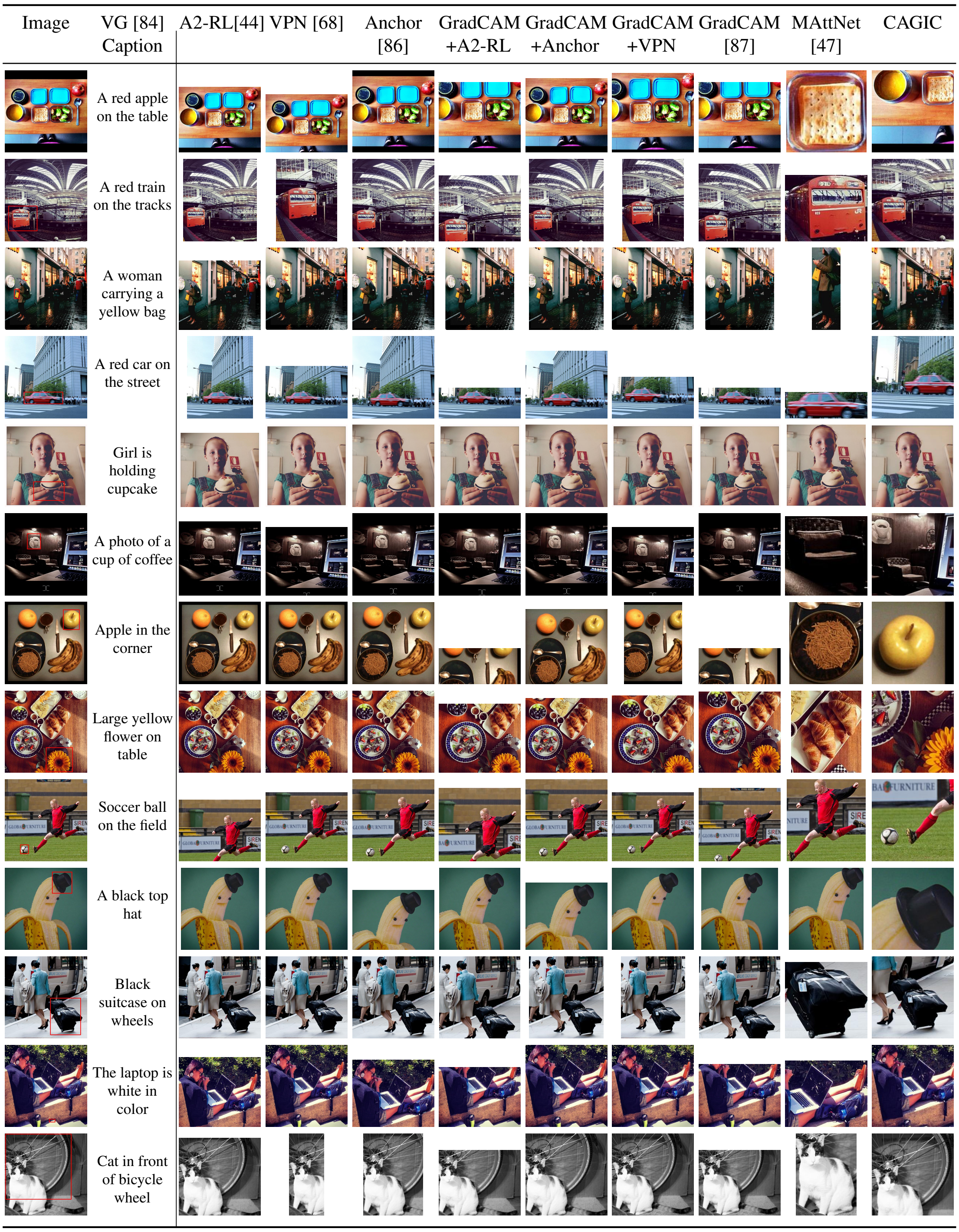


Figure 2.13: **Qualitative examples for the Visual Genome dataset.**

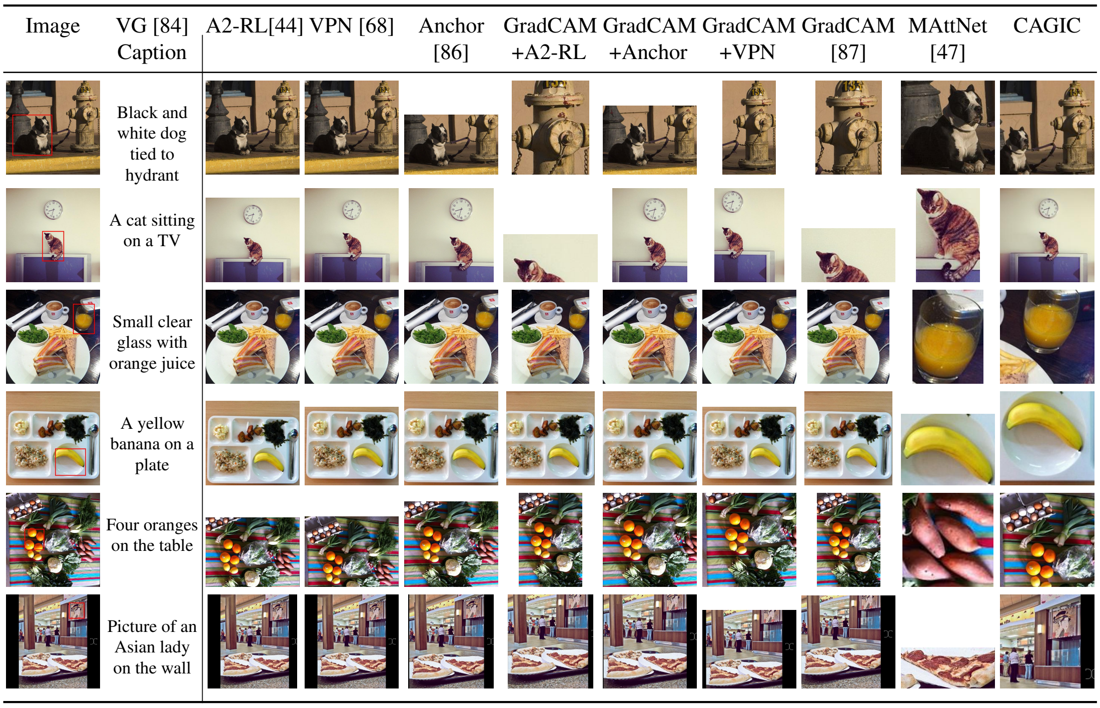


Image
VG [84] Caption
A2-RL[44]
VPN [68]
Anchor [86]
GradCAM +A2-RL
GradCAM +Anchor
GradCAM +VPN
GradCAM [87]
MAttNet [47]
CAGIC
Black and white dog tied to hydrant
A cat sitting on a TV
Small clear glass with orange juice
A yellow banana on a plate
Four oranges on the table
Picture of an Asian lady on the wall

Figure 2.14: **Qualitative examples for the Visual Genome dataset (continued).**

Table 2.4: **Quantitative comparison of the different methods on the Visual Genome dataset [84] using IoU on the output bounding boxes.** Note that these numbers represent object localisation performance, not caption-guided cropping.

Method	Mean ± Std.
Original	0.113 ± 0.136
A2-RL[44]	0.124 ± 0.150
VPN[68]	0.131 ± 0.146
GradCAM+VPN	0.138 ± 0.143
Anchor [86]	0.142 ± 0.161
GradCAM+Anchor	0.146 ± 0.162
GradCAM[87]	0.148 ± 0.146
GradCAM+A2-RL	0.153 ± 0.149
CAGIC	0.218 ± 0.198
MAttNet[47]	**0.471** ± **0.375**

Table 2.5: **Quantitative result of the ablation study using IoU measure on the output bounding boxes.**

Method	Mean ± Std.
CN	0.334 ± 0.069
CN+SA	0.341 ± 0.058
CN+SA+MR	0.356 ± 0.057
CN+SA+MR+MS	0.382 ± 0.062
CAGIC (Full)	**0.416 ± 0.013**

2.4.5 Ablation Study

Framework components

To motivate our design choices, we present a qualitative comparison of our method with various components of our full pipeline disabled; see Figure 2.11.

We compare against five variants. The full method uses the Caption Network (CN) along with Scale Anneal (SA), Multiple Restart (MR), Multi-scale Bilinear Sampling (MS) as well as aesthetics. As shown, we can produce more visually appealing crops of the original image as we introduce aesthetics scores to the system. In the case of CN+SA+MR+MS, the method can find the relevant part of the original image, but the contents are not placed in the centre, as can be noticed in the examples shown in the last two rows. MR helps the approach to centre the content better, while scale annealing prevents drastic scale changes. In short, the full method delivers the best outcome in all cases. It can be observed that our aesthetic-based method gives us the most relevant and best-centred crop of the original image.

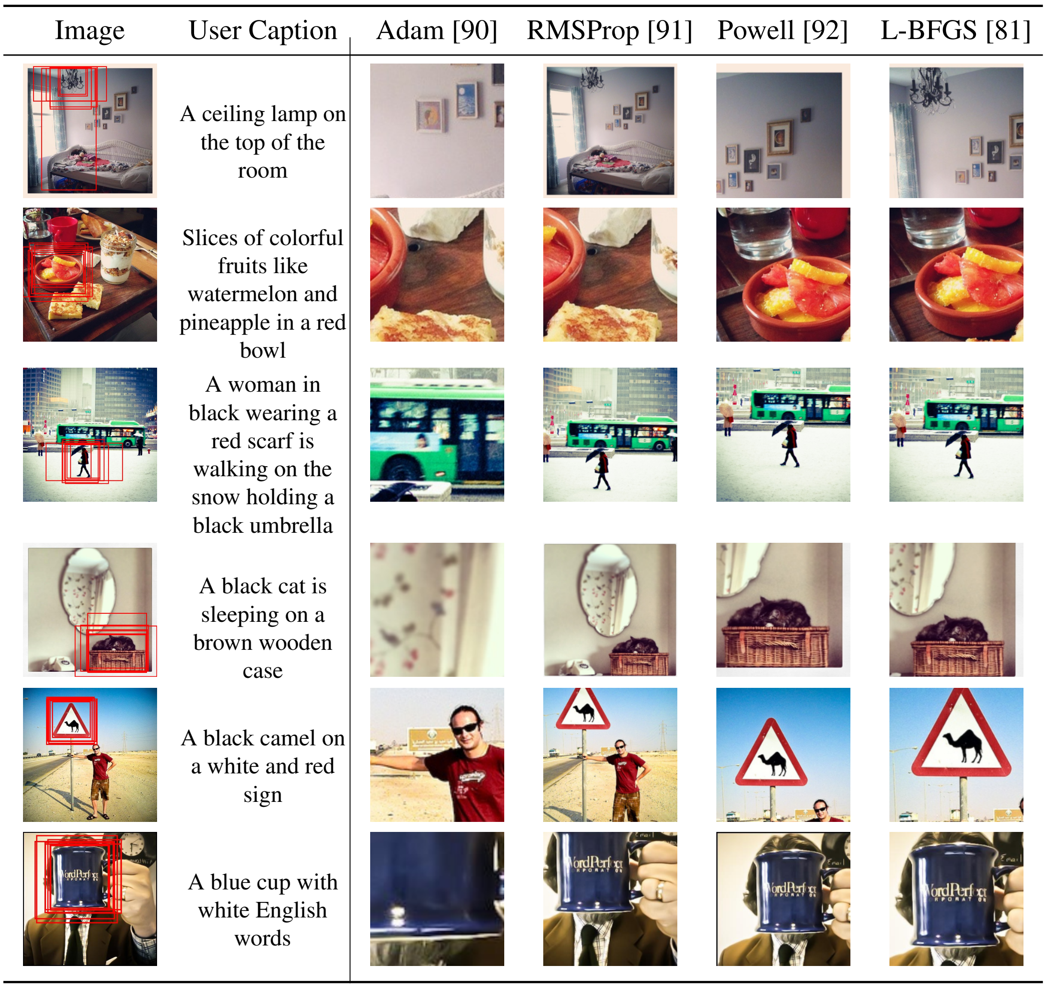
The quantitative results of the ablation study in Table 2.5 show that every component of the proposed method is effective and that using aesthetics on top of the different optimisation and search stabilisation methods improved the results. In addition, we further demonstrate the effectiveness of the aesthetics loss in our framework in Figure 2.12. In our method, the aesthetics ratio (λ) was empirically set to 0.01. As shown, the produced crops are most similar to the ground truth regions when the aesthetics loss is enabled.

Optimisers

Furthermore, as part of our ablation study, we compared the outputs of our method with different optimisers. We motivate our choice of optimiser based on this study. We consider four methods: Adam [90], RMSProp [91], Powell [92], and L-BFGS [81]. The qualitative comparisons of the optimisers are shown in Figure 2.15. We also perform the quantitative comparisons using the IoU measure on the output bounding boxes (See Table 2.6). Based on our experiments, we found out that for our task, L-BFGS method [81] is the best. Note that the runtime was not affected by the type of optimisers used during the experiments.

Table 2.6: **Quantitative results from different optimisers in terms of IoU.**

Method	Mean ± Std.
Adam [90]	0.064 ± 0.066
RMSProp [91]	0.199 ± 0.121
Powell [92]	0.361 ± 0.209
L-BFGS [81]	**0.416** ± **0.013**



A ceiling lamp on the top of the room

Slices of colorful fruits like watermelon and pineapple in a red bowl

A woman in black wearing a red scarf is walking on the snow holding a black umbrella

A black cat is sleeping on a brown wooden case

A black camel on a white and red sign

A blue cup with white English words

Figure 2.15: **Qualitative comparison of different optimisers.**

Pre-defined versus fixed aspect ratio

Our method operates on 1:1 aspect ratio images whose design choice was confirmed by our experimental results discussed in this section. The drawback of extending the method to handle arbitrary aspect ratios is that when optimising for location, scale and aspect ratio simultaneously, the search space becomes so large that it makes convergence difficult. In Section 2.4.3, we compared our method with other state-of-the-art methods which produced rectangular crops using different ranking techniques. Our qualitative and quantitative results show that the proposed method performs best despite the fixed aspect ratio. Therefore, we argue that having an arbitrary aspect ratio would not improve the quality of the image crop.

To support our claim experimentally, we generated candidate output crops, similar to [68], using the following aspect ratios: 16:9, 4:3, 3:2, 1:1, 2:3, 3:4 and 9:16. We ranked the candidate crops based on both their aesthetics score and their $\mathscr{L}_{total}$. Then we measured the IoU score of the best candidates along with the proposed square output crop. Our results show that using the 1:1 aspect ratio results in higher IoU measures (See Table 2.7.).

Table 2.7: **Performance evaluation of the proposed method using different aspect ratios.**

IoU	Mean ± Std.
Aesthetic score	0.390 ± 0.051
$\mathscr{L}_{total}$	0.405 ± 0.053
1:1	**0.416 ± 0.013**

2.4.6 Runtime

Our un-optimised implementation cannot run in real-time — it requires an average of 2.06 seconds per single optimisation iteration. CAGIC requires a total of 200 iterations to produce the desired output crop which takes approximately 6.5 minutes. The runtime of the other deep learning-based methods where the input goes through one pre-trained network is shorter than our optimisation-based method. However, despite our method not being real-time, it was demonstrated that it performs better and it is flexible. Our future work would be to resolve this, similar to how style transfer [93] started off taking minutes per image but is now able to run in real-time [94].

2.5 Conclusion

This Chapter proposes a novel optimisation framework that produces image crops that follow users' descriptions and aesthetic criteria. The main idea behind this study was to demonstrate that this can be achieved without training a specialised network but instead utilising two pre-trained networks on related tasks, namely image captioning and aesthetics measuring. We designed a new cost function considering the two networks' outputs together and performed a search among various bilinear sampling parameters. However, the parameter space is very large and non-convex, so we designed a new scale annealing and multiple restarts search strategy to achieve a stable and efficient solution. We have shown that our proposed method outperforms other existing approaches based solely on saliency-based or caption-based cropping methods.

A big advantage of not requiring training is when we want to alter the model's behaviour later or maybe even add another sub-task to be performed. In this case, a joint-training strategy would require complete re-training of the model, which may not even be a valid option. However, with our method, one has to attach a new loss component to optimise or change the parameterised transform on the input image to include the desired task.

Our ongoing research extends into two directions. The description- and aesthetics-driven captioning will be coupled with attention mechanisms and gaze estimation. On a more conceptual level, we are extending the ideas of solving complex tasks in optimisation frameworks by repurposing existing modular networks trained on auxiliary tasks, thereby seeking efficient alternatives to extensive learning of networks from scratch.

Chapter 3

Gaze Initialized Framework for Description and Aesthetic-based Image Cropping

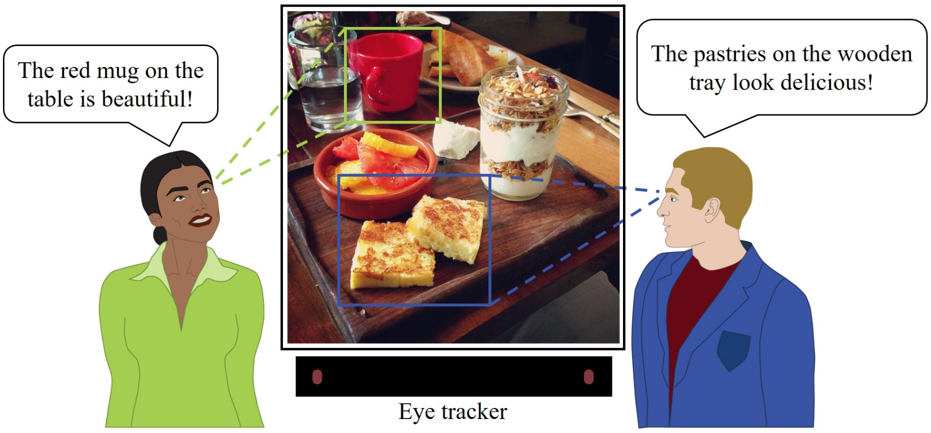

The red mug on the table is beautiful!
The pastries on the wooden tray look delicious!
Eye tracker

Figure 3.1: **Illustration of a scenario where the users are looking at and describing the different image parts.** Our proposed framework utilises human natural language expression in combination with eye movements to localise their region of interest. This multimodal solution can effectively produce high-quality image crops corresponding to the user's intention.

3.1 Introduction

Many researchers throughout the past years have demonstrated the usefulness of multimodal approaches. Regarding speech and description, the existing ambiguity and complexity of natural speech can be compensated by using other modalities, such as hand gestures [95] or gaze [96–98]. While hand gestures or other social cues are not always coupled with the description of the scene, it is inevitable to observe the region of interest to describe it [99].

With the current advancements in eye-tracking research, eye-tracking devices have become more affordable and easily accessible to everyone. Therefore we can take advantage of this extra modality free of cost by placing an eye-tracker in front of the subject and recording their eye movements before and during describing the image region. This way, we can effortlessly collect additional, rich contextual information in addition to the recorded speech or image caption.

Despite its inarguable usefulness of gaze in multimodal communication, this non-verbal mechanism has been less studied as a communication modality [100]. Previous works used fixation points on their own as pointers to implicitly localise the user's region of interest [101] or to detect the object of interest [102]. Recent research [103] demonstrated gaze integration's usefulness in anchoring implicit notes to digital content. Furthermore, Reinholt *et al.* [104] proposed to combine gaze information with speech to identify regions of interest in an image. This assistive system aimed to create detailed descriptions of images with minimal effort from the image creator.

In this work, we tackle the reverse problem of [104], where we obtain the image description, like in Chapter 2, and the gaze information with no additional cost from the user to crop the described part of the image. We propose directly integrating the user's gaze information into a multimodal image-cropping framework to understand their interest better. We expect this additional information to improve the accuracy of the state-of-the-art CAGIC methods and reduce their run time. This framework utilises the gaze information recorded by an eye-tracking device to initialise the iterative search to localise the described area of the image. To the best of our knowledge, this is the first work explicitly tackling the gaze-initialised user description-based image cropping task. Our contributions are four-fold:

- We presented a novel image cropping framework that integrates the user's intentions through explicit (user caption) and implicit (user gaze) input into a multimodal framework optimised to achieve aesthetically pleasing output.
- We studied the usefulness of gaze data collected before, during, and after the caption generation and proposed the *Fixed grid* and *Region proposal* on how to leverage the correlation between them to initialise the image cropping method most efficiently.
- We propose a new multimodal framework to optimise crop parameters adaptively using the novel *Mixed scaling method* and gaze-based initialisation coupled with an *Early termination* technique.
- With the above-mentioned solution, we were able to significantly reduce the run time compared to the state-of-the-art and improve the performance.

Utilising gaze information in a multimodal framework is a complex problem studied by many [105, 106]. We designed the gaze data collection experiment carefully, similarly [104], and performed multiple experiments to confirm the quality of the collected gaze information. We proposed the *Fixed Grid* and *Region proposal* methods to find and define the most looked-at area of the image, which was used as a start region of our optimisation framework.

For the Gaze initialized, description and aesthetics-based image cropping (G-DAIC) method, we modified [33] introduced in Chapter 2 and proposed a *Mixed scaling* method, where based on the size of the initialisation area, the cropping parameter is either shrinking or expanding in every iteration. This proposed alteration is crucial to find the desired image region. Finally, we introduced *Early Termination* into the multimodal framework to reduce the number of iterations required to produce the output image crop.

3.2 Related Works

Related works on image cropping have been discussed in detail in Chapter 2. For more information on the categorisation and existing methods, refer to Section 2.2. In this section, we discuss the existing multimodal and gaze-based image-cropping solutions.

The primary task of the image cropping algorithms is attention-related. The aim of the description-based methods, like [47, 69], is to find the described part of the image, keeping in mind that it might not be the main subject of the image. While aesthetics and attention-based automatic image cropping algorithms, such as [41], focus on localising the most important part and preserving the image's main subject. The attention-based methods can rely on the user's gaze information or artificial attention *e.g.* saliency maps.

3.2.1 Gaze and artificial attention

Gaze-based image cropping

Despite gaze information not being suitable for our task on its own, gaze has been known as a crucial element of our social interactions. This non-audible signal can be interpreted and utilised in many ways; hence, its role has been studied by many during the past decades. Gaze is a well-known primary social clue that can express the subject's emotions [107]. Furthermore, it is a good guide to the subject's interest, and intention [108, 109], and it also functions as a signal for turn-taking [110], and indicator of conversational roles [111, 112]. Gaze in communication is closely related to speech, meaning that in everyday settings, the user has to address speech *w.r.t.* to gaze [100].

Recent eye-tracking technology advancements enabled us to collect high-precision and accurate eye movement data of the subjects [113]. This high-level precision and affordable prices boosted the gaze research field and gaze-based multimodal frameworks. One of the first gaze-based image cropping solutions, a semi-automatic image cropping algorithm using gaze data, has been proposed by Santella *et al.* [60]. This method uses gaze data to identify the important content of the image and generates an image crop based on a set of composition rules. While this method is useful for photo composition, it is not suitable for the description-based image cropping task. This semi-automatic method is designed to identify the main subject of the image. Still, it is not flexible enough to take the user's intention explicitly into account to identify any part of the image.

Artificial attention-based image cropping

The usefulness of artificial attention, such as automatically-generated saliency maps using deep networks, has been demonstrated in attention-based image cropping methods [53, 114, 115]. While these maps are good in indicating the potentially important areas of the image based on its features, they are not sufficient for the description-based image cropping task due to the lack of contextual information. Therefore in this work, we aim to extend [33], and [103] by combining gaze and description information into a multimodal system designed to crop images based on the natural language expression provided by the users.

3.2.2 Multimodal Image Cropping

Fang *et al.* [116] proposed an automatic image cropping method that uses composition, content preservation, and boundary simplicity clues to preserve the image's subject. This work was one of the first learning-based automatic solutions which did not hard code the cropping rules but learned it from online resources. Their study proved that combining different models in one framework can yield much better performance. Following the success of [116], Wang *et al.* proposed a novel deep network solution for attention and aesthetics-aware image cropping, and they also utilised a cascade attention box regression and aesthetic quality classification in [117, 118]. The proposed neural network consists of two branches for predicting attention bounding boxes and analysing aesthetics. This method infers the initial crop as a bounding box covering the visually important area (attention) and then selects the best crop with the highest aesthetic quality from a few cropping candidates generated around the initial crop (aesthetic). Most recently, Horanyi *et al.* [33] proposed a multimodal description and aesthetic-based image cropping framework using explicit user description information in addition to the aesthetics assessment. They re-purposed an existing image caption and aesthetic prediction model into one framework. Through a series of optimisation techniques, they could localise the described part of the image and output an aesthetically pleasing image crop. Through excessive experiments, it was shown that the proposed framework outperforms the other learning-based automatic image-cropping methods. The drawback of this solution is taking a long time to produce an image crop due to the iterative nature of the algorithm.

Gaze and Description-based Image Cropping

As the gaze-based semi-automatic image cropping method does not use any additional clue *w.r.t.*the image context, it is not suitable for description-based image cropping applications. However, it is important to note that gaze has been known as a crucial element of our social interactions, and an important characteristic of gaze in communication is that it is closely connected to speech [100, 119]. Accordingly, an analysis of communication in daily settings has to address speech in relation to gaze. Therefore, it is natural to assume that gaze can be coupled with other modalities, such as user description, for applied computer vision tasks. Our work is the first multimodal, gaze-initialised, and description- and aesthetics-based image cropping solution to the best of our knowledge.

3.3 Methodology

Gaze information could be used as an indicator of the important part of the image. The collected fixation points and the coupling temporal information can tell us which part of the image caught the viewer's eye first and how its attention shifted throughout the recording. We propose to use the gaze-based user attention information to define the most attended area of the input image in addition to the user caption to enhance the performance of the description-based image cropping algorithm.

3.3.1 Gaze-based initialisation of the image cropping

The collected gaze points can directly define the cropping parameters [60] or could be used as complementary input coupled with other modalities in a single multimodal framework. First, from the gathered gaze points, we need to define the centre point of the content and then use predefined rules to obtain the image crops. To find the centre point and the size of the bounding box ($\mathbf{I}_{init}$) around the visually relevant part of the image, we proposed the following two solutions: *Fixed grid* and *Region proposal*.

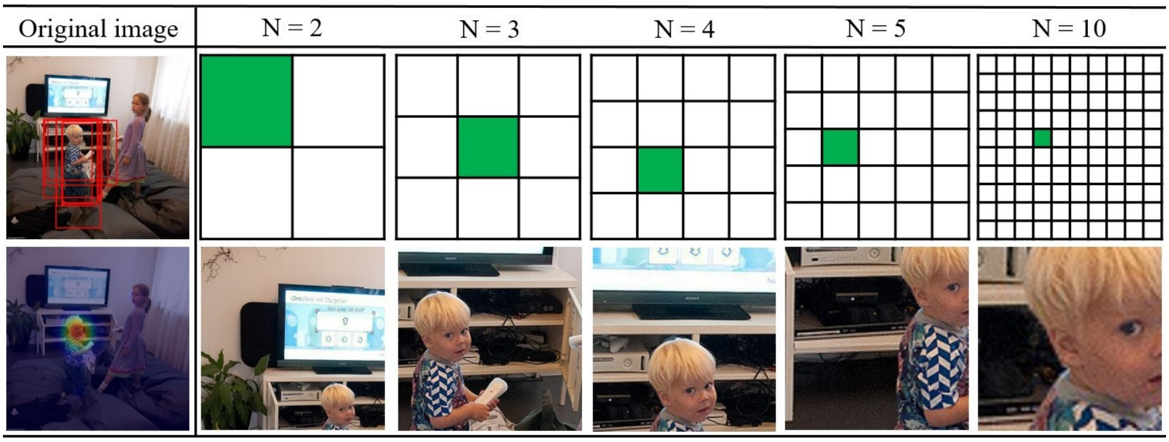


Figure 3.2: **Example outputs of the fixed grid method using different scales.** In the first column, we show the original image, the ground truth bounding box annotations (top), and the heatmap generated using the recorded gaze points (bottom). We visualise the generated $N \times N$ grids, the selected section in green with the highest gaze point count, and the output initialisation region ($\mathbf{I}_{init}$) selected based on the gaze points.

Fixed grid

This solution uses a generated fixed-size grid to divide the image into $N \times N$, uniform grids G_i, and then classify the gaze points ($\mathbf{g}$) into one of the N^2 classes. To select the most attended image region, we counted the number of fixation points within each grid and chose the one with the highest value. Denoted as

$$\mathbf{I}_{init} = argmax(\mathbf{g}(\mathbf{G}_i)), \tag{3.1}$$

where i=1,.., $N \times N$ corresponds to the Grid index. A generated output initialisation region ($\mathbf{I}_{init}$) example visualisation is shown in Figure 3.2. In the first column of the figure, at the top, we show the ground truth bounding box annotations of the image cropping dataset proposed by [33], and at the bottom, the heatmap generated from the collected gaze data. For our experiments, the grids were generated using N=2,3,4,5, and 10, shown in the first row of the figure. In the bottom row of the figure, we show the output of the gaze-based image region selection using the *Fixed grid* method. The qualitative highlights and the quantitative evaluation (See Ablation study, Section 3.4.2, Table 3.5) of the method confirm that the gaze data is useful to initialise the search by roughly localising the centre of the described area; however, it is not flexible enough, hence not suitable to fully preserve the contextual information.

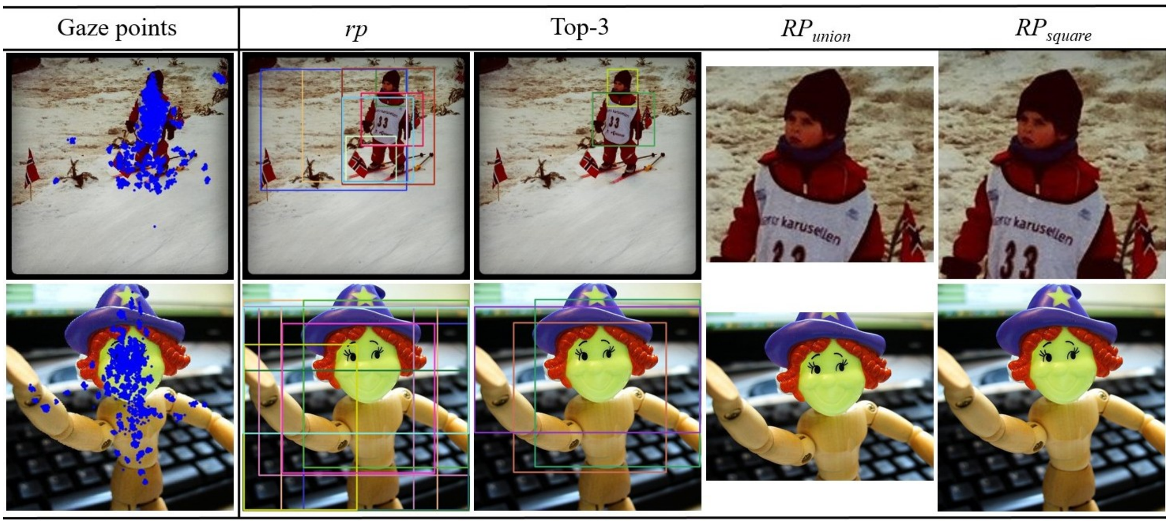


Figure 3.3: **Start region generation from the collected gaze points using the region proposal method.** We show the output region proposal bounding boxes (rp) generated by [120], the selected three boxes with the highest gaze point density (Top-3), and the generated start regions RP_{union} with arbitrary aspect ratio and RP_{square} with 1:1 aspect ratio.

Region Proposal

Furthermore, the results of our ablation study, discussed in Section 3.4.2, show that the selection of N significantly influences the output of the *Fixed grid* method; therefore, to alleviate this problem, we propose exploiting context information by using a region proposal module. This module is implemented by Structured Edge Detector (SED) [120] to get n region bounding boxes ($rp_m, m = 1, ...n$) for each frame. Note that there are large overlaps among the generated rectangles; hence calculating the density is a better measure to identify the most attended image region than counting the number of gaze points. Therefore, for each bounding box, we calculated the gaze point density (d_m) by counting the number of gaze points inside (g_m) and dividing them by the area of the bounding box ($A(rp_m)$).

$$RP_{union} = \bigcup \left(max_3 \{ d_m = \frac{g_m}{A(rp_m)}, m = 1, .., |rp| \} \right), \tag{3.2}$$

where $|rp|$ is the number of bounding boxes generated by [120], d is the density function, and max_3 refers to the top-3 highest value elements of the set. We show an example of this method in Figure 3.3. In the first column, we visualised the collected fixation points; next, we show the generated bounding boxes using SED [120]. Due to the nature of the region proposal algorithm, we selected the three highest-density bounding boxes shown in the third column. To generate the output bounding box for the search initialisation, we merged the

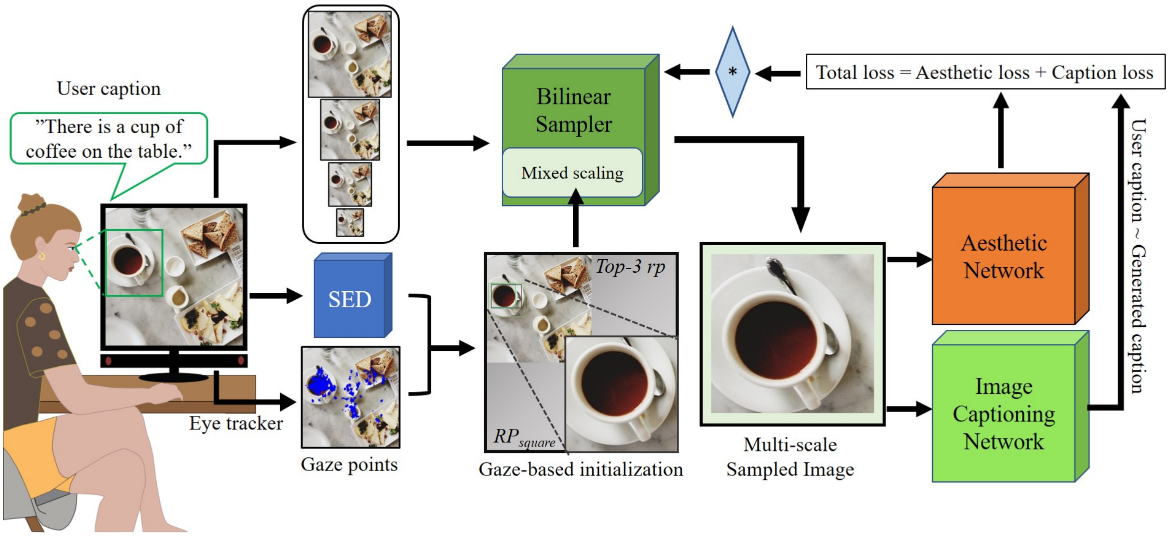

User caption
"There is a cup of coffee on the table."
Eye tracker
SED
Gaze points
Bilinear Sampler
Mixed scaling
Top-3 rp
RP_{square}
Gaze-based initialization
*
Total loss = Aesthetic loss + Caption loss
Multi-scale Sampled Image
Aesthetic Network
Image Captioning Network
User caption ~ Generated caption

Figure 3.4: **Overall framework of the proposed method: G-DAIC.** The search is initialised using the gaze points collected from the subject during the user caption generation and the Region Proposal module (SED [120]). The adaptive *Mixed Scaling* method receives the initialisation region (RP_{square}); meanwhile, the framework takes an image as input, which goes through multi-scale bilinear sampling to produce a cropped image. We then input this cropped region into the image captioning and aesthetic networks. The optimisation ends when the Total loss is below $T_{loss} = 5.23$ threshold by the proposed *Early termination* (*).

selected region by taking the union of the rectangles (fourth column). Then we extended this rectangle of arbitrary aspect ratio into a square (RP_{square}) (last column) initialisation region.

3.3.2 Proposed Framework

The proposed gaze initialised, description and aesthetics-based image cropping framework (G-DAIC) is shown in Figure 3.4. Based on our experimental results, later discussed in Section 3.4.2, we chose the Region Proposal-based RP_{square} method for the gaze-based initialisation.

Mixed Scaling method

Prior methods chose the described part of the image, starting from the full image and iteratively searching for the optimal output crop. This method used a fixed scale and shrunk the sample image crop size every iteration. The gaze-based initialisation depends on the collected gaze points, which are subjective and user dependent. Therefore, the shrinking-only strategy for finding the desired output crop is not optimal. If the initialisation region is too small, the method might not be able to localise the described part of the image. Therefore, we proposed an adaptive scaling strategy based on the size of the generated initialisation area.

Meaning, that based on the size of $\mathbf{I}_{init}$, the algorithm either zooms in (shrink) or zooms out (expand) every iteration with the scale. Mathematically denoted as:

$$s_{mixed} = \begin{cases} +0.98 \text{ (shrink)} & \text{if } \frac{A(\mathbf{I}_{init})}{A(\mathbf{I})} > T \\ -0.98 \text{ (expand)} & \text{otherwise} \end{cases}, \tag{3.3}$$

where A($\mathbf{I}_{init}$) is the size of the gaze-based input initialisation region, A($\mathbf{I}$) is the size of the input image, T is the threshold and s_{mixed} is the scale. The threshold $T = 0.75$ was selected empirically as part of our ablation study (See Section 3.4.2, Table 3.7). This new scaling strategy is a crucial part of the initialised search algorithm as the size of the described part of the image and the calculated initialisation region varies based on the image content, the image caption, and the user's search behaviour.

Iterative optimisation

Once we calculated the RP_{square} and s_{mixed} we input these along with the original image into the *Bilinear Sampler* [49]. This module generates a multi-scale sampled image based on the input information in every iteration. The proposed sample image is chosen based on the crop parameter θ, which is composed of the centre coordinates of the crop x and y, and its scale s_{mixed}. A pre-trained Aesthetic Network [46] is used to generate the *Aesthetic loss* of the sample image, which reflects on the quality of the current image sample. Furthermore, we used an Image Captioning Network [50] to calculate the *Caption loss* from the user caption and the caption generated from the sampled image. The *Total loss* ($\mathscr{L}_{total}(\mathbf{I}, \mathbf{y}, \theta)$) was calculated as the sum of these two loss functions. The optimisation is performed iteratively to minimise $\mathscr{L}_{total}$ until we find the optimal output crop which best reflects the user's intention.

Early Termination

By using gaze-based initialisation, we have a better idea of where the described part of the image might be. Hence, it is reasonable to assume that the method will require fewer iterations (N_{iter}) to find the optimal output image crop. To further address the runtime limitations of [33], we proposed to use an empirical threshold, $T_{loss} = 5.23$, to terminate the iterations early (*Early termination*) after n iterations ($n_{iter} < N_{iter}$) in case the total loss ($\mathscr{L}_{total}(\mathbf{I}, \mathbf{y}, \theta)$) was below a given threshold. This module was integrated into the iterative optimisation cycle.

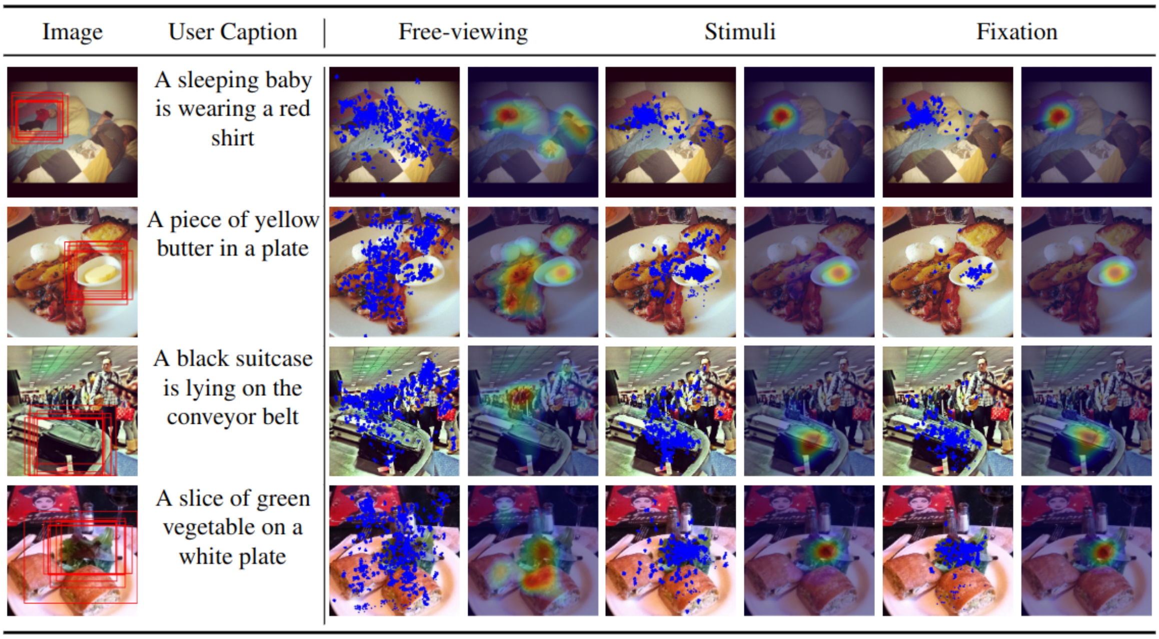


Image
User Caption
Free-viewing
Stimuli
Fixation
A sleeping baby is wearing a red shirt
A piece of yellow butter in a plate
A black suitcase is lying on the conveyor belt
A slice of green vegetable on a white plate

Figure 3.5: **Example images of the extended, multimodal dataset.** The user-defined ground truth bounding box annotations are shown on the original images in red. We show the collected Free-viewing, Stimuli and Fixation gaze points and corresponding heatmaps.

3.4 Experiments

We implement our method in Tensorflow [82]. All experiments are run on an Intel i7- CPU @3.40GHZ, 16 GB RAM, and two NVIDIA TITAN Xp GPU for fair runtime comparison with [33]. The eye-tracking data was collected using a monitor-mounted Tobii Pro Fusion Eye Tracker device and the Tobii Pro Lab v1.145 software.

3.4.1 Multimodal Dataset

The dataset proposed in [33], introduced in Section 2.4.1, was extended with gaze data, shown in Figure 3.5. To analyse the correspondence between the two modalities, gaze and caption, we recorded the eye movements of the participants before and while they performed the caption-based image part localisation task. All data generated during this study are included in Section B.2.

Data collection

Experimental setting. For the data collection, we invited 14 participants to participate in our experiments. Every participant attended our experiment ten times to ensure that they did

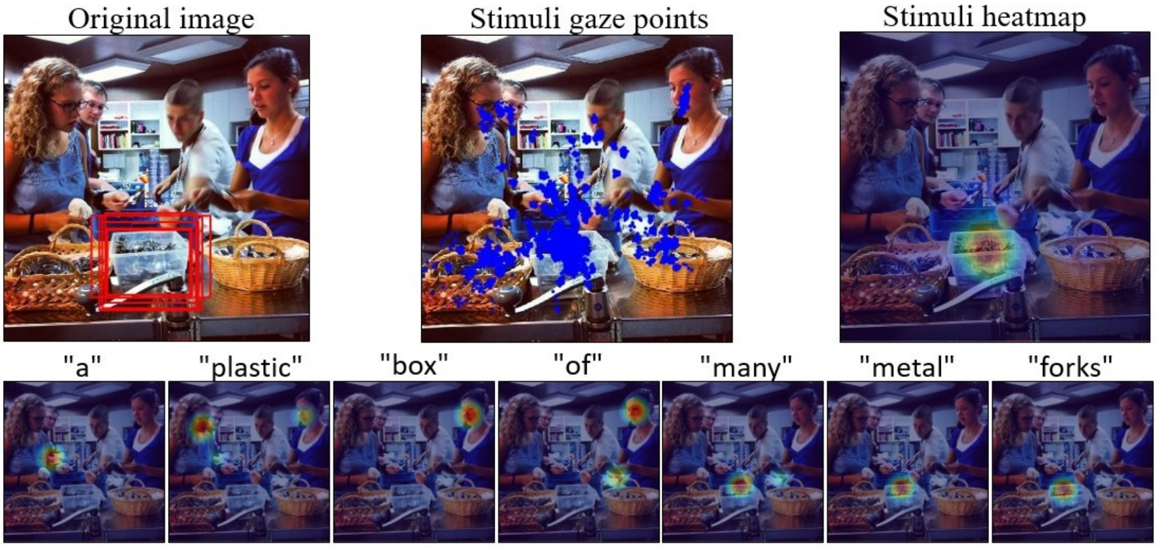


Figure 3.6: **Example visualisation of the eye movements over time during the Stimuli stage.** At the top, we show the original image with the ground truth bounding boxes, the collected gaze points during Stimuli and the corresponding Stimuli heatmap. Below we visualise the gaze points word by word over time.

not get exhausted during the recording, and during each session, the participant observed ten images displayed on the monitor in front of them. The Tobii Pro Fusion eye-tracking device was mounted to this monitor, and before every session, it was calibrated for the user. All the experiments were performed in a laboratory with controlled lighting conditions. The data collection had three stages *Free-viewing*, *Stimuli*, and *Fixation*.

Collected Gaze points. For each image, first, the users observed the image without any given instruction for 10 seconds (*Free-viewing*). We recorded the participants' eye movements during this experiment while they freely observed the previously unseen image. Without instructions, the participants naturally observe the image and spend more time on complex or interesting image parts. Following this stage, we played a recording of the corresponding image caption to the users from [33]. During the second experiment stage *Stimuli*, the participants got to know the contextual information and were asked to follow our instructions. In this phase, the participants were asked to localise the described part of the image while listening to the image caption recording. In the experiment's final *Fixation* stage, they were asked to fixate on the region of interest for 5 seconds. Using the collected gaze points, we generated heat maps corresponding to the image caption. Note that during this stage, the participants were instructed to fixate on the described image part, which is not a single point but an image region; therefore, it is expected that the gaze points will be within a certain area but not limited to a single point.

Table 3.1: **The proportion of detected gaze points inside the ground truth bounding box during Free-viewing, Stimuli and Fixation.**

GT	Free viewing	Stimuli	Fixation	Mean ± std
1	58.78 ± 29.91	87.75 ± 17.91	92.42 ± 19.91	79.65 ± 22.58
2	56.25 ± 32.44	83.75 ± 22.77	90.31 ± 23.31	76.77 ± 26.18
3	54.50 ± 31.22	84.37 ± 21.24	90.37 ± 22.77	76.41 ± 25.07
4	57.67 ± 31.46	85.34 ± 21.17	91.48 ± 22.21	78.16 ± 24.95
5	59.12 ± 29.70	86.96 ± 18.51	92.51 ± 19.94	79.53 ± 22.72
6	50.01 ± 32.69	79.41 ± 26.50	85.26 ± 28.97	71.59 ± 29.39
7	49.04 ± 29.83	81.46 ± 22.63	88.28 ± 25.11	72.92 ± 25.86
Mean ± std	55.06 ± 31.28	84.15 ± 21.86	90.09 ± 23.47	

During Stimuli, the participants listened to the image caption and were asked to localise the described part of the image. Naturally, this is a dynamic part of our experiment. We extended the dataset with a per-word time annotation to study the saccades and divided the recorded gaze points word by word. This new annotation allowed us to investigate how the gaze points shifted between different image regions over time. Figure 3.6 shows an example and the recorded gaze points over time.

Participant information. The participants of the eye-tracking experiment were selected to be diverse in terms of their country of origin, age, sex, and visual acuity. The participants were aged 23-29, seven males and seven females from 6 different countries. Five of them had perfect vision, five of them wore glasses, and four participants used contact lenses during the recordings.

Dataset analysis

We performed multiple experiments to evaluate the quality of the collected gaze points. First, we analysed how well the gaze points correspond to the dataset's ground truth bounding box annotations by calculating the proportion of the points inside these bounding boxes. This is an important measure as it influences the quality of the gaze-based initialisation, hence the overall framework's accuracy. Then we investigated the dynamic nature of the collected Stimuli gaze data. We aimed to understand the participants' eye movements *w.r.t.*the words of the image caption.

Gaze points *w.r.t.*bounding box annotations. We evaluated the different stages of the eye tracking recording individually and compared the recorded points *w.r.t.* the seven ground truth bounding boxes of the dataset. In this experiment, we used the recorded gaze points from every participant. In Table 3.1, we can see that the proportion of the gaze points inside the bounding boxes is very similar for all three stages of the recording. Furthermore, our results show that over 71% of the gaze points were within the bounding boxes for every ground truth bounding box.

The lowest percentage of gaze points inside the ground truth bounding boxes belongs to the Free-viewing stage of the experiment, which is not surprising as the participants were allowed to observe the image without any contextual constraints during this stage. During the Stimuli stage, the number of gaze points within the target area increased by more than 29%, reaching nearly 84% accuracy. Overall, this is a high percentage considering that the proportion was about 90% high during the Fixation stage. In a real-world scenario, users do not tend to fixate on the described part of the image after providing the description. Therefore, while the Fixation points are more aligned with the ground truth bounding box annotations of the dataset, we used the Stimuli points in our experiments for initialisation.

Table 3.2: **The proportion of detected gaze points inside the ground truth bounding boxes *w.r.t.* the subjects during Free-viewing, Stimuli and Fixation.**

User	Free viewing	Stimuli	Fixation	Mean ± std
1	55.98 ± 32.63	87.73 ± 17.65	92.72 ± 19.64	78.81 ± 23.30
2	52.48 ± 28.91	89.27 ± 15.63	94.67 ± 17.05	78.81 ± 20.53
3	53.39 ± 33.01	87.97 ± 18.28	94.18 ± 19.97	78.51 ± 23.76
4	59.34 ± 34.73	76.84 ± 30.38	80.87 ± 33.11	72.35 ± 32.74
5	52.45 ± 30.42	84.70 ± 19.55	86.36 ± 25.05	74.50 ± 25.01
6	47.89 ± 28.73	87.18 ± 18.15	94.67 ± 19.58	76.58 ± 22.15
7	54.84 ± 30.53	86.84 ± 18.72	93.93 ± 18.68	78.54 ± 22.65
8	54.62 ± 30.09	83.11 ± 25.92	89.95 ± 27.36	75.89 ± 27.79
9	57.96 ± 31.37	83.57 ± 22.24	92.23 ± 22.35	77.92 ± 25.32
10	66.00 ± 33.92	86.04 ± 22.10	93.03 ± 23.61	81.69 ± 26.54
11	54.58 ± 29.22	80.87 ± 25.03	88.27 ± 23.59	74.57 ± 25.95
12	57.84 ± 31.33	81.77 ± 23.34	90.39 ± 21.61	76.66 ± 25.43
13	54.46 ± 31.04	82.70 ± 20.62	89.39 ± 23.04	75.52 ± 24.90
14	55.92 ± 31.57	78.44 ± 23.64	79.90 ± 27.02	71.42 ± 27.41
Mean ± std	55.06 ± 31.28	84.15 ± 21.86	90.09 ± 23.47	

Gaze points *w.r.t.*participants. It is important to note that the participants' data had some disagreements, similar to the subjective nature of the user-annotated bounding boxes of the dataset. Therefore, we experimented to better understand the subjective nature of the gaze

Table 3.3: **Word by word temporal behaviour analysis of the collected Stimuli gaze points.** The average percentage is calculated by computing the percentage of the gaze points inside every ground truth bounding box annotation and taking their average. The heatmaps corresponding to each word are visualised in Figure 3.6.

Word	Gaze inside (%)
A	0
Plastic	0
Box	0
Of	0
Many	67.03
Metal	99.34
Forks	100

points. The results of this experiment are shown in Table 3.2. We can observe that some subjects, like User 10, had higher accuracy during all three stages of the experiment than others. However, the tendency among the stages is the same for every user, and the overall percentages are close to each other too.

Temporal information. The eye movements and the description are related and aligned [121]. During the image description, the observer shifts its attention over time as they receive more information regarding the described part of the image. One way to analyse the nature of this transition is by splitting the sentence into words and checking where the subject was looking when the word was used. Therefore, we analysed the gaze points collected during the Stimuli stage using temporal information and related the points to each word of the given image caption. This way, we obtain a linearly ordered sequence of fixation locations encoded using the word and gaze recording timestamps. We collected the length of each word in the caption recordings and visualised the gaze points *w.r.t.*the words, as shown in Figure 3.6. This figure demonstrates the eye movements of the users during the Stimuli stage.

We calculated the gaze inside score (GIS) over time for quantitative analysis of the dynamic behaviour of the eye movements. The word-by-word GIS is calculated as follows:

$$GIS = \frac{\sum\limits_{i=\{1,\ldots,m\}} \frac{gw_i}{gw}}{m}, \tag{3.4}$$

where gw is the number of gaze points recorded during the word, gw_i number of gaze points within $i = \{1,...,m\}$ ground truth bounding boxes of the dataset, where $m = 7$ is the number of ground truth bounding boxes. In Table 3.3, we show the GIS for each word of the caption corresponding to the example shown in Figure 3.6. We performed this analysis across the

Table 3.4: **Quantitative comparison of the heatmaps generated based on human gaze data and by GradCAM [87].**

AUC (Mean ± Std.)	$GradCAM_{no\ caption}$	$GradCAM_{caption}$
Free-viewing	0.556 ± 0.119	0.545 ± 0.132
Stimuli	0.572 ± 0.201	0.620 ± 0.192
Fixation	0.567 ± 0.219	**0.633 ± 0.206**

whole dataset and found that the GIS increased over time. The calculated GIS of every user caption is included in G-DAIC dataset.

Furthermore, based on the computed GIS, we calculated the normalised word index (NWI) of each caption, defined as:

$$NWI = \frac{argmax(GIS_i)}{w}, \tag{3.5}$$

where w refers to the caption length and GIS_i is the calculated GIS of the i-th word of the caption. In other words, NWI calculates the position of the word with the highest GIS in the caption. The normalisation was necessary due to the varying caption length. Note that when multiple GIS scores are equal, the NWI returns the first appearance of the maximum value. The average NWI across the dataset is 0.65. This value indicates that the first time the subject concentrates their attention on the described part of the image typically occurs during the second half of the caption. This means that while the captions' last words have the maximum gaze inside score of 94.65 % of the time, the first occurrence of the subject's looking at the described part of the image happens earlier during the caption generation. This finding shows that using the gaze points collected during Stimuli (entire caption) is beneficial; however, based on the temporal information, irrelevant gaze points could be removed during the start region generation. While this result demonstrates a strong correspondence between the collected gaze points and the image caption, the nature of the GIS increase is highly context-dependent; therefore, it would require more investigation on efficiently utilising the temporal information for G-DAIC.

3.4.2 Ablation Study

Human gaze-based versus artificial attention heatmaps. In this analysis, GradCAM [87] was used in two ways to create a coarse localisation map (artificial attention heatmap) that spotlighted essential areas in the image for predicting the concept. $GradCAM_{no\ caption}$ was used without providing the user caption to the model as part of an image captioning.

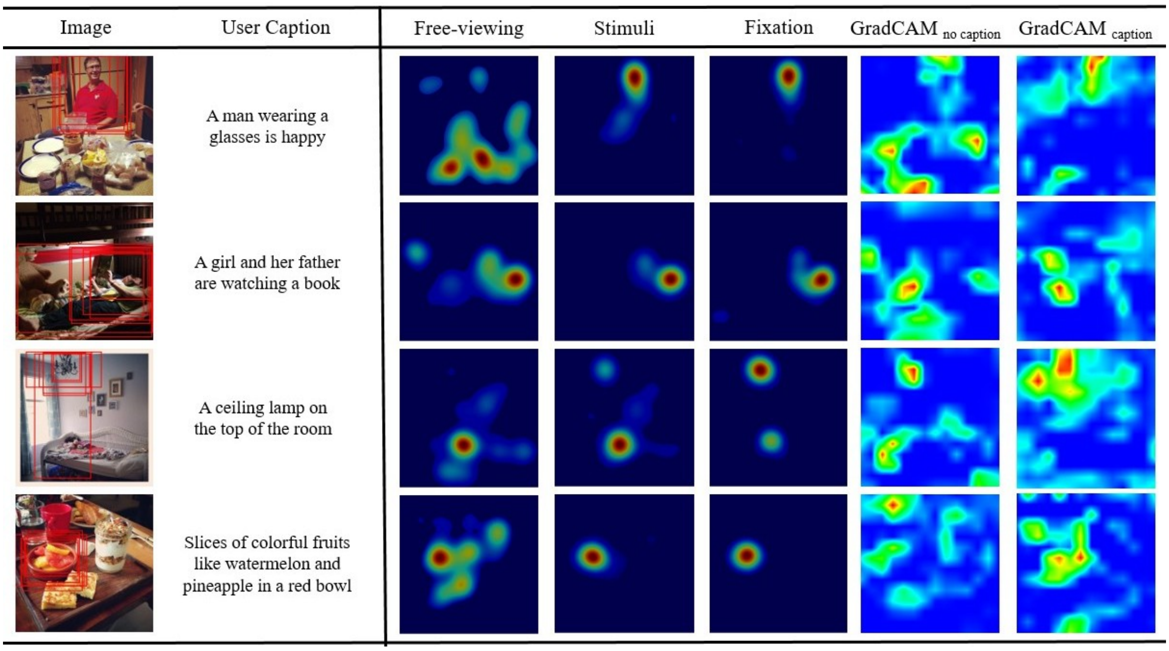


Figure 3.7: **Gaze-based and artificial attention heatmap comparison.** Illustration of the heatmaps generated based on the collected gaze points from the users compared to the artificial attention heatmaps generated by with ($GradCAM_{caption}$) and without ($GradCAM_{no\ caption}$) providing the user caption to GradCAM [87].

$GradCAM_{no\ caption}$ highlighted the image's salient features, which could be used to describe the full image. Note that this image captioning model is designed to describe the entire image, which is different from the user captions. The implementation of this model is available at [1]. $GradCAM_{caption}$ was used as introduced in Section 2.4.2; it was used along with a CN to extract the parts in the image corresponding to the user caption.

Figure 3.7 shows example heatmaps corresponding to the collected gaze points and the generated artificial attention heatmaps. We can see that the artificial attention heatmaps are sparse and poorly aligned with the ground truth bounding box locations compared to the gaze-based heatmaps. In an attempt to quantitatively measure the similarity among these heatmaps, we calculated the average AUC scores across the dataset between the heatmaps generated based on the recorded gaze points (Free-viewing, Stimuli and Fixation) and the artificial attention heatmaps. The average AUC scores are summarised in Table 3.4. We found that the GradCAM heatmaps were generated without the caption, which corresponded to the free-viewing situation when the users were unaware that the caption corresponded vaguely to the human gaze-based heatmaps. $GradCAM_{caption}$ made more similar predictions to the

[1] https://github.com/ramprs/grad-cam#image-captioning

Table 3.5: **Quantitative comparison of the different start region generation methods using IoU measure (Mean ± Std.) on the output bounding boxes.** The gaze data used in this comparison was collected during Stimuli.

Method	$GradCAM_{nocaption}$	$GradCAM_{caption}$	Gaze data
$Fixed_{N=10}$	0.013 ± 0.003	0.013 ± 0.004	0.049 ± 0.013
$Fixed_{N=5}$	0.056 ± 0.012	0.059 ± 0.010	0.171 ± 0.007
$Fixed_{N=4}$	0.080 ± 0.011	0.094 ± 0.014	0.219 ± 0.017
$Fixed_{N=3}$	0.119 ± 0.012	0.150 ± 0.007	0.280 ± 0.005
$Fixed_{N=2}$	0.179 ± 0.009	0.223 ± 0.007	0.326 ± 0.006
RP_{union}	0.277 ± 0.009	0.290 ± 0.012	0.355 ± 0.009
RP_{square}	0.284 ± 0.010	0.294 ± 0.012	**0.361** ± **0.008**

human heatmaps. The maximum AUC achieved was 0.63, which corresponded to the fixation map.

Different initialisation techniques

In Table 3.5, we compare how the different start regions (See more in Section 3.3.1) correlate to the described part of the image. The start regions were generated based on the collected gaze points during Stimuli and the output heatmap of $GradCAM_{nocaption}$ and $GradCAM_{caption}$. Without performing any optimisation or image cropping, we compare the proposed image regions with the ground truth annotations of the dataset using the Intersection over Union (IoU) measure.

The presented results show that human gaze-based initialisation results in start regions that correlate more to the desired image region than the artificial attention-based initialisation. This tendency was present regardless of the type of initialisation method. Furthermore, aligned with the results presented in Table 3.4, the calculated IoU score of the start regions generated by $GradCAM_{caption}$ was higher than the ones generated by $GradCAM_{nocaption}$.

The results of this experiment show that the gaze data is useful to initialise the framework; however, more is needed for the description and aesthetics-guided image-cropping task. Furthermore, we found that the Region Proposal-based RP_{square} initialisation method was the most reliable for start region generation, per our findings in Section 2.4.5.

Human gaze and artificial attention-based initialisation

We evaluated the performance of the G-DAIC framework using artificial attention and human gaze-based initialisation to compare their usefulness. Based on the results presented in

Table 3.6: **Quantitative comparison of the different attention information for start region definition using IoU measure (Mean ± Std.) on the output bounding boxes.** To obtain these results, we used the proposed Mixed scaling method for every experiment.

Attention type	$GradCAM_{caption}$	Free-viewing	Stimuli	Fixation
$Fixed_{N=10}$	0.022 ± 0.005	0.042 ± 0.016	0.075 ± 0.018	0.073 ± 0.015
$Fixed_{N=5}$	0.084 ± 0.013	0.146 ± 0.015	0.242 ± 0.007	0.237 ± 0.007
$Fixed_{N=4}$	0.140 ± 0.018	0.214 ± 0.017	0.294 ± 0.011	0.289 ± 0.013
$Fixed_{N=3}$	0.272 ± 0.020	0.271 ± 0.013	0.365 ± 0.012	0.349 ± 0.013
$Fixed_{N=2}$	0.214 ± 0.013	0.246 ± 0.015	0.328 ± 0.015	0.327 ± 0.018
RP_{union}	0.298 ± 0.013	0.332 ± 0.013	0.369 ± 0.018	0.368 ± 0.014
RP_{square}	0.307 ± 0.021	0.309 ± 0.013	**0.433 ± 0.011**	0.408 ± 0.013

Table 3.4 and 3.5, we used the artificial attention heatmaps generated by $GradCAM_{caption}$ to initialise our framework. Note that gaze point collection from the users during Free-viewing and Stimuli does not require additional effort from the user; therefore, it is unobtrusive.

Gaze information from different stages. In Table 3.6, we report the IoU scores when the Mixed scaling method was performed using different gaze data (See Section 3.4.1). Namely, we compare the differences when using Free-viewing, Stimuli, and Fixation information in the optimisation framework. In this comparison, we found that using the gaze points from the Stimuli stage results in the highest similarity with the human ground-truth annotations of the dataset. Furthermore, regardless of the chosen gaze-based initialisation technique, we can observe that the computed IoU scores are the lowest when relying on the gaze points collected during Free-viewing. We found that the algorithm's performance using the Stimuli and Fixation information for initialisation is very similar. This is because the users were aware of the contextual information during both stages of the experiment. The difference between them might come from the fact that while the Stimuli stage is shorter and more active during the Fixation stage, the participants' attention could wander around over time, potentially influencing the initialisation by introducing Free-viewing like gaze points.

Human versus artificial attention. Finally, Table 3.6 compared also the performance of the G-DAIC framework initialised by human and artificial attention. The results invariably confirmed that the artificial attention-based initialisation ($GradCAM_{caption}$) performed worse than when we used the gaze-based initialisation. Note that the caption-aware $GradCAM_{caption}$ initialisation proved to yield lower IoU scores compared to even the Free-viewing gaze points-based output crops. This result aligns well with our qualitative (Figure 3.7) and quantitative (Table 3.4) findings regarding the artificial attention-based start regions.

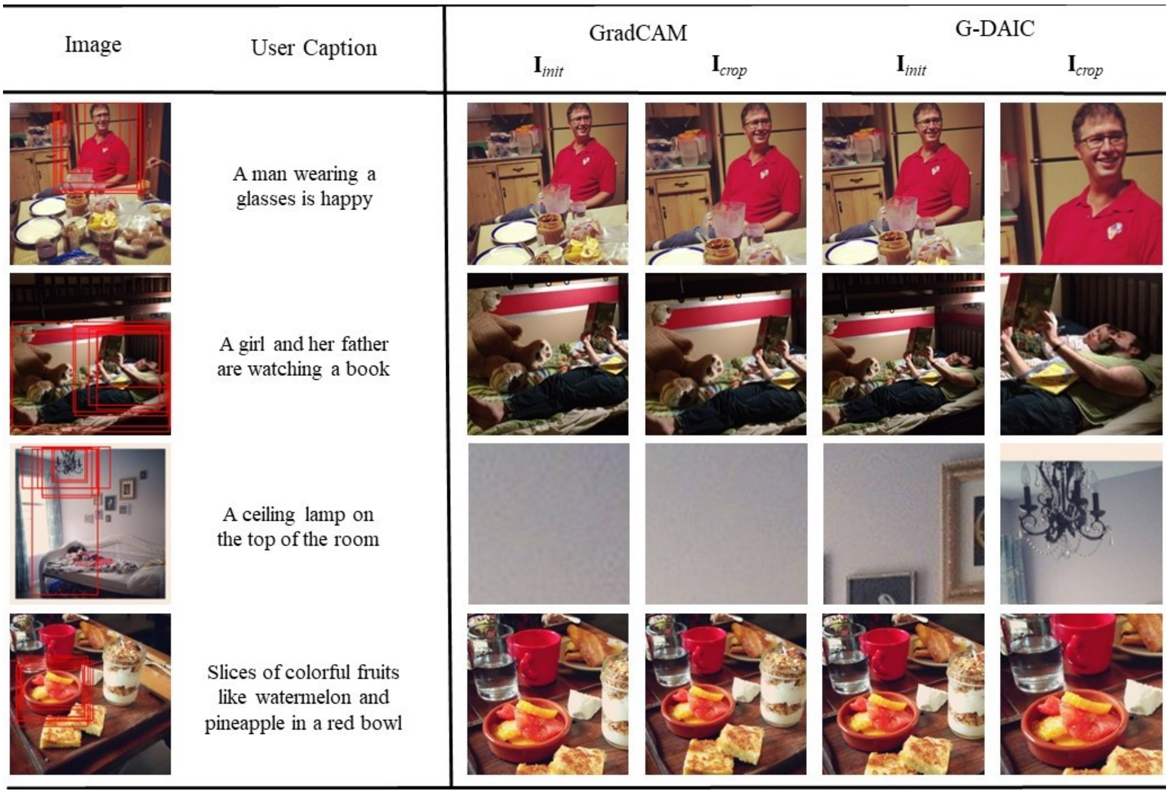


Image
User Caption
GradCAM
G-DAIC
$\mathbf{I}_{init}$
$\mathbf{I}_{crop}$
$\mathbf{I}_{init}$
$\mathbf{I}_{crop}$
A man wearing a glasses is happy
A girl and her father are watching a book
A ceiling lamp on the top of the room
Slices of colorful fruits like watermelon and pineapple in a red bowl

Figure 3.8: **Qualitative comparison of the human gaze and artificial attention heatmap initialised output crops.** Example illustration of the generated gaze and $GradCAM_{caption}$ -based I_{init} regions and the I_{crop} output crops.

Qualitative highlights of this experiment are shown in Figure 3.8. In this figure, we show the generated RP_{square} start regions based on the collected Stimuli gaze points and the heatmap of $GradCAM_{caption}$ and the final output crops of the proposed method. The qualitative comparison shows that the human gaze data collected during Stimuli is more useful and preferable for start region generation than the artificial attention heatmaps.

Different Mixed Scaling method thresholds

Furthermore, we analysed the performance of the Mixed scaling method using different thresholds. We used a 0.75 threshold level in our experiments, as mentioned in Section 3.3.2. This threshold level was chosen empirically based on the results of our experiments presented in Table 3.7. In this table, we show the performance of the fixed grid and region proposal-based gaze initialisation methods using the Mixed method with different threshold levels.

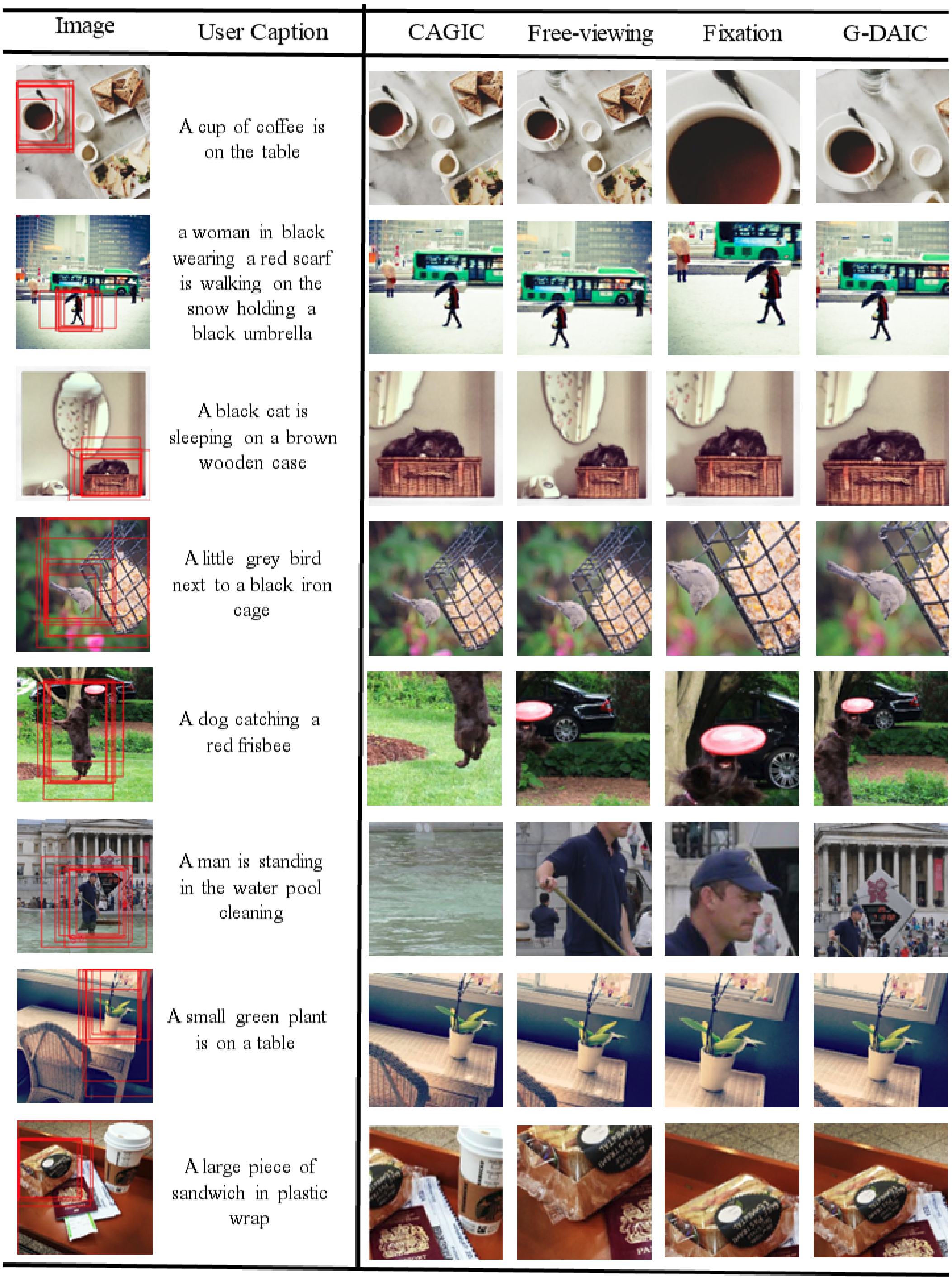


Image
User Caption
CAGIC
Free-viewing
Fixation
G-DAIC
A cup of coffee is on the table
a woman in black wearing a red scarf is walking on the snow holding a black umbrella
A black cat is sleeping on a brown wooden case
A little grey bird next to a black iron cage
A dog catching a red frisbee
A man is standing in the water pool cleaning
A small green plant is on a table
A large piece of sandwich in plastic wrap

Figure 3.9: **Qualitative comparison highlights.** The cropped images obtained by CAGIC, two baseline methods using Free-viewing and Fixation-based gaze initialisation and the proposed method (G-DAIC) using Stimuli-based initialisation. The user-defined ground truth bounding box annotations are shown on the original images in red. The proposed method well crops the images as the user described.

Table 3.7: **Quantitative comparison of the different thresholds (T) of the Mixed scaling method using IoU measure (Mean ± Std.) on the output bounding boxes.**

T	0.125	0.250	0.375	0.500	0.625	0.750	0.875
$Fixed_{N=10}$	0.076 ± 0.025	0.070 ± 0.016	0.075 ± 0.017	0.085 ± 0.007	0.075 ± 0.019	0.075 ± 0.018	0.073 ± 0.019
$Fixed_{N=5}$	0.152 ± 0.011	0.243 ± 0.008	0.241 ± 0.008	0.246 ± 0.004	0.240 ± 0.008	0.242 ± 0.007	0.242 ± 0.005
$Fixed_{N=4}$	0.195 ± 0.018	0.289 ± 0.015	0.301 ± 0.008	0.300 ± 0.012	0.296 ± 0.009	0.294 ± 0.011	0.289 ± 0.010
$Fixed_{N=3}$	0.259 ± 0.007	0.258 ± 0.010	0.357 ± 0.011	0.371 ± 0.010	0.369 ± 0.010	0.365 ± 0.012	0.367 ± 0.013
$Fixed_{N=2}$	0.305 ± 0.014	0.301 ± 0.009	0.298 ± 0.014	0.314 ± 0.011	0.317 ± 0.017	0.328 ± 0.015	0.311 ± 0.019
RP_{union}	0.352 ± 0.016	0.362 ± 0.014	0.377 ± 0.014	0.361 ± 0.017	0.378 ± 0.021	0.369 ± 0.018	0.363 ± 0.020
RP_{square}	0.359 ± 0.007	0.368 ± 0.009	0.382 ± 0.007	0.419 ± 0.010	0.428 ± 0.011	**0.433 ± 0.011**	0.431 ± 0.012

3.4.3 Comparison with the state-of-the-art

Qualitative evaluation

In Figure 3.9, we show qualitative output examples produced by CAGIC [33], the Mixed initialisation method using Free-viewing and Fixation gaze points, and finally, the proposed method G-DAIC. In this figure, we demonstrate that using gaze data from different stages of the eye-tracking experiment in the Mixed initialisation method results in very different image crops. We can observe that initialising the optimisation framework based on the Free-viewing gaze information often returned a larger region of the image as the output image crop. Opposite to the Free-viewing output when the initialisation was based on the Fixation, the image crops tend to be tightly cropped around the subject. Overall, the crops generated by G-DAIC correspond to the captions and are aesthetically pleasing. Compared to the images of CAGIC, we found that the subjects of the caption were more centralised, and we could crop tight enough around the described image region without losing the contextual information provided by the user caption.

Quantitative evaluation

As part of our ablation study in Section 3.4.2, we compared the quantitative performance of the proposed method using different initialisation methods, threshold levels of the proposed Mixed Scaling method, and different types of gaze information. In this section, we present the experiment's results using different scaling methods (Shrink, Expand, and Mixed) and the outcome of the user studies.

Different scaling methods. In Table 3.8 we show the IoU scores *w.r.t.*different scaling methods. Note that during this experiment, we used the gaze points collected during the stimuli stage based on our ablation study's findings. CAGIC [33] iteratively zooms into the described part of the image (Shrink) and does not have gaze-based initialisation (See more in Section 3.3). When the image cropping framework is initialised, we use three scaling

Table 3.8: **Quantitative comparison of the different scaling methods using IoU measure (Mean ± Std.) on the output bounding boxes.** All the gaze-based initialisation methods use the Stimuli gaze points.

Scaling Method	Shrink	Expand	Mixed
$Fixed_{N=10}$	0.042 ± 0.021	0.074 ± 0.016	0.075 ± 0.018
$Fixed_{N=5}$	0.151 ± 0.015	0.241 ± 0.008	0.242 ± 0.007
$Fixed_{N=4}$	0.188 ± 0.015	0.303 ± 0.011	0.294 ± 0.011
$Fixed_{N=3}$	0.262 ± 0.010	0.365 ± 0.010	0.365 ± 0.012
$Fixed_{N=2}$	0.302 ± 0.013	0.286 ± 0.019	0.328 ± 0.015
RP_{union}	0.352 ± 0.015	0.330 ± 0.017	0.369 ± 0.018
RP_{square}	0.369 ± 0.007	0.388 ± 0.011	**0.433 ± 0.011**
CAGIC	0.416 ± 0.013	-	-

methods to find the region of interest. Namely, we zoomed in or out in every iteration despite the size of the initialisation region. Alternatively, in Section 3.3.2, we proposed the adaptive Mixed scaling method, which flexibly decides to use Shrink or Expand *w.r.t.*the size of the initialisation area. Our results show that the Mixed scaling method using Region Proposal initialisation provides the highest IoU value among the compared methods, exceeding even the score of the state-of-the-art method, CAGIC.

Overall, based on the quantitative evaluations, we found that using gaze-based initialisation is useful, and with the proposed additions, the new framework G-DAIC was able to outperform the state-of-the-art method. Furthermore, we found that for the initialisation, it is best to use the gaze points collected during the Stimuli stage and that the proposed adaptive Mixed scaling method is better suited for our multimodal framework than the previously proposed Shrinking only method. Among the proposed initialisation methods, we found that the Region Proposal-based RP_{square} initialisation was the most successful in every experiment.

User study

Cycle Crop-Caption consistency. To measure the success of the proposed image-cropping framework, we performed a quantitative comparison of how well each method preserved the contextual information of the given user caption. We asked five users who did not participate in our previous experiments, to describe the output image crops of G-DAIC. The caption similarity scores were calculated using the same natural language processing metrics as in [33] and introduced in Section 2.4.3 for a fair comparison. The comparative results are presented in Table 3.9. This experiment demonstrated that G-DAIC was the most efficient

Table 3.9: **Comparison of user intention presence.** We ask users to caption cropped images and compare with natural language metrics how similar they are with the original desired caption.

NLP Metric	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE_L	CIDEr
GradCAM[87]	0.2728	0.1410	0.0874	0.0531	0.1182	0.2790	0.6973
MAttNet[47]	0.1718	0.0937	0.0603	0.0355	0.1132	0.2947	0.7154
CAGIC[33]	0.3424	0.1876	0.1017	0.0631	0.1702	0.2970	0.9054
G-DAIC	**0.3519**	**0.1963**	**0.1065**	**0.0654**	**0.1747**	**0.3072**	**0.9536**

Table 3.10: **User study quantitative result.** Qualitative comparisons among the state-of-the-art image cropping methods with the original image and G-DAIC were compared through a human survey and evaluated by aggregation.

	Original Image	MAttNet [47]	GradCAM [87]	CAGIC [33]	G-DAIC
Aggregated percentage (%)	18.07	18.73	19.13	20.87	**23.20**

at preserving the contextual information provided by the user according to every metric. Meaning that the image captions generated from the output crops were more similar to the original user caption when we used G-DAIC compared to all the other baseline methods.

Aesthetic assessment. Due to the subjective nature of this research area, we performed a user study to compare the quality of the image crops provided by the baseline methods and G-DAIC. We performed an online user study, asking the participants to select the best-looking output image crop among the five images shown in a randomised order. We asked 15 users, resulting in 1500 decisions, to choose one among the original image and the output image crops of MAttNet, GradCAM, CAGIC , and G-DAIC. Table 3.10 shows the aggregated percentages. This experiment shows that using the aesthetics information in the optimisation framework improved the quality of the output image crops. Furthermore, based on the users' votes, the most popular choice was the proposed method, G-DAIC.

3.4.4 Runtime

Finally, we aimed to reduce the runtime (t) of the description and aesthetics-based image cropping method compared to the state-of-the-art optimisation method [33] by adding a new modality to the framework. Therefore, beyond the quantitative and qualitative experiments, we measured the runtime of G-DAIC and the baseline image cropping methods. The results of this experiment are shown in Table 3.11.

Table 3.11: **Runtime comparison of the baseline algorithms with the proposed method using a fixed number of iterations (G-DAIC $N_{iter} = 25$), *Early termination* (G-DAIC n_{iter}) and the $GradCAM_{caption}$-based (G-DAIC $GradCAM_{caption}$).**

Method	Runtime (sec)
A2RL [44]	0.150
VPN [68]	0.008
Anchor [86]	0.005
GradCAM [87]	0.030
MAttNet [47]	0.020
CAGIC [33]	412
G-DAIC ($GradCAM_{caption}$)	23.42
G-DAIC (N_{iter})	40.92
G-DAIC (n_{iter})	32.52

In agreement with our hypothesis, the gaze-based initialisation successfully reduced the runtime by 92.11% compared to the results presented in Section 2.4.6 and [33]. This runtime improvement is especially impressive, considering that both the quantitative and the qualitative evaluations confirmed that G-DAIC outperformed all the baseline methods. Therefore, the proposed method is faster and more accurate at localising the described part of the image and producing an aesthetically pleasing image crop.

Fixed number of iterations.

This runtime improvement was achieved in three steps using two methods. Firstly, the time to run one iteration (t_{iter}) was reduced from 2.06 to 1.637 seconds by optimising the code. This modification resulted in an average 20.53% runtime decrease per iteration. Secondly, we maximised the number of iterations (N_{iter}) by 87.5% to 25 instead of 200, assuming that the provided gaze-based initialisation provides a better starting point for our search than starting from the original image. This resulted in a 90.07% runtime decrease, where producing a single image crop took 40.92 seconds (Table 3.11, G-DAIC (N_{iter})).

Early Termination.

Finally, we used the Early Termination strategy to end the optimisation after n_{iter} instead of N_{iter} when the calculated Total loss ($\mathscr{L}_{total}(\mathbf{I}, \mathbf{y}, \theta)$) was below $T_{loss} = 5.23$. The time spent to generate a single image crop, therefore, is calculated as follows:

$$t = \begin{cases} t_{iter} \times n_{iter} & \text{if } \mathscr{L}_{total}(\mathbf{I}, \mathbf{y}, \theta) <= T_{loss} \\ t_{iter} \times N_{iter} & \text{otherwise} \end{cases} \tag{3.6}$$

This solution allowed us to save 513, equal to 20.52% of the iterations across the whole dataset. The Early Termination further reduced the average runtime from 40.92 to 32.52 seconds/image (Table 3.11, G-DAIC (n_{iter})). Overall the average total time to produce a single output crop takes 32.52 seconds which is 92.11% faster than the original method.

The runtime achieved by G-DAIC is significantly faster than the other iterative optimisation-based description and aesthetics-based image cropping algorithm CAGIC; however, due to the iterative nature of the proposed algorithm, it is still significantly slower but better in terms of performance than the other baselines. Finally, we evaluated the runtime of G-DAIC when initialised by $GradCAM_{caption}$. While this method achieved 34% faster runtime compared to G-DAIC (n_{iter}), the quantitative and qualitative results presented in Section 3.4.2 showed that its performance was significantly lower compared to the human gaze-based initialisation. Note that the models in our framework were not fine-tuned nor specifically trained for this task. The overall runtime could be further improved by using new modalities and potentially replacing existing or new pre-trained models in the framework.

3.5 Conclusion

This chapter proposes a new gaze-initialized multimodal optimization method for the description and aesthetics-based image cropping problem. The main motivation was to reduce the runtime of the state-of-the-art image cropping method [33] while preserving the accuracy of the previously introduced solution. Hence, we designed a new framework, which initialized the image cropping algorithm based on the subject's eye movements recorded during the image description generation. In this work, we proposed two solutions, *Fixed Grid* and *Region Proposal*, on integrating and utilizing the gaze information into the proposed multimodal framework. Furthermore, we implemented the adaptive *Mixed method* for localization of the described image region based on the size of the gaze-based initialization area. Finally, we proposed the *Early termination* of the optimization to significantly reduce the runtime. In this chapter, we have shown that with the gaze-based initialization of the description and aesthetic-based image cropping method, we were able to outperform the state-of-the-art methods significantly and in addition, we reduced the runtime by more than 92%.

Following the success of the multimodal approach, a potential extension of this work would be to include another important social clue, hand gestures, into the framework. Furthermore, we believe that utilizing the temporal information and better understanding the eye movements *w.r.t.* the caption could result in a more reliable gaze-based initialization using the gaze points recorded during the caption generation.

Chapter 4

Where Are *They* Looking in the 3D Space?

Figure 4.1: **Attention target estimation example use case visualisation.** The driver performs gaze following of the pedestrians to infer their intention and prevent a potential collision.

4.1 Introduction

We live in a multi-dimensional world and experience our environment through our senses. Vision is one of the most dominant ways we experience the world. Our sense of orientation within a given context, location, and place is influenced by an animated 3D model of the world our brains construct from the cues around us. Image-based gaze target estimation aims to infer what the subjects in the image scene are looking at from a single RGB image. Human gaze following and gaze target estimation in the wild are fundamental for visual navigation. Furthermore, this information is important to evaluate intentions and predict human behaviours in various social contexts [122]. For these reasons, gaze analysis has widely been used in neurophysiology studies [123, 124], relevant saliency prediction [125, 126] and social awareness tracking [127–129]. Humans are naturally good at understanding the actions of others and estimating where they are looking by leveraging prior knowledge. They can infer the pose and the person's orientation and reconstruct the 3D image scene based on a single view. By looking at the image, they can understand what the person is doing and predict their intentions. They can even guess what it would look like from another viewpoint. We can do this because all the previously seen objects and scenes have enabled us to build prior knowledge and create mental models of object appearance.

Earlier in Figure 1.1, we showed one of the many possible scenarios when people are trying to estimate other people's attention targets from a third-person view. In this figure, we showed a couple observing three actors on the television. Figure 4.1 shows another example of an everyday scenario where the human gaze following is critical to safe driving. In this instance, the vehicle's driver is approaching a junction where two pedestrians are about to cross the road. A crucial part of safe driving is adequate situational awareness of the driver. This includes predicting the actions of other road users, such as pedestrians and cyclists within their proximity and on their path. Gestures such as pedestrians looking around for traffic could indicate their intention to cross the road and potentially the driver's path. Furthermore, this enables the driver to determine whether the pedestrian has spotted them, which would help to prevent a potential collision. To do so, the driver observes the pedestrians from a third-person view and estimates their individual and joint gaze direction and target.

From the neurocognitive perspective, gaze perception is performed by humans to discriminate the gaze direction of others [130] as part of various social interactions, such as gaze following. This action is a vital part of the cognitive functions that allow people to learn via observation [131]. Once they successfully perceive the gaze direction of others, they can utilise this knowledge through their social cognition system in various ways [132], for example, to engage in joint attention with the observed person. By definition, joint attention happens when a gaze leader looks at a particular object which induces gaze followers to orient their attention to the same target. In the field of computer vision, the task of joint attention target estimation is often referred to as shared attention in the literature [30, 133, 134]. While these terms are similar, they are subtly different from each other [122, 132]. In this thesis, we define joint and shared attention terms according to the neurocognitive perspective as in [122] and treat gaze perception and joint attention as parts of shared attention. Shared attention requires both the initiator and the responder to be aware that they are observing the same object, unlike joint attention.

Recent work showed the ability to estimate the individual's gaze target directly from images using neural networks. We differentiate between single and joint attention target estimation methods based on the number of subjects involved in this process. A key step towards single attention target estimation was the work by Recasens *et al.* [135], which demonstrated the ability to detect the attention target of each person within a single image. This image-based method did not consider human attention over time and the cases when the target of the subject's attention was outside the image frame. This approach was later extended to handle the issue of out-of-frame gaze targets [136]. Afterwards, Chong *et al.* [30] proposed a spatio-temporal approach to gaze target prediction, which models gaze dynamics from video data. These single-target estimation approaches are attractive because they can leverage head pose features and the saliency of potential gaze targets to resolve ambiguities in gaze estimation. However, unlike humans, they only use 2D information to estimate the point of interest. For the first time, Fang *et al.* [32] proposed a method using depth prior, 3D gaze estimation and 2D field-of-view (FOV) estimation for gaze target prediction. This was essential to more realistic gaze target estimation in 3D space. However, in reality, the FOV of people is not two-dimensional, as a healthy person can observe things in front of them within a 3D cone.

In social scenarios, we often infer the gaze target of two or more people simultaneously. To solve this task, an inefficient way is to use the single target estimation models and estimate the gaze target of every individual in the scene one by one and then combine these estimates to find the joint attention target of the scene. Fan *et al.* [133] proposed to infer the joint attention target in third-person social scene videos using a spatial-temporal neural network to overcome the limitations of the single target estimation methods. This solution was based on a head detector module, region proposal, and saliency estimation. Later, Sumer *et al.* [137] proposed an end-to-end solution without using any temporal information, face detector or head pose estimator to detect and localise joint attention. Both existing joint attention target estimation methods rely solemnly on 2D data, making the models more prone to errors, such as physically impossible predictions where the subjects are estimated to look within their blindspot.

This study extends the previous approaches by developing a model for 3D FOV-based co-attention target estimation by jointly using 2D and 3D clues and temporal information. Our motivation is to create a model that can translate the image as humans do and estimate where the subject is looking in the 3D space. Introducing strong 3D clues into this framework helps the model to handle occlusion and other challenging cases better later introduced in Section 4.3. Our contributions are fourfold:

- We propose a novel joint attention target estimation model which mimics how humans observe their environment using 2D and 3D clues.
- We trained a spatial model that can utilise the full scope of the 3D information of the 3D space provided by the monocular depth estimator. The predicted 3D FOV of the subjects are used as a probability map instead of a fixed angle, hard thresholded FOV cone to make the model more robust against the potential 3D gaze direction estimation errors.
- This is the first work to incorporate depth information into a joint attention target model and investigate its usefulness in the case of both joint and single attention target estimation tasks.
- The proposed joint attention target estimation approach outperformed the state-of-the-art single and joint attention target estimation methods. The results of the methods were compared on a large-scale image benchmark dataset and two video datasets.

In this chapter, we rely on an implicit social clue to infer multiple users' common gaze target point in the scene. We present the extensive preliminary experiments and the analysis of the weaknesses of the existing attention target estimation methods, which formed the proposed solution's basis. We aim to address the physically impossible predictions of the models, where the subjects are predicted to observe a point within their blind spot. Inspired by the working of the human visual system, we proposed to incorporate depth information into the attention target estimation pipeline. By fully utilising the depth prior generated by a monocular depth estimator module [138] combined with the subject 3D orientation, we predicted a 2D probability map indicating the pixel-wise co-attention target likelihood on the image frame. To the best of our knowledge, this is the first work to fully utilise depth information in a joint attention target estimation framework.

4.2 Related Works

Attention target estimation methods can be categorised based on the number of people involved in the social interaction in third-person social images or videos.

Single attention target (SAT). This class of methods focuses on a single subject within the scene and aims to infer their visual attention target location based on the visual information.

The pioneering work of this research field was proposed by Recasens *et al.* [139]. The proposed deep model was the first to learn to find the gaze target in the image through two pathways. The input of the scene saliency pathway is an RGB image designed to estimate the saliency of the scene. The subject's gaze direction was estimated through the gaze pathway, which takes the face crop of the subject and its spatial location within the original image as the input. The image dataset proposed in this paper serves as the primary large-scale benchmark of this research field. Despite the promising results presented in this work, the proposed method did not handle out-of-frame targets or modelled the temporal dynamics of attention. Chong *et al.* [30] proposed a spatio-temporal model to address these limitations. The authors proposed a video attention target dataset and extended the image dataset with out-of-frame annotations. These methods were designed to estimate the attention target location of a single subject within the 2D image. Other related works include [140–142].

Joint attention target (JAT). Fan *et al.* [133] proposed a method to infer the joint attention target of two or more people in the scene. The method takes an image frame as input and, through a head detector, a gaze estimation module, and a region proposal module, generates a joint attention spatial heatmap. Furthermore, the authors presented this task's first large-scale third-person social scene video dataset. An end-to-end Joint attention target (JAT) estimation method was developed by Sumer *et al.* [137]. A frequent common mistake of the presented SAT and JAT methods is that they do not utilise 3D clues for scene understanding. Therefore, the target estimates are often located within the subject's blind spot.

Depth-aware attention target This limitation was addressed by the latest SAT estimation works proposed by Fang *et al.* [32] and Bao *et al.* [143]. [32] proposed an image-based SAT estimation model. This work does not consider temporal information and despite the authors developed a 3D gaze direction estimation module, the FOV generator only relies on 2D gaze and head direction and a view angle of 60°. [143] proposed to reconstruct the 3D scene to 3D point cloud using relative depth estimation and 3D human pose estimation. They selected the front-most 3D points along the predefined visual rays to find the final gaze target position. These works are the first and currently only existing works which took a step towards integrating the depth clue into the attention target detection model.

The work presented in this Chapter is the first one to use depth information for the JAT estimation task. We propose formulating the 3D FOV as a probability map instead of hard thresholding the values along a pre-defined angle or visual rays. This new way of representing the FOV allows more room for inaccuracies coming from the 3D gaze direction estimator. Furthermore, our method combines the depth information with the subjects' pose to produce a joint 3D FOV probability map to predict joint gaze targets of multiple subjects within the scene.

4.3 Research questions and preliminary experiments

To better understand the potential problems which may occur during joint attention target estimation, we conducted a series of experiments on the single attention target estimation task. Considering that joint attention is the combination of two or more subjects' simultaneous attention, we believe that analysing the gaze targets individually is beneficial for our task. Therefore, this section introduces our findings on the single attention target estimation field using static images and videos. We used [30] as a baseline and took incremental steps to investigate the weaknesses of the state-of-the-art gaze target prediction models. More specifically, during our preliminary experiments, we investigated the following research questions:

Q1 **Different temporal modes:** Is there a better temporal mode we could use instead of the Convolutional Long Short-Term Memory network (Conv-LSTM)?

Q2 **Different attributes:** Could we improve the performance by relying on new attributes, such as body position?

Q3 **Non-local network:** Would Non-local networks improve scene understanding by predicting a better attention map?

Q4 **Non-local block variants:** Which type of Non-local block is the most suitable for our task?

Q5 **Alternative head position encoding:** Could we encode the head position more effectively?

Q6 **Inside the head bounding box:** Should we allow the model to predict the subject to look inside their head bounding box?

4.3.1 Dataset and evaluation metrics

Benchmark datasets - Single Attention Target Estimation

GazeFollow dataset. [135] A widely used dataset for predicting the gaze target of the subjects is the GazeFollow benchmark dataset [135], which contains static images. See example images in Figure 4.2. Amazon Mechanical Turkers annotated the head and gaze locations inside the images of 130,339 people in 122,143 images. 10 different people annotate each image of the test set. The diversity of these annotations well reflects the subjective and

Figure 4.2: **Single attention target estimation benchmark dataset highlights.** On the left we show images of the GazeFollow image dataset [139] and on the right sample image frames of video sequences of the VideoAttentionTarget dataset [30].

complex nature of the gaze target attention estimation task. This dataset does not handle cases when the gaze target is outside the image frame.

VideoAttentionTarget (VAT) dataset. [30] The VideoAttentionTarget video dataset [30] is specifically designed for modelling the gaze target in videos. We show an example of randomly sampled image frames of different video sequences in Figure 4.2 to demonstrate the diversity of the dataset. For each video clip, the annotators provided the head bounding boxes as well as the gaze target of each person with the indication of whether the person was looking outside the video frame.

Evaluation Metrics

In our experiments, we evaluate the performance of the single attention target estimation models on the GazeFollow and VAT benchmark datasets using the following three performance measures: AUC, Distance, and Out-of-Frame AP.

- **AUC:** Each cell in the spatially-discretised image is classified as a gaze target or not. The ground truth comes from thresholding a Gaussian confidence mask centred at the human annotator's target location. The final heatmap provides the prediction confidence score evaluated at different thresholds in the receiver operating characteristic curve (ROC). The area under the curve (AUC) of this ROC curve is reported.

- **Distance:** Pixel-wise normalised L2 distance between the ground truth target location and the pixel of the maximum value in the predicted heatmap.

- **Out-of-Frame AP:** The gaze target estimation model learns a scalar α which quantifies whether the person's focus of attention is located inside or outside the frame, with higher values indicating in-frame attention. The average precision (AP) is computed for the prediction score from the scalar α against the ground truth computed in every frame.

Note that AUC and Distance are computed whenever an in-frame ground truth gaze target (the heatmap always has a maximum). Also, the ten ground truth annotation locations of the GazeFollow dataset were averaged, and the *average L2 distance* was calculated *w.r.t.*this new ground truth location. We found that the average position was completely off from the actual gaze target in cases where the ground truth annotations disagreed. The *minimum L2 distance* was calculated as the minimum distance from all the ground truth gaze locations. We also show the performance of the annotators (Human performance) across all three measures of the datasets. This is done by comparing annotator predictions in all pairs and averaging them.

4.3.2 Experimental analysis of the research questions

Q1. Different temporal modes

In the proposed framework, Chong *et al.* used a Conv-LSTM [144] to integrate temporal information from a sequence of frames. In this work, they did not compare the performance of the proposed model using other temporal networks; therefore, in our study, we used Peephole LSTM [145], Gated Recurrent Unit (GRU) neural network [146] and Temporal Convolutional Network (TCN) [147]. We trained the models on the VAT dataset using initialisation weights from the spatial model. In Table 4.1, we show the performance of the trained models. In this comparison, we found that by using TCN, we achieved better results than the original method in terms of AUC and L2 distance. The AP measure was slightly below the performance of [30], which we further investigated.

Table 4.1: **Quantitative comparison of the different temporal modes using AUC, L2 distance and AP measure on the VideoAttentionTarget dataset.**

Temporal mode	AUC ↑	L2 distance ↓	AP ↑
Conv-LSTM [144]	0.860	0.134	**0.853**
PeepholeLSTM [145]	0.837	0.149	0.853
Conv2DGRU [146]	0.855	0.132	0.847
TCN [147]	**0.880**	**0.132**	0.847
VideoAttention [30]	0.860	0.134	0.853
Human performance	0.921	0.051	0.925

In our qualitative evaluation (See Figure 4.7), we found that the new model using TCN has three major weaknesses: bias towards humans, occlusion, and ambiguous ground truth annotations. Specifically, the model tends to predict physically impossible gaze targets *w.r.t.*the subject. Namely, the predicted gaze target points were within the blind spots of the subjects. Finally, we found that the model has a bias towards humans.

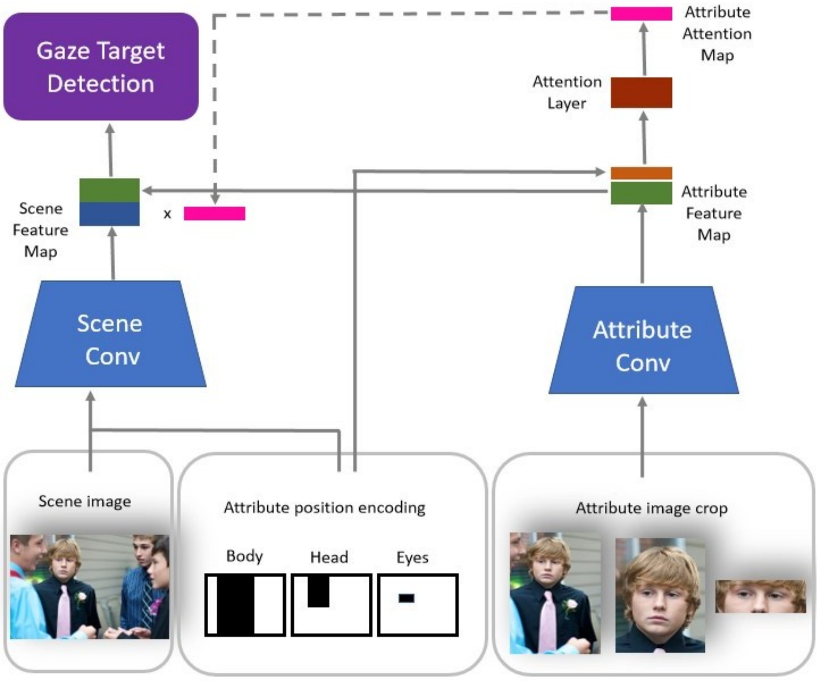


Figure 4.3: **Visualisation of the SAT framework using different attributes.**

Q2. Different attributes

The existing gaze target estimation methods only rely on the head position of the subject; however, when humans try to estimate others' gaze targets, they rely on various clues, including the subject's eye and body position and other contextual clues. Inspired by our findings in Chapter 3, we tried to extend the single attention target estimation method by introducing new attributes. Therefore, we proposed to use different attributes, which are body, head and eyes and their combinations to provide more information on the subject during training under the assumption that these new attributes will help the model estimate the subject's gaze target. To do this, we extended the architecture of [30] by adding new attribute branches to the scene branch (See Figure 4.3). Note that we used separate Attribute Conv modules for each module, then we multiplied the Scene Feature map with every computed Attribute attention map and concatenated every Attribute Feature map to this product which was the input of the Gaze Target Detection Module. The quantitative results are shown in

Table 4.2: **Quantitative comparison of the spatial models trained using the eye location, head position and body position attributes of the subject and their combinations on the GazeFollow dataset.**

Attributes	AUC ↑	Avg distance ↓	Min distance ↓
Eye only	0.856	0.144	0.241
Body only	0.814	0.185	0.257
Head + eye	0.877	0.122	0.194
Head + body	0.834	0.202	0.207
Head + eye + body	0.851	0.130	0.238
Head only	**0.896**	**0.106**	**0.170**

Table 4.2. Our quantitative results showed that introducing new attributes does not improve the model's performance compared to the model trained using only the head position.

Q3. Non-local network

One of the biggest challenges of Gaze Following is to estimate a correct attention map *w.r.t.*the subject's location within the scene. Saliency estimation methods can better understand the whole image scene but do not reflect the spatial dependence *w.r.t.*the subject's position. Recently, Non-local neural networks were proposed by Wang *et al.* [148] for capturing long-range dependencies. Essentially, the proposed non-local operation computes the response at a position as a weighted sum of the features at all positions. Inspired by the similarity between our tasks and the use case of this network, our hypothesis was that it would help us better model the spatial relationship between the subject and its gaze target point; hence we integrated this block into our current framework.

We defined two bottlenecks for our experiments. Following the architecture of [30], we used a ResNet-50 as a backbone of one of our architectures. In this case, we defined five blocks (Stage 1-5 in Figure 4.4.), each with [3,4,6,3,2] layers, respectively. The other backbone has five blocks with one layer each. Since this architecture is much smaller, we later refer to this backbone as *light*. We inserted the Non-local blocks between every layer (*between layers*) or between every block (*between blocks*) into our backbones (See in Figure 4.4). We removed the Head conditioning branch of the original architecture proposed in [30] and only worked with the main scene branch.

In [148], the authors proposed four different non-local blocks: Gaussian, embedded gaussian, dot-product and concatenation. However, due to computational power limitations, we could only use the *Gaussian* and the *Embedded Gaussian* Non-local blocks. We show the quantitative comparison between different backbones, block placements, and block types

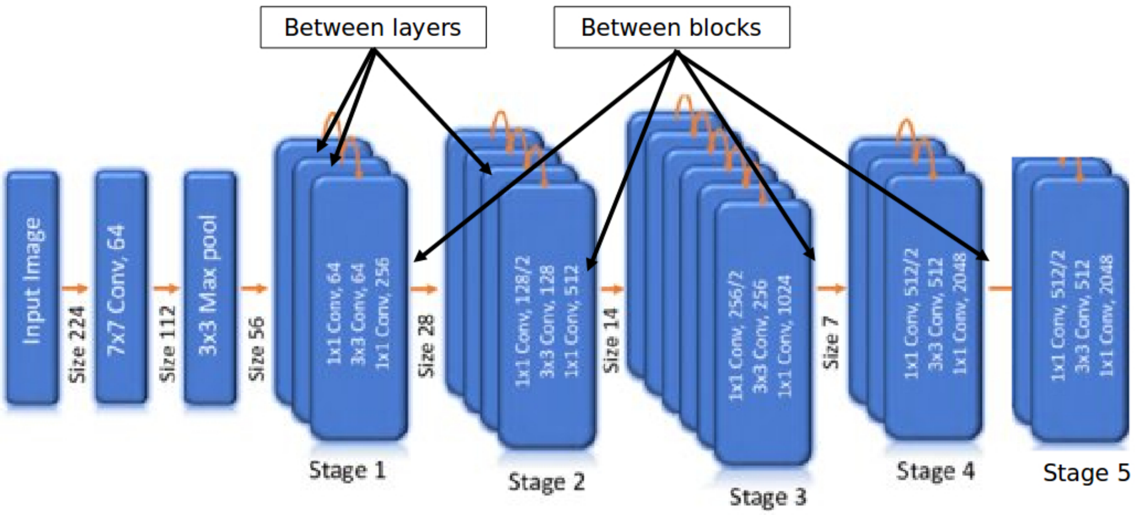


Figure 4.4: **Visualisation of different NL block placements between the layers and the blocks of the backbone.**

Table 4.3: **Quantitative comparison of different spatial models on the GazeFollow dataset.**

Backbone	Spatial model	AUC ↑	Avg dist ↓	Min Dist ↓
Light	Between layers, Gaussian	0.783	0.308	0.229
	Between blocks, Gaussian	0.855	0.245	0.170
	Between layers, Emb. gaussian	0.758	0.324	0.245
	Between blocks, Emb. gaussian	0.752	0.321	0.244
ResNet-50	**Between layers, Gaussian**	0.907	0.173	0.105
	Between blocks, Gaussian	0.851	0.259	0.182
	Between layers, Emb. gaussian	0.866	0.233	0.158
	Between blocks, Emb. gaussian	0.636	0.306	0.240
VideoAttention [30]		0.921	0.137	0.077
Human performance		0.924	0.096	0.040

in Table 4.3. and 4.4. Our experimental results show that the Gaussian non-local blocks performed better for our task for both backbones. Furthermore, placing the non-local blocks between the layers improved the performance more in most cases than placing them between them. Note that when we place the blocks between the layers in the ResNet-50 backbone, we use 18 blocks instead of 4, and in the case of *light*, we use five instead of 4 blocks.

Inspired by the design choice of [148], we ran an experiment where we used the ResNet-50 architecture and inserted 4 Non-local blocks to stage 2 and 6 blocks at stage 3 between each layer (Model92: ResNet-50, 4-6 between layers, Gaussian) . We also attempted to

Table 4.4: **Quantitative comparison of different spatiotemporal models on the VAT dataset.**

Backbone	Spatiotemporal model	AUC ↑	L2 distance ↓	AP ↑
Light	Between layers, Gaussian	0.770	0.277	0.752
	Between blocks, Gaussian	0.752	0.281	0.747
ResNet-50	**Between layers, Gaussian**	0.854	0.171	0.843
	Between blocks, Gaussian	0.842	0.205	0.809
Chong *et al.* [30]		0.860	0.134	0.853
Human performance		0.921	0.051	0.925

connect the Head conditioning branch to this model (Combined model: ResNet-50, 4-6 between layers, Gaussian with head branch). The quantitative evaluation of the proposed spatial models is shown in Table 4.5. We found that the Model92 architecture had a higher AUC score meaning that the estimated gaze heatmap had a better overlap with the ground truth than the state-of-the-art method [30]. However, the average and minimum distances are slightly higher.

We also evaluated this model's performance on the VAT dataset in Table 4.6. We evaluated the spatial Model92 without fine-tuning; however, despite the AUC measure being higher than the state-of-the-art's, the AP score was very low. This is because the Gazefollow dataset does not have in/out labels; therefore, our model could not learn when the subject looked outside the image frame during spatial training. To improve the AP score, we finetuned Model92 without temporal training using only the binary cross entropy loss because the AUC score was high. Furthermore, we also trained a spatiotemporal model using TCN. The results of our experiments show that there is still room for improvement to achieve the performance of [30]'s full model or the human annotators. Note that Model92 does not use the head branch; hence we also included the performance of [30] in Table 4.6 without the head branch as well as without the temporal training. Compared to these models, our proposed model performed better without any finetuning.

During our qualitative evaluation found the following similarities among the cases with the highest L2 distance: the heatmap is sparse, the max value is not in the correct image region, the estimated gaze target is in a physically impossible location (behind the person, on the subject's face *e.t.c.*), the disagreement between the ground truth annotations is very large. For the cases with the lowest error, we found that the heatmap is concentrated around the target object, the scenario is typically human-human or human-object interaction, and the ground truth annotations are very close.

Table 4.5: **Quantitative comparison of different spatial models on the GazeFollow dataset.**

Spatial model	AUC ↑	Avg distance ↓	Min distance ↓
Model92	**0.921**	0.147	0.083
Combined	0.865	0.201	0.135
Chong *et al.* [30]	0.921	**0.137**	**0.077**
Human performance	0.924	0.096	0.040

Table 4.6: **Quantitative comparison of different models on the VAT dataset.** Note that there is no Head branch integrated in the Model92 architectures.

Model	AUC ↑	L2 distance ↓	AP ↑
Model92 no finetune	0.889	0.148	0.638
Model92 fine-tune, no temporal	0.889	0.148	0.821
Model92 spatiotemporal	0.829	0.170	0.842
Chong *et al.* [30] no head	0.758	0.258	0.714
Chong *et al.* [30] no temporal	0.854	0.147	0.848
Chong *et al.* [30]	0.860	0.134	0.853
Human performance	0.921	0.051	0.925

Q4. Non-local block variants

Non-local blocks proved to be useful in our experiments. Quantitative results on the benchmark dataset show that the Compact Generalised Non-local network proposed by Kaiyu *et al.* [149] and the Efficient Attention by Shen *et al.* [150] proposed recently performed better than [148]. We trained Model92 using these non-local block variants. We show the new model's performance in Table 4.7. on the GazeFollow dataset. All the variants performed similarly; however, for our task, the original Non-local blocks [148] seemed the best choice based on our results.

Q5. Alternative head position encoding

In Figure 4.5 we show an example input of the existing SAT methods. Namely, the original image which is the full RGB image of the scene, the head crop of the subject based on the ground truth head bounding box information. Studies found that providing the head position as a spatial reference along with the scene helped the model learn faster, therefore, the methods use the head position encoding in addition.

Table 4.7: **Quantitative comparison of different NL block variants on the GazeFollow dataset.**

Non-local block	AUC ↑	Avg distance ↓	Min distance ↓
CG-NLx [149]	0.919	0.151	0.086
EA-NL, 1 head [150]	0.921	0.151	0.086
EA-NL, 2 head [150]	0.920	0.154	0.089
Original [148] (Model92)	**0.921**	**0.147**	**0.083**

Table 4.8: **Quantitative comparison of different head position encoding on the GazeFollow dataset.**

Positional encoding	AUC ↑	Avg distance ↓	Min distance ↓
1D addition	0.672	0.236	0.309
1D multiplication	0.780	0.228	0.306
2D	0.887	**0.128**	0.197
Model92	**0.921**	0.147	**0.083**

The head position *w.r.t.*the image scene is derived from the given head bounding box information. In our example, we displayed the head bounding box over the original image in green. Existing method encode this location information as an image, where the bounding box is a black rectangle on a white image. See example visualisation of the head channel within the original image frame in the top row of Figure 4.5. The resolution of the encoding is fixed, in our experiments its 224×224.

When the subject is far from the camera, the currently used head position image would only have a small bounding box (black rectangle) due to the difference between the original image and the resolution of the head position encoding. The smaller the rectangle on the original image, the more likely the information gets lost during transformations. Therefore, we proposed a new head position encoding, shown in Figure 4.5. The proposed solution uses the bounding box coordinates to represent the position of the subject's head within the original image frame. Our solution used the coordinates of the edges of the rectangles and formed sequences of evenly spaced numbers between the edge coordinate values with the desired encoding resolution. See example visualisation at the bottom of Figure 4.5.

In Table 4.8. we show quantitative results using the Model92 architecture with different head position encodings. First, we multiplied the generated x coordinates of the head bounding box with the y coordinates (1D multiplication), next we multiplied the x coordinates with the size of the bounding box and added the y coordinates (1D addition); lastly, we stored

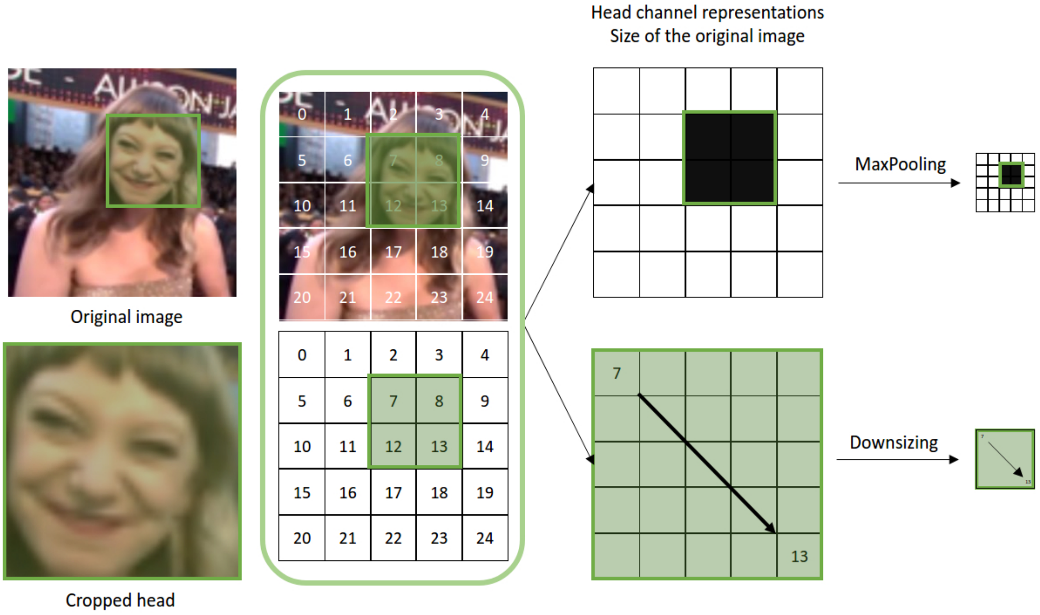


Figure 4.5: **Representation of the proposed new positional head position encoding.**

both the x and y coordinates (2D). Our results show that the new 2D head position encoding is better than the 1D representations we tried, but none gave us better results than the original spatial encoding.

Q6. The attention target is within the subject's head bounding box

A person cannot look at their face; therefore, the attention target is unlikely to be within the subject's head bounding box. To test our hypothesis, we studied the ground truth head and co-attention bounding box annotations of the VideoCoAttention (VCA)[133] joint attention target estimation dataset. See Section 4.5.1 for details on the VCA dataset. We performed an experiment to show how the prediction accuracy changes when we exclude the head bounding box area of the subjects within the scene.

First, we calculated the frequency of the ground truth co-attention bounding box annotation's intersection with the subjects' head bounding boxes (See Table 4.9a). We found that the frequency of the intersection occurrence is different among the dataset's training, validation and test set. The highest frequency of intersections was counted in the training set, where 6.42% of the ground truth co-attention annotations intersected with at least one of the head bounding boxes of the frame. While this number is relatively high, we also measured the average AUC within the intersection, and we found that the AUC score associated with these points was very low. Meaning, that while there exists an intersection between the head

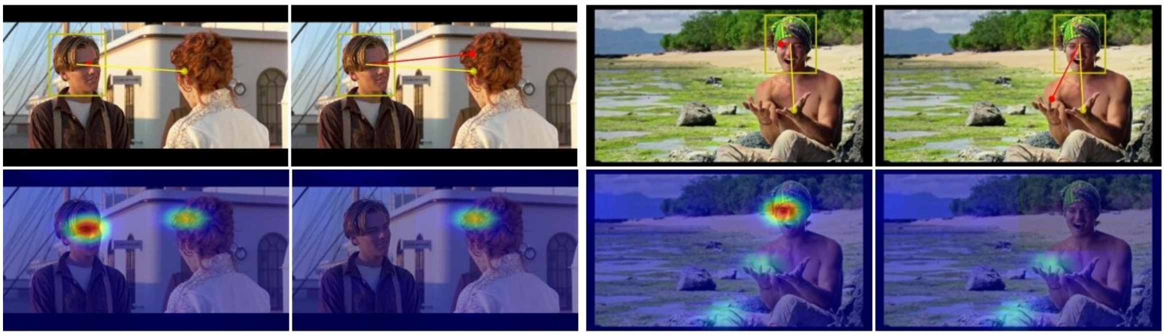


Figure 4.6: **Qualitative highlights of the head bounding box exclusion experiment.** We show two examples of the recurring error where the model predicts the user to look within their head bounding box at their own face. By excluding the area of their head bounding from the search field the gaze target estimation accuracy improved. The observed subject's head bounding box is highlighted in yellow, the ground truth annotation is marked as a yellow circle or blue dot, and the estimated gaze target estimate is shown as a red circle.

Table 4.9: **Quantitative comparison on the VideoCoAtt dataset [133] of the SAT and JAT accuracy with and without the head bounding box region.**

(a) Frequency and average AUC score of the head and co-attention bounding box intersections in the train, validation and test sets of VCA.

	Train	Validation	Test
Frequency (%)	6.42	3.67	2.55
Average AUC	0.004	0.002	0.002

(b) Results of the quantitative evaluation on the VCA dataset including and excluding the head bounding box image region for the single and joint attention target prediction.

	Single Attention		Joint Attention	
	L2 dist (px) ↓	Inout (%) ↑	L2 dist (px) ↓	Inout (%) ↑
Include	67.38	54.85	56.48	53
Exclude	**64.58**	**56.48**	**48.70**	**66.53**

and co-attention bounding boxes, the points within this area are not the co-attention target points.

Therefore, we performed an experiment where we excluded the head bounding box area of the image and measured the single and joint attention target estimation accuracy in terms of L2 distance and Out-of-Frame AP (denoted as Inout). In Table 4.9b, we show that after excluding the points within the head bounding box regions from the attention target estimation, the results improved in the case of both single and joint attention target estimation. Qualitative highlights are shown in Figure 4.6.

4.3.3 Summary

We found the following answers to our questions through the preliminary experiments presented in this section:

1. We compared the single attention target estimation model's performance with different temporal modes, including Conv-LSTM, PeepholeLSTM, GRU and TCN. Our results showed that using TCN as the temporal part of the model achieved better performance *w.r.t.*AUC and the L2 distance measure, meaning that the predicted probability heatmap was more similar to the ground truth annotations and the target estimate's distance from the ground truth annotation was smaller. Regarding AP, it is better to use Conv-LSTM to predict whether the subject is looking inside or outside the image frame.

2. We conducted experiments using the following attributes and some of their combinations: eye position, head position and body position. The positions within the image were given as bounding boxes. We extended the framework of [30] and replaced the head position with other attributes or included new ones. Inspired by our findings in Chapter 3, we assumed that introducing additional attributes would improve the model's performance. We found that using the eye position achieved better performance than the body. This is because the eye position is more closely related to the person's gaze direction than the body position. However, overall we found that using head position only resulted in a better estimate using all three measures. We think this is because the subjects' eye location in the benchmark datasets is often not visible; therefore, it might be misleading to the model.

3. To capture the long-range dependencies and therefore calculate a better attention map *w.r.t.*the subject's position, we used Non-local layers. We found that inserting some Non-local layers between specific layers of the ResNet-50 of the spatial model achieved a better AUC score than the state-of-the-art method on the GazeFollow benchmark dataset. Furthermore, we found that in comparison with [30] ablation study results, the proposed model outperformed different versions of the previous model respectively by a large margin. However, after the temporal training, we could not surpass the performance of the existing method. The weaknesses and findings regarding this model are discussed later at the end of this subsection.

4. Motivated by the success of the spatial model using Non-local blocks between the layers of the backbone, we further evaluated the performance of our framework using three different Non-local block variants. Our experiments showed that the Non-local blocks proposed by [148] are the best suited to our task.

5. We proposed three head positional encoding alternatives to the one commonly used in this research field. Our experiments showed that the original positional encoding method is better regarding AUC and Min distance measures; therefore, we followed the standard head positional encoding in our following experiments.

6. Lastly, we proposed to exclude the image region of the subject's head bounding box annotation from the potential attention target estimate locations. This idea is motivated by the fact that a person can not look at their face. The results of this experiment showed that both the individual and the joint attention target estimates accuracy improved when we excluded the head bounding box area.

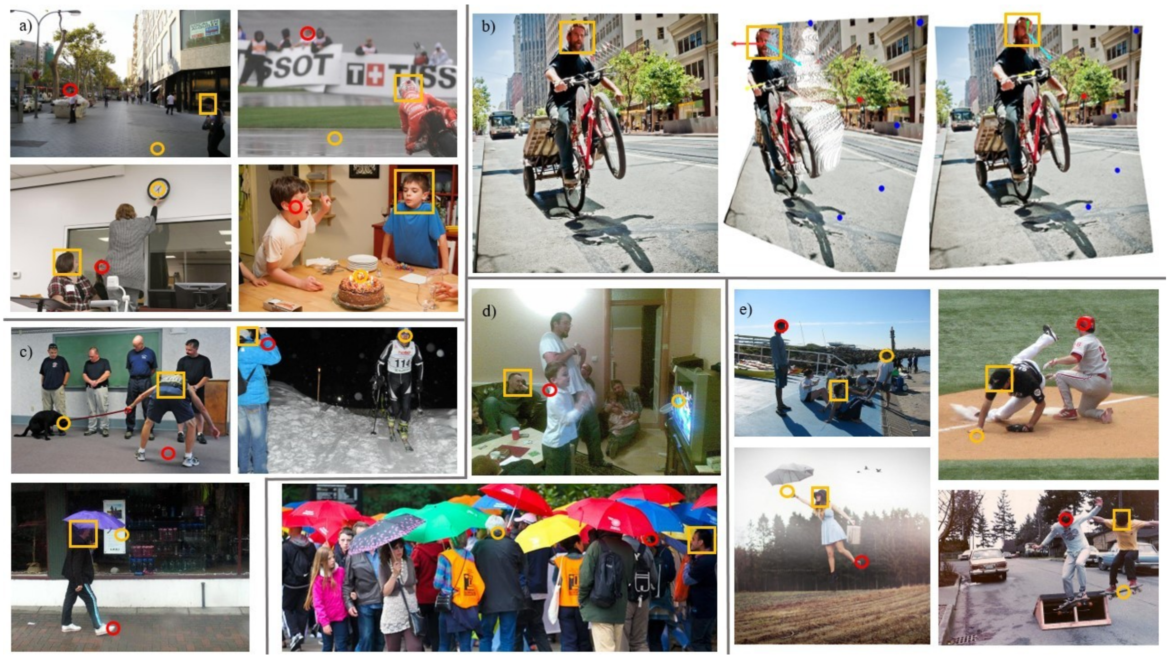


Figure 4.7: **Visualisation of different failure cases**, including human bias (a), ground truth annotation ambiguity (b), occlusion of the subject (c) and the gaze target (d), and physically impossible estimates (e). The observed subject's head bounding box is highlighted in yellow, the ground truth annotation is marked as a yellow circle or blue dot, and the estimated gaze target estimate is shown as a red circle.

Finally, after careful analysis of our quantitative and qualitative results, we found the following problems with the state-of-the-art and the proposed methods:

- **Bias towards humans:** In Figure 4.7 a), we show a common mistake of our model. There are many cases in the benchmark datasets where the gaze target of the subject is not another person but an object or the ground in front of them or somewhere between two potential gaze target locations. These cases are especially common in the video dataset, where we observe the user's eye movements frame by frame. During the observed time, the user often shifts their gaze between people or objects. Our analysis showed that our model is biased towards humans, and it is more likely to predict the subject to be looking at another person within the image instead of an insignificant location.

- **Ambiguous annotations:** We studied the annotation of the GazeFollow image dataset, where we have ten annotations for each test image to better understand the ambiguity of the gaze target annotations (example shown in b)). This variation among human annotations originates from the subjective nature of the task. In part b) of the figure, we visualised the ground truth annotations of the selected subject's estimated gaze target location in blue and their average in red in 3D using a prior depth map. We show that annotating the gaze target locations on a 2D image can result in inaccuracies. For example, in b) several ground truth annotations and their average fall behind the subject.

- **Occlusion:** Based on the qualitative results, we identified the two most common causes of failed gaze target estimates, which are:

  - **Occlusion of the subject:** In Subfigure c), we show examples where the occlusion of the subject's face caused the prediction error. Generally, when the model is used to estimate someone's gaze target from the back, the estimate is often different from the ground truth annotations. Note that for these cases, the manually annotated ground truth target locations are not well aligned, meaning that even the human annotators could not agree on the gaze target, which highlights the complexity of this case.
  - **Occlusion of the gaze target:** In other cases, we found that the gaze target selected by the human annotators was occluded by, for example, another person within the image scene (See Subfigure d)). This scenario is not uncommon in the existing benchmark datasets. While it is not correct to annotate that the subject at the top of Subfigure d) is looking at the television even when the other person is standing in front of him, it makes sense regarding the ongoing activity.

- **Physically impossible estimates:** Finally, we found a very common problem where the model predicted the gaze target to be behind the subject or in a physically impossible position. In reality, humans can only look at target points within their field of view, which is defined as the part of their visual field that can be viewed instantaneously [151]. This error may occur due to an incorrect head pose or gaze direction estimate or when the most probable target is located behind the subject.

4.4 Methodology

In this Section, we introduce a novel framework using a monocular depth estimator and a 3D FOV-based probability map to estimate the joint attention target while minimising the physically impossible gaze target estimate. Our assumption is that by integrating relative depth information of the scene and the calculated joint 3D FOV of the subjects in our framework, the model will learn to differentiate between the FOV of the subject and the blind spot. The framework of the proposed method is shown in Figure 4.8.

The input of this method, following the results of our preliminary experiments in Section 4.3, is the original image, the head positional encoding. We use an existing monocular depth estimation method [138] for relative depth map generation from the single image input and a gaze direction estimation method [152] to estimate the subjects' 3D gaze direction.

4.4.1 Framework

Our framework comprises three major components: a *Relative Depth Prior Module*, a *3D Field-of-View Module*, and finally, a *Joint Attention Target Prediction Module*.

Relative Depth Prior Module

This module is the core of the proposed method, as the generated depth map is the input of both the 3D Field-of-View and the Joint Attention Target Prediction Modules. For our task, we are primarily interested in the order of the objects and where they are located *w.r.t.*each other. Therefore, instead of estimating the absolute depth, we used an existing monocular depth estimation network [138] to estimate the relative depth map of the scene. Relative depth is the ratio between the depth of two points, which is useful to determine which point is closer to the camera [153].

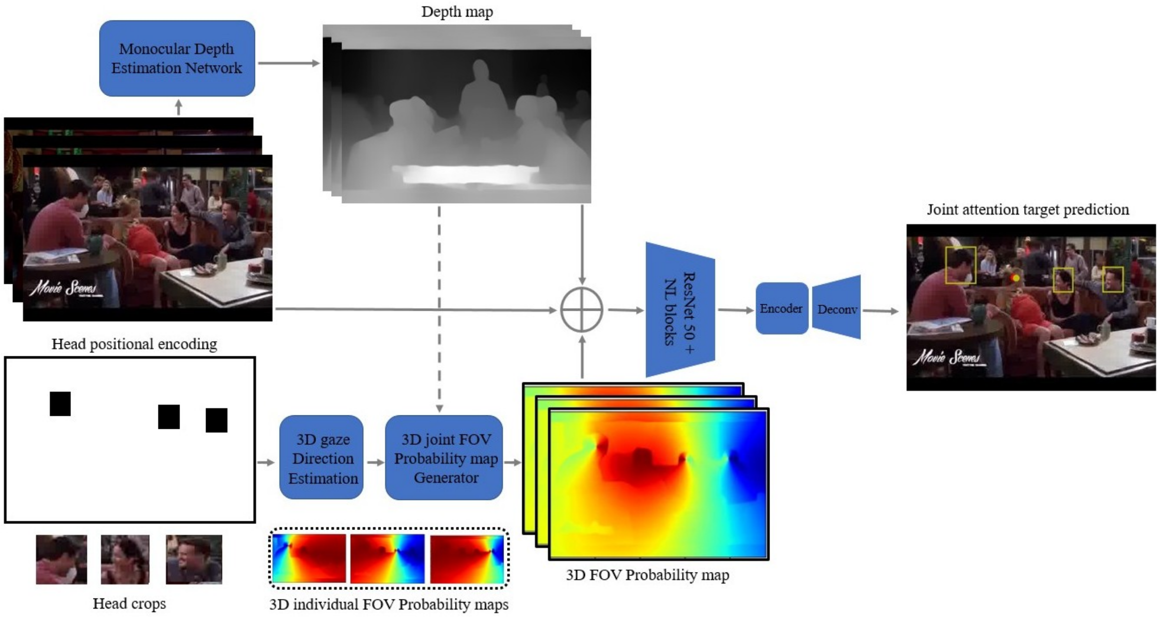

Monocular Depth Estimation Network
Depth map
Joint attention target prediction
ResNet 50 + NL blocks
Encoder
Deconv
Head positional encoding
3D gaze Direction Estimation
3D joint FOV Probability map Generator
3D FOV Probability map
Head crops
3D individual FOV Probability maps

Figure 4.8: **Overall framework of the proposed JAT method.** The input of the framework is the RGB image and the head bounding box annotation of the subjects of interest. Head crops of the subjects are generated based on their head positional encoding and used as the input of the 3D gaze target estimator. The depth map, generated by the monocular depth estimation network, is used as the input of the 3D field-of-view (FOV) probability map generator alongside the estimated 3D gaze direction. The generated depth and 3D FOV probability maps and the original RGB image are then inputted into the Joint attention target estimation module to predict the location of the joint attention target of the selected subjects in the scene. The ground truth attention target location is shown in yellow, and the estimate of the proposed method is in red.

3D Field-of-View Module

The crop of the subjects' heads in the scene is used to estimate their 3D head orientation using an existing 3D gaze estimator module. The 3D direction estimate combined with the 2D spatial positional encoding of the head bounding boxes allows us to generate the subjects' 3D individual FOV (shown at the bottom of Figure 4.8). We generate a shared 3D FOV probability map for each image, including every subject. Based on the assumption that a person is more likely to look within their 3D FOV cone than to their blind spot, we assigned a higher probability for the joint attention targets to be within the intersections of the subject's 3D FOV cones, and we penalise the predictions which fall outside of the cones. In addition, based on our preliminary experimental results presented in Section 4.3.2 we assigned the lowest probability score to the subjects' head bounding box region. The generator outputs are

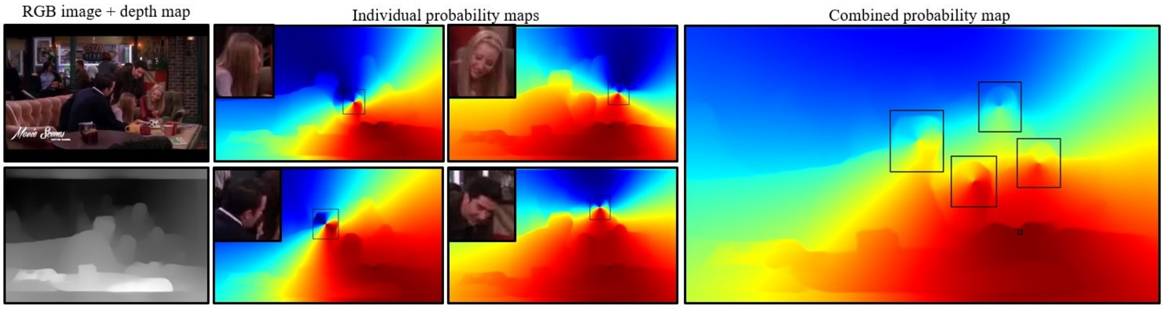


Figure 4.9: **Example joint 3D FOV probability map of the subjects looking at the same attention target.** We show the original input image and corresponding monocular depth map in the first column. Then we show the individual probability maps of each subject. Finally, on the combined joint attention probability map, we show the ground truth head bounding boxes and the attention target location in black.

joint 3D FOV probability maps corresponding to the input images. The individual probability map generation is mathematically denoted as:

$$M_{ind} = min\_max\_scaler\left(\frac{(i-h_x, j-h_y, k-h_z)\cdot(g_x, g_y, g_z)}{\left\|i-h_x, j-h_y, k-h_z\right\|_2 \cdot \left\|g_x, g_y, g_z\right\|_2}\right), \tag{4.1}$$

where ind={0,...,n} is the index of the subjects in the scene, (i,j,k) is the coordinate of each point in M_{ind},(h_x, h_y, h_z) is the centre of the head bounding box, (g_x, g_y, g_z) is the estimated 3D gaze direction, and min_max_scaler() is the transformation that scales each value of the probability map between zero and one. Then we set the values within the subject's head bounding box ($[x_{min}, x_{max}, y_{min}, y_{max}]$) to be equal to zero based on Section 4.3.2.

$$M_{ind}[x_{min}, x_{max}, y_{min}, y_{max}] = 0 \tag{4.2}$$

Finally, the joint attention probability map values are calculated as the average of the individual probability maps.

$$M_{FOV} = mean(M_{ind}) \tag{4.3}$$

An example visualisation of the generated individual and joint 3D FOV probability maps is shown in Figure 4.9. Four subjects are in the image, their head crops are highlighted at the left top corner of the individual heatmaps, and the positional encoding is visualised as black bounding boxes on the joint probability map. The ground truth JAT point is visualised on the joint probability map in black. We can see that the highest probability map values correspond to the area where the ground truth point is located within the scene. Furthermore, by using the relative depth prior information (shown at the left bottom of the figure), we can see the clear difference between the pixel values of the blind spot of the individuals and the FOV.

Joint Attention Target Prediction Module

Finally, we defined a JAT prediction module to localise the attention target point of the individuals in the scene. The input of this module is a series of scene images and the corresponding relative depth maps, and the calculated 3D joint FOV probability maps. These inputs are concatenated and fed into an encoder consisting of a ResNet-50 [154] followed by an additional residual and average pooling layer combined with NL layers [148] motivated by our findings discussed in Section 4.3.2. In between the layers of the second and third residual blocks, we included 3 and 5 NL layers, respectively. The concatenated features are encoded using two convolutional layers in the Encoder. A deconvolutional network composed of four deconvolution layers upsamples the features calculated by the Encoder into a full-sized feature map. We found that by combining the scene and the subject-dependent information using these inputs, we can find the most probable gaze target location on the image from the joint FOV of the subjects.

4.4.2 Implementation details

We implement our method in PyTorch. All experiments are run on an Intel i9-CPU @ 3.30GHz, 125 GB RAM, and four NVIDIA GeForce RTX 2080Ti GPUs.

The input RGB and generated depth images are resized to 224×224 and normalised. As described in [139], we used random flip, colour jitter, and crop augmentations. We also added noise to the head position and the 3D gaze direction during training to minimise the influence of localisation errors. The ground truth heatmap was generated using the output 3D direction estimate calculated by [152] and by adding Gaussian weight around the centre of the target for supervision. We implemented two loss functions during training: heatmap and in-frame loss. We used MSE loss to compute the heatmap loss ($\mathscr{L}_h$) and binary cross-entropy loss for the in-frame loss ($\mathscr{L}_f$). The total loss $\mathscr{L}$ used for training is a weighted sum of these two: $\mathscr{L} = w_h \mathscr{L}_h + w_f \mathscr{L}_f$.

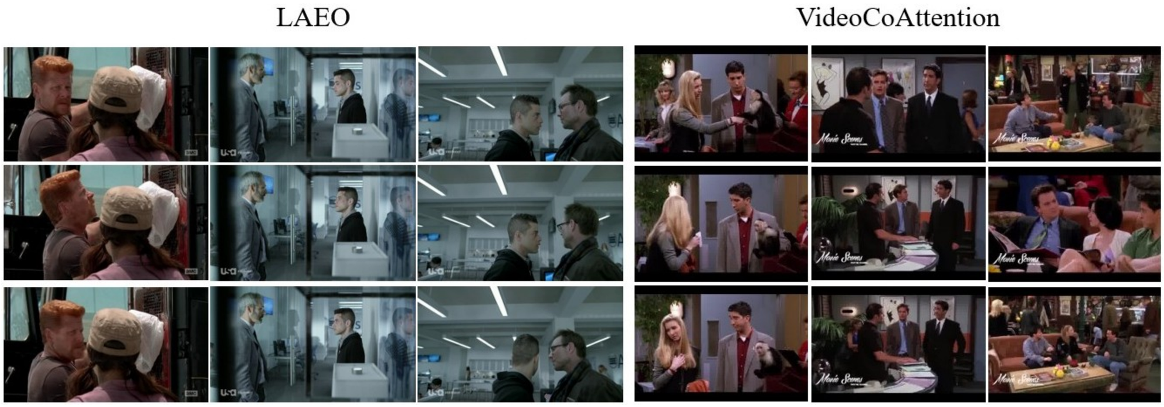


Figure 4.10: **Social interaction detection benchmark dataset highlights.** On the left we show example image frames of video sequences from the LAEO [129] dataset and on the right, we show examples from the VideoCoAttention JAT estimation benchmark dataset [133].

4.5 Experiments

4.5.1 Dataset and baselines

Benchmark datasets - Social interaction detection

In Section 4.3.1, we introduced the existing single attention target estimation datasets: the GazeFollow image [139] and VideoAttentionTarget [30] video benchmark datasets. Two popular benchmark datasets are available for social interaction detection: LAEO [129] and VideoCoAttention [133] video datasets.

Looking At Each Other (LAEO) dataset [129] The video dataset proposed by Marin *et al*. was used to train a model which can analyse one-to-one social interactions between subjects. The primary question they were trying to answer was whether the subjects looked at each other (See examples in Figure 4.10). The data consist of three types of annotations: a binary label indicating the presence of any pair of people looking at each other, the head bounding boxes of the subjects present at the scene, and, if they exist, the indices of the subjects looking at each other. The dataset is limited to human-human interactions and bounding box-level annotations; no pixel-wise gaze target point is available. The dataset does not extend to cases with more than two participants; therefore, while there are multiple subjects in the scene, joint attention, as we defined it in this thesis, does not exist in this dataset.

VideoCoAtt dataset [133]. A more detailed, larger video dataset proposed by Fan *et al*. was proposed for training models to estimate the joint attention target of the subjects in the video frames. This large-scale, diverse dataset consists of 492,100 frames from 380 video sequences of 20 different shows. For every frame, they collected the bounding box of joint attention. The attention targets occluded or outside of the frame were not annotated. In addition, they collected the head bounding boxes of the currently engaged subjects within the image frame. The drawbacks of this dataset are that not every person is annotated within the scene, and only one attention target bounding box is identified per image frame. VideoCoAtt dataset highlights are shown in Figure 4.10.

4.5.2 Additional Evaluation Metrics

The task of Attention Target Detection consists of two subtasks: spatial location prediction and temporal interval detection. To evaluate and compare the performance of the joint attention target prediction models, we used the L2 distance for the localisation task and reported the Prediction Accuracy for the detection task.

- **L2 distance:** Using the predicted joint, joint attention confidence map, we compute the distance between the pixel location of the maximum confidence and the centre of the ground truth bounding box.
- **Prediction Accuracy:** We regarded the given frame with joint attention when the predicted confidence map's maximum value was above a threshold adopted from [133]. The Prediction Accuracy is calculated as the percentage of the frames with correct joint attention estimation.

4.5.3 Baselines

The proposed method was evaluated using three benchmark datasets on both the single and joint attention target detection tasks. We used the datasets and methods introduced in Section 4.3 for the single target estimation evaluation. On the joint attention Target estimation task, we compare our method against the following methods on the VideoCoAtt dataset [133]: **Fan [133]** the first method proposed to infer joint attention in social scene videos, **VAT [30]** a single attention target estimation method used on videos, and **Attention Flow [137]** an End-to-End Joint Attention Estimation method.

4.5.4 Comparison with the state-of-the-art

For the most exhaustive comparison, the proposed joint attention target estimation model is evaluated and compared against both single and joint attention target estimation methods. We present the quantitative and qualitative results of our experiments using three previously introduced benchmark datasets.

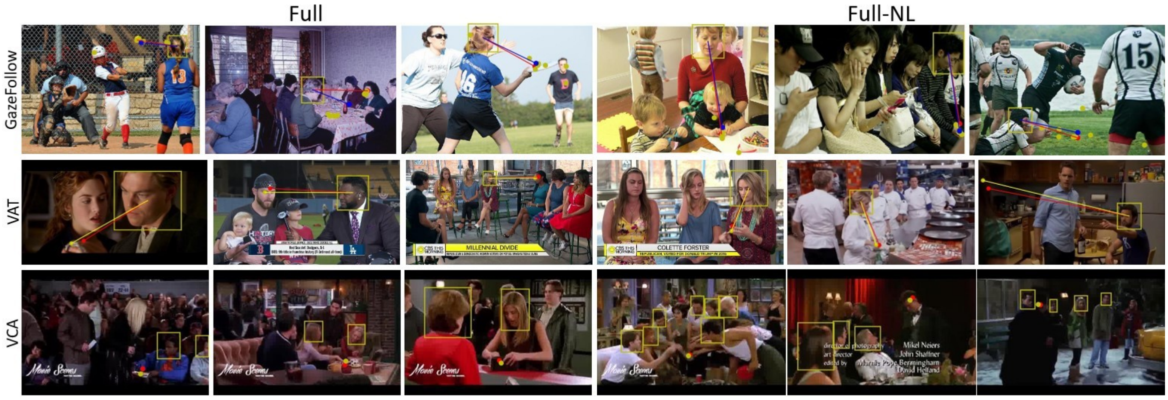


Figure 4.11: **Qualitative highlights of the proposed method with NL layers (Full-NL) and without (Full)** on three attention target estimation datasets: GazeFollow, VAT, and VCA. In the presented examples, the head bounding box of the observed subjects and the ground truth annotations are marked as yellow, the average is shown in blue, and the estimated gaze target estimate is shown in red.

Qualitative results

The qualitative highlights of the proposed Full and Full-NL methods on three benchmark datasets, GazeFollow, VAT and VCA, are shown in Figure 4.11. These examples were selected to demonstrate the efficiency of our method in different scenarios, *e.g.* in case of occlusion and ground truth ambiguity. Furthermore, we selected cases when the gaze target was not another person in the scene to address the human bias problem mentioned earlier in Section 4.3.

The attention target estimate of our methods is shown in red, the ground truth annotations and the head bounding boxes of the observed subjects are yellow, and for the GazeFollow dataset, we show the average ground truth annotation location in blue. These examples show that even in challenging cases *e.g.* when the subject's face is occluded or not visible or when the person is surrounded by many people around them and is looking at an object or in the middle of the field, our models successfully identified the attention target location.

Qualitative results are shown in the last two columns of Figure 4.12, 4.13 and 4.14. In the first column of these figures, we show the original input image, the prior depth map and the 3D FOV probability map for each example. Following these in the penultimate column, we show the predicted heatmap of Full and Full-NL. In the last column, we visualised the output target prediction, the ground truth annotation and the subject's head bounding box. We observed that the two models produce very similar heatmaps reflected in the calculated AUC scores presented in Section 4.5.4. We found that the Full-NL model generated more confident and correct heatmaps when the scenario was complex (See Figure 4.12 and 4.13

first examples). This might be due to the NL layers' ability to capture long-range spatial dependencies. For less complicated cases, *e.g.* last two rows of Figure 4.13, we observed that the use of the NL layers introduced unwanted dependencies.

Quantitative results

Here, we present the results of the quantitative evaluation. Note that the evaluation metrics differ for each benchmark dataset. For more details on the metrics, see Section 4.3.1 and 4.5.2. The type of attention target estimation tasks and benchmark datasets organise the results in this section.

Table 4.10: **Quantitative evaluation and ablation study results and comparison with the state-of-the-art methods on the GazeFollow SAT estimation image dataset.** Gaze direction estimation error shows the range of the random noise added to the 3D gaze direction of the subjects before the probability map generation.

Method	AUC ↑	Min distance ↓	Avg distance ↓
Model 92 (Sec. 4.3.2)	0.921	0.147	0.083
Scene only	0.889	0.143	0.213
Scene + depth	0.894	0.136	0.205
Scene + prob	0.928	0.036	0.084
Full (scene + depth + prob)	**0.932**	**0.036**	**0.082**
NL + Scene only	0.883	0.148	0.216
NL + Scene + depth	0.894	0.136	0.204
NL + Scene + prob	0.925	0.033	0.082
Full-NL (NL + scene + depth + prob)	**0.926**	**0.028**	**0.075**
Full gaze dir error ± 13.5°	0.930	0.052	0.100
Full gaze dir error ± 30°	0.927	0.047	0.097
Full-NL gaze dir error ± 13.5°	0.932	0.039	0.087
Full-NL gaze dir error ± 30°	0.929	0.049	0.099
HGTTR [155]	0.905	0.065	0.138
VideoAttention [30]	0.921	0.077	0.137
DAM [32]	0.922	0.067	0.124

SAT estimation on the GazeFollow dataset. The quantitative results on the GazeFollow dataset are shown in Table 4.10. We compared the performance of our method in the SAT estimation task with the latest methods HGTTR [155], VideoAttention [30], and DAM [32] on this image benchmark dataset. Among these, HGTTR and DAM were specifically designed to solve this task and similar to our solution, DAM used partial, relative depth prior

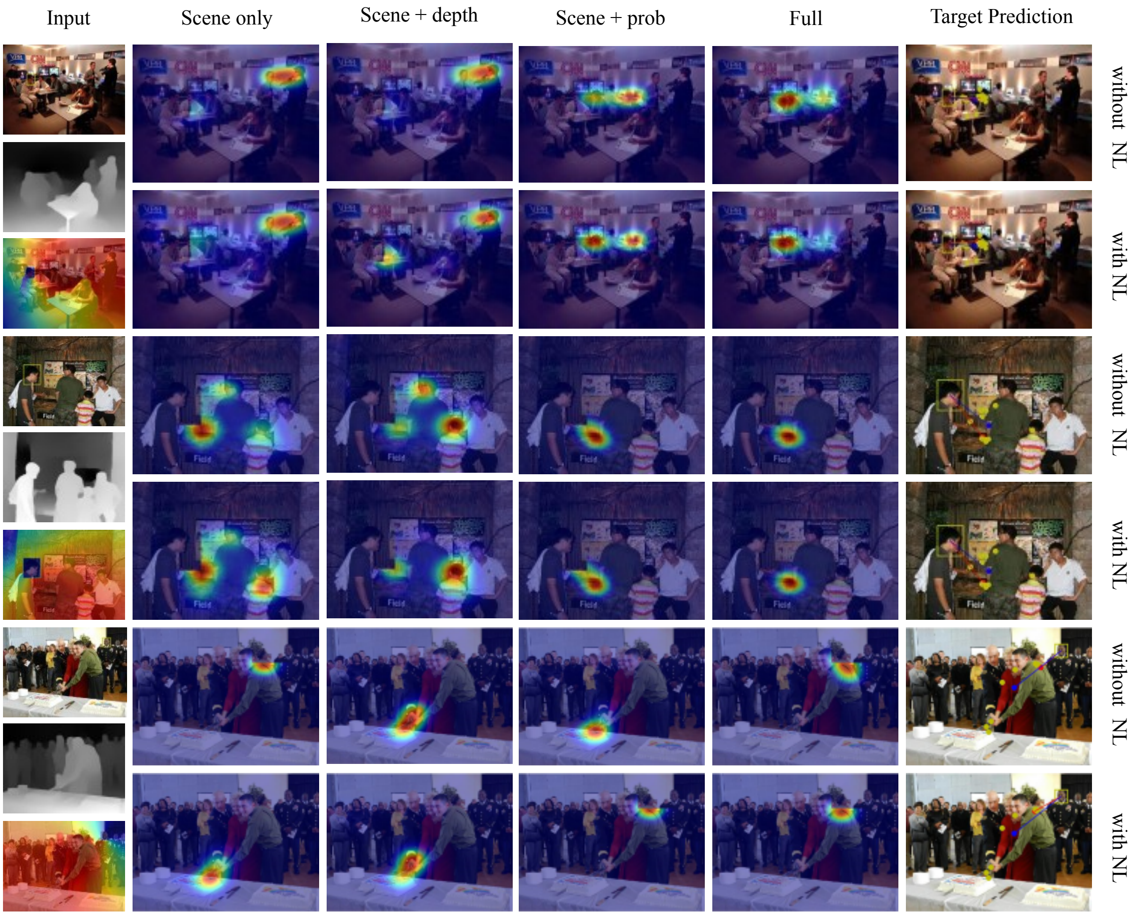


Figure 4.12: **Qualitative results of ablation study on the GazeFollow SAT benchmark image dataset.** The input of the Full and Full-NL proposed models and their variants, the RGB image, the generated prior depth map and the corresponding calculated 3D FOV probability map are shown in the first column. We visualised the generated output heatmap of every variant (Scene only, Scene+depth, Scene+prob) and the Full method (Scene+depth+prob) and, finally, the gaze target prediction of the Full method. In the target prediction visualisation and the input image, the head bounding box of the observed subjects and the ground truth annotations are marked as yellow, and the average is shown in blue, and the estimated gaze target is shown in red.

information in their method. The results highlighted in Table 4.10, show that in terms of all the evaluation metrics, the proposed framework with and without the NL layers outperformed all the existing methods. Furthermore, when using the NL layers in the framework, 58.20% minimum distance and 39.52% average distance relative improvement were achieved by our method on the GazeFollow dataset.

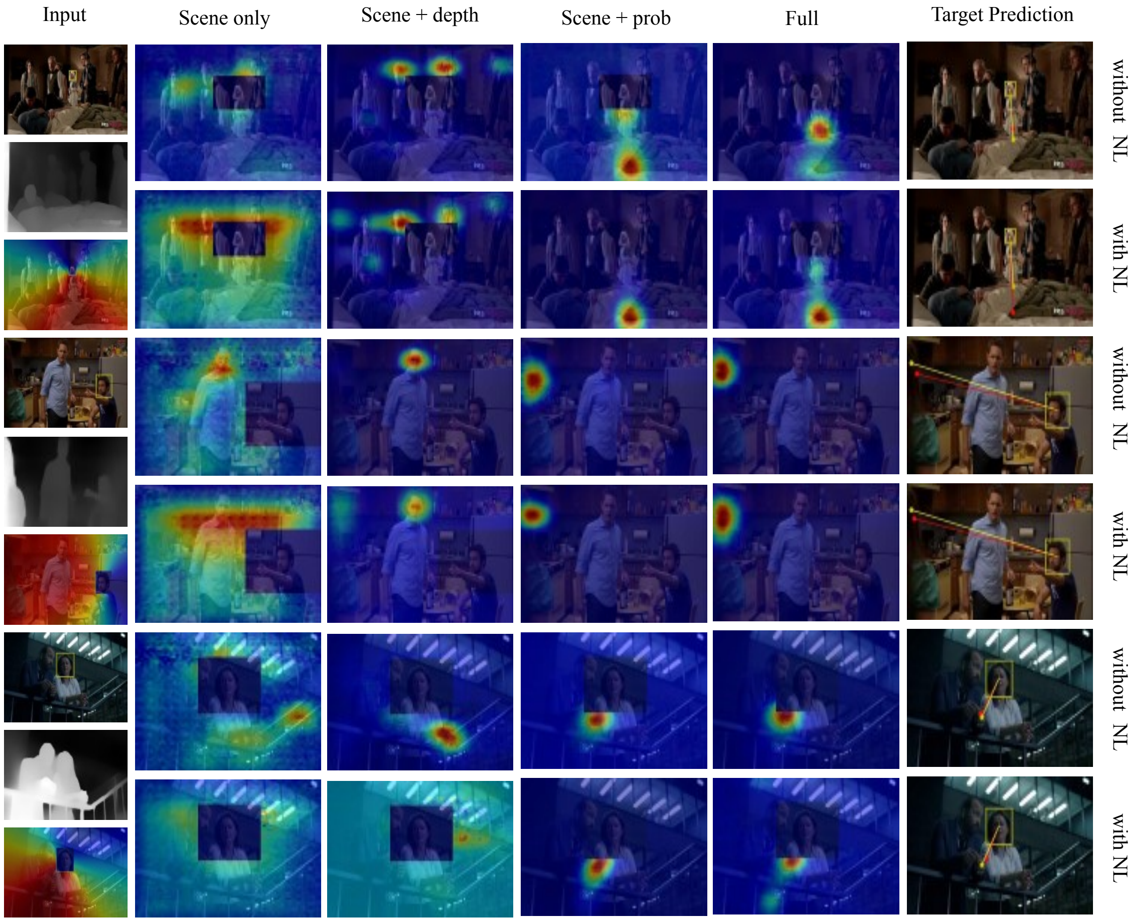


Figure 4.13: **Qualitative results of ablation study on the VAT SAT benchmark video dataset.** The input of the Full and Full-NL proposed models and their variants, the RGB image, the generated prior depth map and the corresponding calculated 3D FOV probability map are shown in the first column. We visualised the generated output heatmap of every variant (Scene only, Scene+depth, Scene+prob) and the Full method (Scene+depth+prob) and, finally, the gaze target prediction of the Full method. In the target prediction visualisation and the input image, the head bounding box of the observed subjects and the ground truth annotations are marked as yellow, and the estimated gaze target is shown in red.

Note that the performance of the proposed method with (Full-NL) or without (Full) the additional NL layers is very similar. We found that in the presence of relative depth prior information, the NL layers did not contribute towards the performance significantly. Furthermore, we included the results of Model 92 (See Section 4.3.2) for reference. In comparison to Model 92, we found that the proposed 3D FOV probability map was more useful than the inserted NL layers. This is not surprising as the probability map contains subject-specific information.

Table 4.11: **Quantitative evaluation and ablation study results and comparison with the state-of-the-art methods on the VAT SAT estimation video dataset.** Gaze direction estimation error shows the range of the random noise added to the 3D gaze direction of the subjects before the probability map generation.

Method	AUC ↑	L2 distance ↓
Scene only	0.711	0.306
Scene + depth	0.728	0.313
Scene + prob	0.935	0.082
Full (scene + depth + prob)	**0.937**	**0.077**
NL + Scene only	0.713	0.318
NL + Scene + depth	0.743	0.334
NL + Scene + prob	0.944	0.082
Full-NL (NL + scene + depth + prob)	**0.951**	**0.074**
Full gaze dir error ± 13.5°	0.930	0.134
Full gaze dir error ± 30°	0.911	0.122
Full-NL gaze dir error ± 13.5°	0.943	0.093
Full-NL gaze dir error ± 30°	0.914	0.153
VideoAttention [30]	0.860	0.134
HGTTR [155]	0.904	0.126
DAM [32]	0.905	0.108

SAT estimation on the VAT dataset. Furthermore, we compared our solution with the same methods on the SAT estimation using the VAT video benchmark dataset. The results are shown in Table 4.11. We compared our performance with the previously mentioned HGTTR [155], VideoAttention [30], and DAM [32] methods. Both proposed methods were more efficient at estimating the gaze target than the state-of-the-art methods. We found that Full-NL outperformed the performance of the Full method in terms of both AUC and L2 distance measures. Full-NL improved the AUC score by 4.84 % and the L2 distance by 31.48 %.

JAT estimation on the VCA dataset. Finally, the quantitative results on the VCA video benchmark dataset are shown in Table 4.12. We compared the JAT estimation performance of our method with Fan [133], Sumer [137], VideoAttention [30], and HGTTR [155]. Among these state-of-the-art methods Fan, Sumer and HGTTR were trained to estimate the attention target location of multiple subjects in the scene. The results showed that the proposed method with and without the NL layers significantly outperformed all the state-of-the-art methods in terms of the L2 distance metric. We were able to reduce the distance error by 71.74 % compared to the result report by [155]. The NL layers proved useful in achieving the best

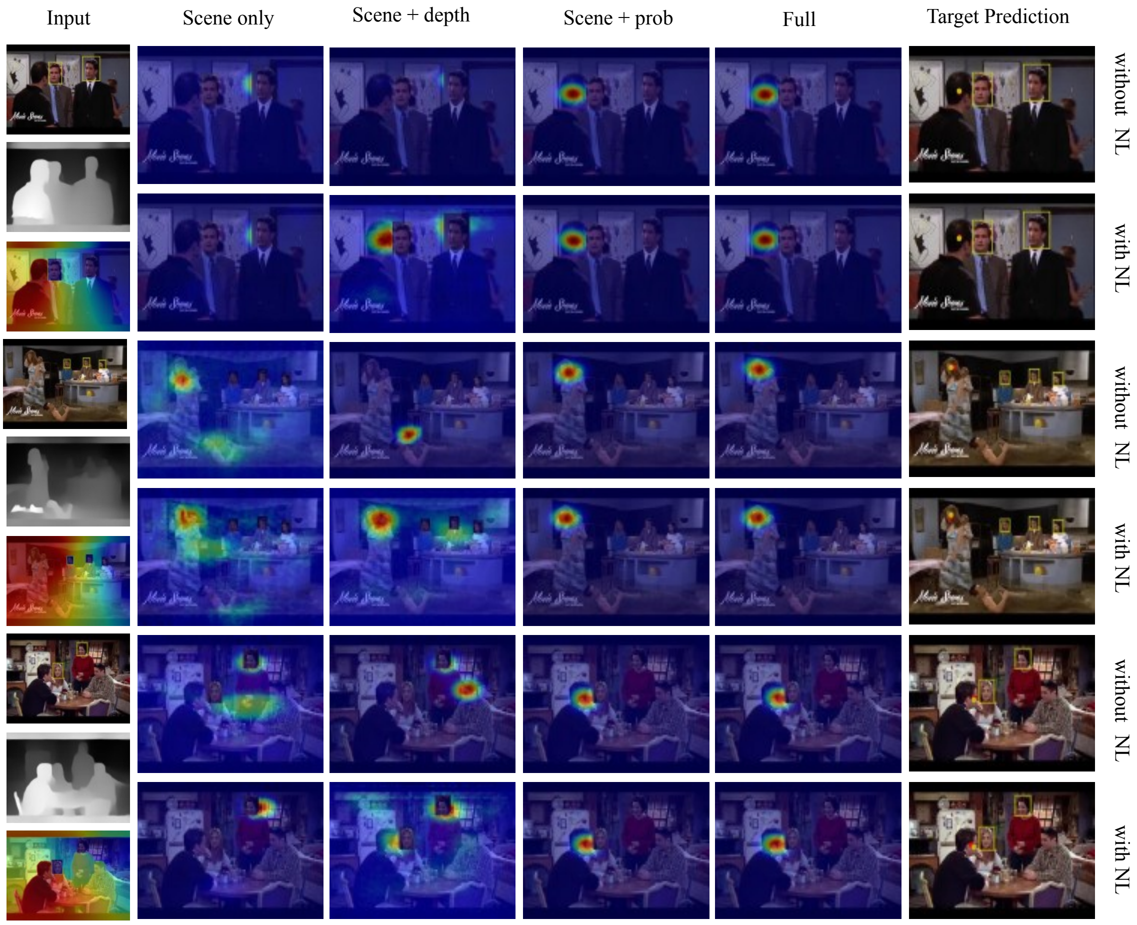


Figure 4.14: **Qualitative results of ablation study on the VCA JAT benchmark video dataset.** The input of the Full and Full-NL proposed models and their variants, the RGB image, the generated prior depth map and the corresponding calculated 3D FOV probability map are shown in the first column. We visualised the generated output heatmap of every variant (Scene only, Scene+depth, Scene+prob) and the Full method (Scene+depth+prob) and, finally, the gaze target prediction of the Full method. In the target prediction visualisation and the input image, the head bounding box of the observed subjects and the ground truth annotations are marked as yellow, and the estimated gaze target is shown in red.

prediction accuracy. Overall, the proposed method achieved state-of-the-art performance in terms of all evaluation metrics on this dataset too.

In summary, the quantitative results confirmed that the JAT estimation method proposed in Section 4.4 achieved state-of-the-art performance across all the benchmark datasets and their evaluation metrics on the SAT and JAT estimation tasks. Furthermore, the comparison between Full and Full-NL across the datasets shows that the usefulness of the NL layers is context and complexity dependent.

Table 4.12: **Quantitative evaluation and ablation study results and comparison with the state-of-the-art methods on the VCA JAT estimation video dataset.** Gaze direction estimation error shows the range of the random noise added to the 3D gaze direction of the subjects before the probability map generation.

Method	L2 distance ↓	Pred. Acc. ↑
Scene only	139	13.0
Scene + depth	130	41.5
Scene + prob	16	90.0
Full (scene + depth + prob)	**13**	90.2
NL + Scene only	145	17.1
NL + Scene + depth	145	31.8
NL + Scene + prob	14	93.0
Full-NL (NL + scene + depth + prob)	**13**	**93.2**
Full gaze dir error ± 13.5°	21	83.5
Full gaze dir error ± 30°	24	50.1
Full-NL gaze dir error ± 13.5°	19	80.8
Full-NL gaze dir error ± 30°	34	66.2
Fan [133]	62	71.4
Sumer [137]	63	78.1
VideoAttention [30]	57	83.3
HGTTR [155]	46	90.4

4.5.5 Ablation Study

To study the contribution and effectiveness of different components of the proposed method, we trained several models with different parameters. In this section, we discuss the findings of these experiments on three benchmark datasets.

Spatial model components

We trained the following variations of the proposed full spatial method: Scene only, Scene+depth, Scene+probability map, Scene+depth+probability map, and their variants, including the non-local layers in the encoder. Qualitative highlights are shown in Figure 4.12, 4.13, and 4.14. Note that the observations discussed below are accurate for all the benchmark datasets.

Across all the benchmark datasets, we found that the Scene only variant performed the worst compared to the other variants. The heatmaps in the second column of the qualitative highlights figures also confirmed that the predicted output heatmap of this module alone, most of the time, did not overlap with the gaze target area of the image, and it was widespread and not confident. This is because the model was unaware of any subject-specific information; therefore, it relied solely on the scene information to estimate the subject-dependent attention target location.

We also found that when we combined the scene information with the output of the prior depth map of the monocular depth estimator, the performance of the trained Scene+depth models improved slightly. The estimated output heatmaps of these models (See the third column of qualitative figures) were more successful in localising the FOV of the subject. These heatmaps were more confident; however, they often misplaced the gaze target as it was selected based on the Scene information. Therefore, despite this improvement, as the input of these models was still subject-independent, their results were not satisfactory.

The proposed 3D FOV probability map contains depth and subject-dependent information. Introducing subject-dependent information into the model significantly improved the performance quantitatively and qualitatively. The output heatmaps of Scene+prob, shown in the fourth column of Figure 4.12, 4.13, and 4.14, are concentrated around the correct gaze target location, the location of the maximum of these heatmaps shifted significantly from the predictions of the Scene only and Scene+depth models.

Finally, we can see that explicitly using the prior depth map as an input and not just as a part of the probability map further improved the results. While the improvement was moderate compared to the Scene+prob performance. However, the Scene+depth+prob with (Full-NL) or without (Full) the NL layers proved to be the most efficient in estimating the attention target of single or multiple subjects within the scene.

In summary, the ablation study confirmed that all the modules included in the Full model (Scene+depth+probability map) are useful and contribute to the proposed solution's performance. We demonstrated that relying only or too much on the scene information is insufficient to estimate the subject's gaze target location.

Gaze direction estimation error

The proposed probability map's role in the outstanding performance of the proposed method has been demonstrated through the previous experiments. The input of the 3D FOV probability map is the 3D gaze target estimate of the observed subject. To test the robustness of the proposed method against gaze direction prediction errors, we trained two variants of the Full and Full-NL models under extreme error levels.

During this experiment, we generated the 3D FOV probability map using additional random noise added to the subjects' estimated gaze direction. We chose the noise levels to reflect the average error (± 13.5°) of the state-of-the-art 3D gaze direction estimation method

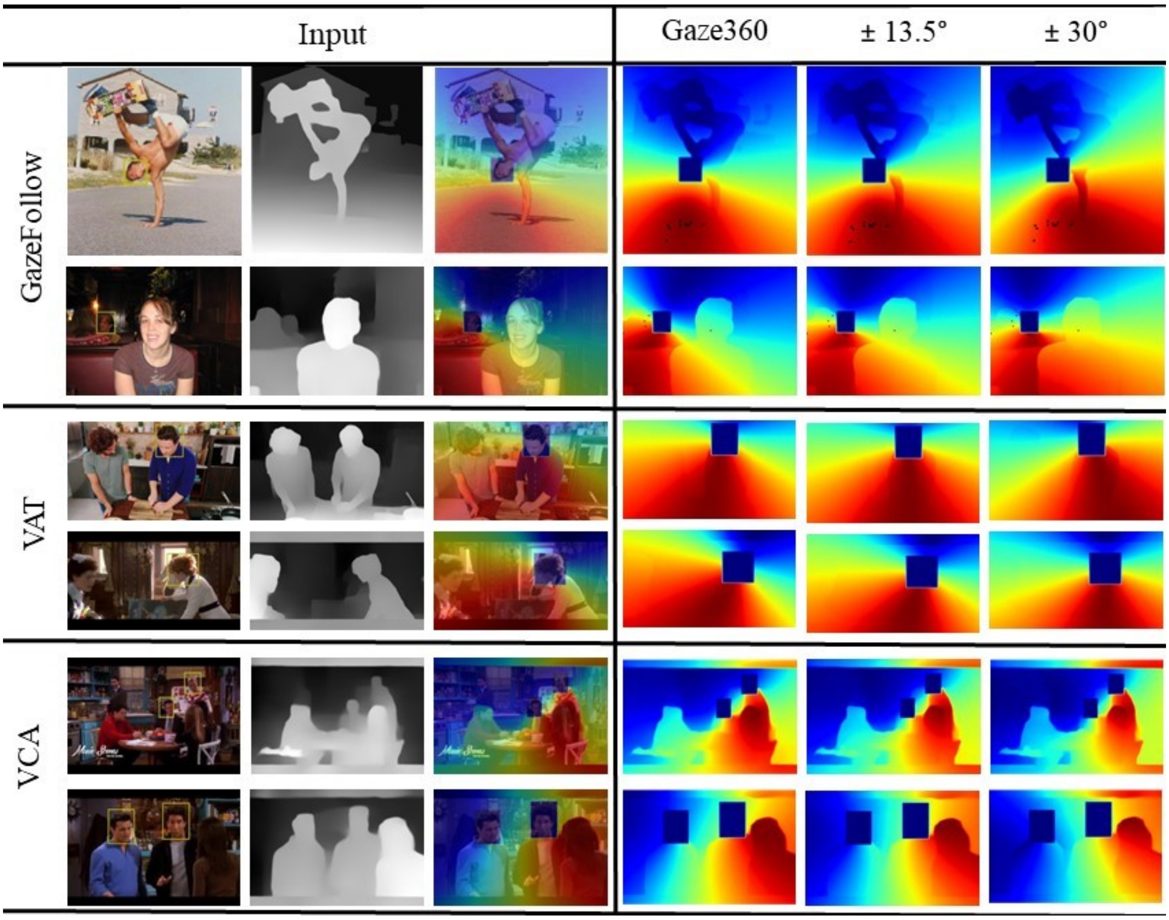


Figure 4.15: **Visualisation of the joint 3D FOV probability map effected by gaze direction error.** We show the generated probability map using the 3D gaze estimator Gaze360 [152] and when additional 13.5 and 30 degree gaze direction error was added.

[152] and to reflect the human's horizontal central vision range (± 30°). We show example probability map variants generated with additional gaze direction error in Figure 4.15.

While adding NL layers to the proposed method did not improve the performance significantly under moderate gaze estimation error in the previously presented experiments, our results show that the Full-NL models were more robust against the additional noise than the Full models. Furthermore, we found that in the case of the large-scale GazeFollow image dataset (See Table 4.10) and the VAT video dataset (See Table 4.11) the proposed model surpassed the performance of the state-of-the-art methods even when we added ± 30° gaze direction estimation error, which is more than double the existing 3D gaze estimators' average angular error. These results on the SAT estimation task are especially outstanding as the proposed 3D FOV probability map is the most useful in improving the robustness of the attention target estimation when there is more than one subject within the scene.

4.6 Conclusion

In this chapter, we proposed a novel joint attention target estimation framework which was developed to fully utilise the 3D clues of the scene efficiently. Following the findings of our preliminary experiments, we aimed to tackle the human bias and physically impossible predictions, which are the major flaws of the previously proposed models. To achieve this, we proposed to combine a novel 3D field-of-view-based joint attention probability map with the scene and depth information. Extensive qualitative and quantitative analysis on three benchmark datasets shows that the proposed method achieved favourable performance compared to both the state-of-the-art single and joint attention target estimation approaches. The demonstrated outstanding performance of the proposed method proved our hypothesis that using 3D clues for the third-person view attention target estimation is advantageous.

Chapter 5

Conclusion

This thesis investigated the benefits of integrating verbal and non-verbal social clues into different computational frameworks designed to solve human-centred computer vision tasks. In the experiments presented in this work, natural language expressions and human gaze information and their combinations were used. This work aimed to demonstrate the usefulness of multimodal systems and explicit and implicit human input.

The first objective of this work was to use the user's verbal input for description-based automatic image cropping. We sought a solution to a highly subjective task during this experiment by re-purposing different existing models. Besides producing an output crop which best preserves the user caption information, we aimed to generate aesthetically pleasing, high-quality output.

The experimental results presented in Chapter 2 proved that integrating user captions into the automatic image cropping is advantageous and that despite the challenges of this task, the proposed CAGIC method can produce image crops that the users preferred compared to those produced by the previous methods. Human annotation generation, such as user captions and corresponding ground truth image crops, for this task is highly time-consuming. Therefore, due to the lack of training data, we proposed re-purposing existing models and integrating them into a novel multimodal optimisation framework instead of training a new model specifically designed for this task. The input of the framework consists of an image and the user caption describing a part of the image. To process and utilise the verbal human input, we integrated an existing CN [50] into the proposed framework. In every iteration of the optimisation, we select a new image part based on different optimisation techniques presented in Section 2.3.3, and we use this image region as the input of the CN to generate a corresponding description. A new caption loss function was designed to measure the similarity between the original user caption and the generated caption in every iteration. Besides the CN we also used an Aesthetic Network [46] to generate an aesthetic score of the selected image region. Combining the calculated caption and aesthetic scores ensured that the final output reflected the user's intention and high quality. Through this work, we demonstrated that integrating the user caption into the CAGIC framework led to favourable results and state-of-the-art performance. Due to the algorithm's iterative nature, this solution's main limitation is that the runtime required to produce a single output crop is large. Since in every iteration, we slowly converge from the full input image to the region of interest, this method currently takes approximately 5 minutes to generate an image crop based on the user's description. As the user can describe any part of the image regardless of its relative area to the image size, it is crucial to use several iterations to ensure that we find the correct crop size and location.

The second contribution of this work investigated the usefulness of a combined gaze and description-based multimodal framework for the task mentioned above. Motivated by the success of the proposed CAGIC framework, we aimed to tackle its runtime limitation. We implicitly sought to collect extra information from the users during the image description without them performing additional actions such as pointing or clicking. A complex cross-modal alignment exists between language and vision during image description [121, 156]. During description generation, the perceived visual information guides the description generation [157]; consecutively, the task influences the user's eye movements [158]. Therefore, we chose to couple the natural language descriptions generated by the user with gaze data collected during the image description. We used an unobtrusive monitor-mounted eye-tracking device to record the gaze data surreptitiously.

The study presented in Chapter 3 demonstrated the usefulness of additional user input through the image cropping performance and runtime improvement. We used the gaze information to initialise the proposed G-DAIC framework. Utilising human gaze information is challenging as it could introduce noise and further subjectivity into the experiments. We proposed different start region generation methods discussed in Section 3.3.1 to prevent this. With the rough initialisation of the image crop search and the proposed early termination strategy (See Section 3.4.4), we were able to reduce the number of iterations required to produce the output image crops by 90.07%. Beyond the overall 92.11% runtime decrease, we demonstrated state-of-the-art performance through extensive qualitative and quantitative analysis. To achieve this positive outcome, we further proposed an adaptive scaling method in Section 3.3.2. This new scaling procedure was essential due to the design differences between CAGIC and G-DAIC. As opposed to iteratively converging from the full image to the described image part, we had to account for the possibility of the calculated start region of G-DAIC being smaller than the region of interest. Therefore, in some cases, it was inevitable to expand the search field instead of shrinking it. The proposed solution utilised an experimentally chosen, pre-defined threshold value to determine the nature of the scaling (shrinkage or expansion) based on the size of the start region *w.r.t.*the original image size.

Encouraged by the rich information collected through the non-verbal social clue and its usefulness in the proposed G-DAIC multimodal system, we further aimed to investigate how gaze information could be used to localise the interest of the subjects. As the final contribution of this work, we aimed to rely on only implicit social clues to estimate the joint attention target of the subjects within the 3D scene from a third-person view. The goal of the emerging research field of human-centred computer vision, attention target estimation, is to estimate what the subjects are looking at in the wild. Using visual data from the wild is challenging as, for example, the subjects are completely unrestricted, and occlusions can occur frequently. Therefore, we need large-scale data to train a model to solve a complex problem in a challenging environment. The image and video datasets presented and used in this study were collected from the web and TV shows for this task, and humans annotated them. Due to the diversity of the visual data, we designed a deep neural network that efficiently learns subject-dependent attention targets. As in our everyday life, we are a part of a 3D environment; we leverage additional information from the depth of the 3D scene and utilise it to make better estimates. When we deal with image and video data, this depth information is not automatically provided; therefore, previously existing methods did not or not fully used this information. The preliminary experiments highlighted several cases challenging the existing attention target estimation models, such as occlusion, bias towards humans and physically impossible predictions.

We aimed to overcome these challenges through the work presented in Chapter 4. We presented our solution: a depth-aware joint attention target estimation model designed to predict the location of the joint interest point of two or more people in the scene. The input of our model is an image frame or video sequence and the head bounding box annotations indicating the position of the subjects within the scene looking at the same attention target. The proposed framework is the combination of the Relative Depth Prior Module, the 3D FOV Module, and the JAT Prediction Module. These components are introduced in Section 4.4 in detail. Using the depth prior module, we generated the corresponding depth map of every frame generated by an existing monocular depth estimator [138]. The 3D gaze direction of each subject was estimated by [152]. The quantitative and qualitative ablation study results showed that the most significant part of the proposed method is the 3D FOV Module. This novel fusion of the scene and subject-dependent 3D information significantly improved the model's performance. Note that unlike the 2D FOV information used by [32] and the 3D FOV formulation of [143] which are limited by a predefined angle or number of visual rays used to generate the subject's FOV, we formulated the 3D FOV as a probability map and assigned a probability score to each pixel. This formulation of the probability map made our method robust against 3D gaze prediction estimation errors. This advantage has been demonstrated and discussed in Section 4.5.5. For the most comprehensive evaluation of our JAT method, we compared our performance on the GazeFollow SAT image, the VAT SAT video, and on the VCA JAT benchmark datasets. The proposed JAT method achieved state-of-the-art performance not only on the JAT task but also outperformed the existing SAT method on both SAT benchmark datasets.

In summary, the work presented in this thesis has taken steps towards the three aims of this project. The methods proposed in the thesis achieved state-of-the-art performances across multiple benchmark datasets and in different human-centred computer vision tasks. Through these works, I have confirmed the effectiveness of user input integration into the computational models. While the results presented in this thesis are encouraging, these solutions have some technical limitations.

First, the optimisation of the CAGIC and G-DAIC methods rely on the caption loss. This loss term is calculated using the output of CN and computing its similarity to the user's input caption. However, the captions generated by CN are limited by its fixed vocabulary size of 10,000 words [50]. This limitation results in errors where the user, for example, attempts to describe the part of the image containing a "Baileys" (Irish cream liqueur brand), but the CN can only generate the words "bottle", "wine" or "champagne". For most humans describing objects by their brand is satisfactory automatic caption generation methods cannot handle these cases, resulting in a high caption loss assigned to the region of interest. Therefore, although we took a significant step towards integrating natural language descriptions into multimodal frameworks, there are exceptions where our solutions fail. This limitation is possible to address due to the re-purposing natural of the CAGIC and G-DAIC methods. Namely, the proposed frameworks use an existing CN, which is replaceable by a novel caption generation model when researchers in the field of Natural Language Processing develop a better and more complex one. Furthermore, CAGIC does not consider the order of the words which could potentially change the meaning of the sentence and prevent the method from correctly utilising the user description. This is especially problematic in case of referring expressions, such as the captions of Ref-COCO [83]. Finally the loss function introduced in Equation 2.4 could be replaced by NLP metrics used later in this thesis during the caption-crop consistency calculation.

In addition, we showed qualitative highlights of these methods in Chapter 2 and 3. While the results produce by CAGIC and G-DAIC are better than the other baseline methods in terms of both quantitative and qualitative measures, there is room for improvement and further analysis might be required to better understand the pitfalls of the methods. In Figure

2.10 we showed that CAGIC is able to produce different output crops based on the user captions from the same image. In the last column of this figure we showed two examples where the method successfully identified the object of the user description, however, the positioning of the image crop is not ideal as they are not centralised. Based on these examples, it is possible that CAGIC is biased towards colours, such as "red" and "green" in the example user captions of Subfigure c and f. Due to the limited amount of data currently available, this cannot be proved at the moment. In Figure 2.13 and 2.14, we show additional qualitative highlights of CAGIC. Note that as mentioned in Section 2.4.1 the captions used in this experiment are not optimal for our task. While the results are not optimal is every case, note that in several occasions CAGIC produced better results compared to, the method trained on this dataset, MattNet [47].

Furthermore, the proposed G-DAIC method builds on the existing relationship between language description and gaze. The temporal connection between the modalities was introduced in Section 3.4.1. Our solution is currently limited by using the gaze points collected during Stimuli without taking advantage of the temporal information. Machine translation alignment models could be trained on a linearly ordered sequence of visual and linguistic units to find the temporal alignment between the recorded fixation points and the language expression [121]. For training such a model on our data, one might use the tokens and the corresponding gaze points based on the timestamp annotations we provided (See example in Section B.2). In this work, we did not attempt to train an alignment model for our task due to the existing challenges and limitations of this research area described in [99]. Namely, the dataset proposed in [33, 34] is too small to train a sequential alignment model. However, incorporating the temporal information, as discussed in Section 3.4.1, might prove useful to generate more reliable start regions for the G-DAIC method.

Finally, occasionally the proposed joint attention target estimation method's predictions for two consecutive image frames of a given video might vary significantly compared to the actual attention target shift between them. Our model relies only on spatial information to estimate the subjects' interest points; therefore, fluctuations can occur between the image frames. We think that the method could benefit from the temporal information of the video sequences; however, this has not been included in our work due to the limited length of the uncut image sequences of the VCA dataset. Alongside the temporal information of the scene, a robust temporal monocular depth estimator, such as [159, 160], could replace the one used in our framework for better alignment between the input image and the generated prior depth map. Furthermore, the method does not estimate out-of-frame cases as [30] does due to the lack of in/out annotations of the VCA dataset. However, this might be necessary to investigate in the future to build a model to predict the joint attention target of people in the wild. Finally, the confidence score estimated by the 3D gaze direction estimation module could be further utilised to fine-tune the proposed 3D FOV probability map.

This thesis investigated the automatic image cropping and joint attention target estimation tasks in detail. Outside of these applications, the proposed caption-based solutions could support learning and teaching by highlighting the described part of the slides during presentations. The combination of verbal and non-verbal social clues, such as gaze and speech, could improve the performance of smart home devices. The devices could better interpret the user's interest by relying on two modalities. Finally, the third-person view subject attention information could be utilised for action recognition and to detect potentially malicious behaviour. The methods presented in this thesis took a step towards better understanding the potential of user input integration into human-centred computer vision tasks and attempted to perform tasks as humans would.

Bibliography

[1] Ben Shneiderman. *Human-Centered AI.* Oxford University Press, 2022.

[2] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. Who is the" human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–32, 2019.

[3] Jan Auernhammer. Human-centered ai: The role of human-centered design research in the development of ai. 2020.

[4] Terry Winograd. *Bringing design to software*. ACM, 1996.

[5] Thomas Huang. Computer vision: Evolution and promise. 1996.

[6] David Daniel Cox and Thomas Dean. Neural networks and neuroscience-inspired computer vision. *Current Biology*, 24(18):R921–R929, 2014.

[7] Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon JL Billinge, et al. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8(1):1–26, 2022.

[8] Xizhao Wang, Yanxia Zhao, and Farhad Pourpanah. Recent advances in deep learning. *International Journal of Machine Learning and Cybernetics*, 11(4):747–750, 2020.

[9] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. In *Machine learning techniques for multimedia*, pages 21–49. Springer, 2008.

[10] Jiayi Luo and Rongjun Yu. Follow the heart or the head? the interactive influence model of emotion and cognition. *Frontiers in psychology*, 6:573, 2015.

[11] Grant Soosalu, Suzanne Henwood, and Arun Deo. Head, heart, and gut in decision making: development of a multiple brain preference questionnaire. *Sage Open*, 9(1): 2158244019837439, 2019.

[12] Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A Fox. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*, 2020.

[13] Felipe Zago Canal, Tobias Rossi Müller, Jhennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sobieranski. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582:593–617, 2022.

[14] Primesh Pathirana, Shashimal Senarath, Dulani Meedeniya, and Sampath Jayarathna. Eye gaze estimation: A survey on deep learning-based approaches. *Expert Systems with Applications*, 199:116894, 2022.

[15] Mukhiddin Toshpulatov, Wookey Lee, Suan Lee, and Arousha Haghighian Roudsari. Human pose, hand and mesh estimation using deep learning: a survey. *The Journal of Supercomputing*, 78(6):7616–7654, 2022.

[16] D Munteanu and R Ionel. Voice-controlled smart assistive device for visually impaired individuals. In *2016 12th IEEE international symposium on electronics and telecommunications (ISETC)*, pages 186–190. IEEE, 2016.

[17] Omer Saad Alkhafaf, Mousa K Wali, and Ali H Al-Timemy. Improved prosthetic hand control with synchronous use of voice recognition and inertial measurements. In *IOP Conference Series: Materials Science and Engineering*, volume 745, page 012088. IOP Publishing, 2020.

[18] Mirela Popa, Leon Rothkrantz, Zhenke Yang, Pascal Wiggers, Ralph Braspenning, and Caifeng Shan. Analysis of shopping behaviour based on surveillance system. In *2010 IEEE International Conference on Systems, Man and Cybernetics*, pages 2512–2519. IEEE, 2010.

[19] William D Wells and Leonard A Lo Sciuto. Direct observation of purchasing behaviour. *Journal of Marketing Research*, 3(3):227–233, 1966.

[20] Ismail Haritaoglu and Myron Flickner. Attentive billboards: Towards to video-based customer behaviour understanding. In *Sixth IEEE Workshop on Applications of Computer Vision, 2002.(WACV 2002). Proceedings.*, pages 127–131. IEEE, 2002.

[21] Zawar Hussain, Quan Z Sheng, and Wei Emma Zhang. A review and categorization of techniques on device-free human activity recognition. *Journal of Network and Computer Applications*, 167:102738, 2020.

[22] HR Chennamma and Xiaohui Yuan. A survey on eye-gaze tracking techniques. *arXiv preprint arXiv:1312.6410*, 2013.

[23] Andy J King, Nadine Bol, R Glenn Cummins, and Kevin K John. Improving visual behaviour research in communication science: An overview, review, and reporting recommendations for using eye-tracking methods. *Communication Methods and Measures*, 13(3):149–177, 2019.

[24] Amer Al-Rahayfeh and Miad Faezipour. Eye tracking and head movement detection: A state-of-art survey. *IEEE journal of translational engineering in health and medicine*, 1:2100212–2100212, 2013.

[25] Matteo Cognolato, Manfredo Atzori, and Henning Müller. Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances. *Journal of rehabilitation and assistive technologies engineering*, 5:2055668318773991, 2018.

[26] Stefan E Huber, Markus Martini, and Pierre Sachse. Patterns of eye blinks are modulated by auditory input in humans. *Cognition*, 221:104982, 2022.

[27] Aasef G Shaikh, Shlomit Ritz Finkelstein, Ronald Schuchard, Glen Ross, and Jorge L Juncos. Fixational eye movements in tourette syndrome. *Neurological Sciences*, 38 (11):1977–1984, 2017.

[28] Joanna N Lahey and Douglas Oxley. The power of eye tracking in economics experiments. *American Economic Review*, 106(5):309–13, 2016.

[29] Guobin Wan, Xuejun Kong, Binbin Sun, Siyi Yu, Yiheng Tu, Joel Park, Courtney Lang, Madelyn Koh, Zhen Wei, Zhe Feng, et al. Applying eye tracking to identify autism spectrum disorder in children. *Journal of autism and developmental disorders*, 49(1):209–215, 2019.

[30] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5406, 2020.

[31] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*, 2021.

[32] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11390–11399, 2021.

[33] Nora Horanyi, Kedi Xia, Kwang Moo Yi, Abhishake Kumar Bojja, Aleš Leonardis, and Hyung Jin Chang. Repurposing existing deep networks for caption and aesthetic-guided image cropping. *Pattern Recognition*, 126:108485, 2022.

[34] Nora Horanyi, Yuqi Hao, Aleš Leonardis, and Hyung Jin Chang. G-daic: A gaze initialized framework for description and aesthetic-based image cropping. *Proceedings of the ACM on Human-Computer Interaction*, (ETRA), 2023. doi: https://doi.org/10.1145/3591132.

[35] Nora Horanyi, Linfang Zheng, Eunji Chong, Aleš Leonardis, and Hyung Jin Chang. Where are they looking in the 3d space? *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (CVPRW), 2023.

[36] Weihua Huang, Chew Lim Tan, and Jiuzhou Zhao. Generating ground truthed dataset of chart images: Automatic or semi-automatic? In *GREC*, 2007.

[37] Irene Anindaputri Iswanto and Bin Li. Visual object tracking based on mean-shift and particle-kalman filter. *Procedia computer science*, 116:587–595, 2017.

[38] Wei-Ta Chu, Chia-Hsiang Yu, and Hsin-Han Wang. Optimized comics-based storytelling for temporal image sequences. *IEEE Transactions on Multimedia*, 17(2): 201–215, 2014.

[39] J. Chen, G. Bai, S. Liang, and Z. Li. Automatic image cropping: A computational complexity study. pages 507–515, June 2016. doi: 10.1109/CVPR.2016.61.

[40] Yueying Kao, Ran He, and Kaiqi Huang. Automatic image cropping with aesthetic map and gradient energy map. pages 1982–1986. IEEE, 2017.

[41] Marcella Cornia, Stefano Pini, Lorenzo Baraldi, and Rita Cucchiara. Automatic image cropping and selection using saliency: An application to historical manuscripts. In *Italian Research Conference on Digital Libraries*, pages 169–179. Springer, 2018.

[42] G. Guo, H. Wang, C. Shen, Y. Yan, and H. M. Liao. Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression. *IEEE Transactions on Multimedia*, 20(8):2073–2085, Aug 2018. ISSN 1520-9210. doi: 10.1109/TMM.2018.2794262.

[43] Ning Shan, Daniel Stanley Tan, Melkamu S Denekew, Yung-Yao Chen, Wen-Huang Cheng, and Kai-Lung Hua. Photobomb defusal expert: Automatically remove distracting people from photos. *IEEE Transactions on Emerging Topics in Computational Intelligence*, (99):1–11, 2018.

[44] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-RL: Aesthetics Aware Reinforcement Learning for Image Cropping. pages 8193–8201, 2018.

[45] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. pages 226–234. IEEE, 2017.

[46] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. Learning to compose with professional photographs on the web. In *ACM International Conference on Multimedia*, pages 37–45. ACM, 2017.

[47] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, pages 1307–1315, 2018.

[48] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, pages 817–834. Springer, 2016.

[49] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial Transformer Networks. pages 2017–2025, 2015.

[50] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. 2015.

[51] Jaesik Park, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Modeling photo composition and its application to photo re-arrangement. pages 2741–2744. IEEE, 2012.

[52] Jingwei Huang, Huarong Chen, Bin Wang, and Stephen Lin. Automatic thumbnail generation based on visual representativeness and foreground recognizability. pages 253–261, 2015.

[53] Nehal Jaiswal and Yogesh K Meghrajani. Saliency based automatic image cropping using support vector machine classifier. In *ICIIECS*, pages 1–5. IEEE, 2015.

[54] Jiwon Choi and Changick Kim. Object-aware image thumbnailing using image classification and enhanced detection of roi. *Multimedia Tools and Applications*, 75 (23):16191–16207, 2016.

[55] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *ACM international conference on Multimedia*, pages 1105–1108. ACM, 2014.

[56] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. pages 3395–3402, 2015.

[57] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Imnage Processing*, 24(11):4185–4196, 2015.

[58] Wenguan Wang, Jianbing Shen, Ling Shao, and Fatih Porikli. Correspondence driven saliency transfer. 25(11):5025–5034, 2016.

[59] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. (1):20–33, 2018.

[60] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. Gaze-based interaction for semi-automatic photo cropping. In *SIGCHI conference on Human Factors in computing systems*, pages 771–780. ACM, 2006.

[61] Fred Stentiford. Attention based auto image cropping. In *Workshop on Computational Attention and Applications, ICVS*, volume 1, pages 253–261. Citeseer, 2007.

[62] Bongwon Suh, Haibin Ling, Benjamin B Bederson, and David W Jacobs. Automatic thumbnail cropping and its effectiveness. In *ACM symposium on User interface software and technology*, pages 95–104. ACM, 2003.

[63] Md Baharul Islam, Wong Lai-Kuan, and Wong Chee-Onn. A survey of aesthetics-driven image recomposition. *Multimedia Tools and Applications*, 76(7):9517–9542, 2017.

[64] Peng Lu, Hao Zhang, XuJun Peng, and Xiang Peng. Aesthetic guided deep regression network for image cropping. *Signal Processing: Image Communication*, 2019.

[65] Eunbin Hong, Junho Jeon, and Seungyong Lee. Cnn based repeated cropping for photo composition enhancement. 2017.

[66] Peng Wang, Zhe Lin, and Radomir Mech. Learning an aesthetic photo cropping cascade. pages 448–455. IEEE, 2015.

[67] Luming Zhang, Mingli Song, Yi Yang, Qi Zhao, Chen Zhao, and Nicu Sebe. Weakly supervised photo cropping. *IEEE Transactions on Multimedia*, 16(1):94–107, 2014.

[68] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomír Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: learning photo composition from dense view pairs. In *CVPR*, pages 5437–5446, 2018.

[69] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2601–2610, 2019.

[70] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. 2014.

[71] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation. pages 2407–2415, 2015.

[72] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. pages 4651–4659, 2016.

[73] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. pages 3156–3164, 2015.

[74] Dexin Zhao, Zhi Chang, and Shutao Guo. A multimodal fusion approach for image captioning. *Neurocomputing*, 2018.

[75] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[76] Xinyuan Qi, Zhiguo Cao, Yang Xiao, Jian Wang, and Chao Zhang. The accurate guidance for image caption generation. In *PRCV*, pages 15–26. Springer, 2018.

[77] Shuang Bai and Shan An. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, 2018.

[78] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. 2014.

[79] G. Song, B. B. Avants, and J. C. Gee. Multi-start method with prior learning for image registration. pages 1–8, Oct 2007. doi: 10.1109/ICCV.2007.4409159.

[80] Wojciech Kwedlo. A new random approach for initialization of the multiple restart em algorithm for gaussian model-based clustering. *Pattern Analysis and Applications*, 18(4):757–770, 2015.

[81] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[82] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A System for Large-Scale Machine Learning. In *USENIX Conference on Operating Systems Design and Implementation*, pages 265–283, 2016.

[83] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014.

[84] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[85] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[86] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[87] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. pages 618–626, 2017.

[88] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomáš Vojíř, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE transactions on PAMI*, 38 (11):2137–2155, 2016.

[89] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[90] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[91] G Hinton, N Srivastava, and K Swersky. Lecture 6d-a separate, adaptive learning rate for each connection. *Slides of Lecture Neural Networks for Machine Learning*, 2012.

[92] Michael JD Powell. A fast algorithm for nonlinearly constrained optimization calculations. In *Numerical analysis*, pages 144–157. Springer, 1978.

[93] L. A. Gatys, A. S. Ecker, and M. Bethge. Image Style Transfer Using Convolutional Neural Networks. 2016.

[94] J. Johnson, A. Alahi, and L. Fei-fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. pages 694–711, 2016.

[95] Adam S Williams and Francisco R Ortega. Understanding gesture and speech multimodal interactions for manipulation tasks in augmented reality using unconstrained elicitation. *Proceedings of the ACM on Human-Computer Interaction*, 4(ISS):1–21, 2020.

[96] Malcolm Slaney, Rahul Rajan, Andreas Stolcke, and Partha Parthasarathy. Gaze-enhanced speech recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3236–3240. IEEE, 2014.

[97] Paul P Maglio, Teenie Matlock, Christopher S Campbell, Shumin Zhai, and Barton A Smith. Gaze and speech in attentive user interfaces. In *International Conference on Multimodal Interfaces*, pages 1–7. Springer, 2000.

[98] Matheus Vieira Portela and David Rozado. Gaze enhanced speech recognition for truly hands-free and efficient text input during hci. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: the Future of Design*, pages 426–429, 2014.

[99] Ece Takmaz, Sandro Pezzelle, Lisa Beinborn, and Raquel Fernández. Generating image descriptions via sequential cross-modal alignment guided by human gaze. *arXiv preprint arXiv:2011.04592*, 2020.

[100] Ülkü Arslan Aydin, Sinan Kalkan, and Cengiz Acartürk. Speech driven gaze in a face-to-face interaction. *Frontiers in Neurorobotics*, page 8, 2021.

[101] Caroline PC Chanel, Raphaëlle N Roy, Frédéric Dehais, and Nicolas Drougard. Towards mixed-initiative human–robot interaction: Assessment of discriminative physiological and behavioral features for performance prediction. *Sensors*, 20(1):296, 2020.

[102] Liangzhe Yuan, Christopher Reardon, Garrett Warnell, and Giuseppe Loianno. Human gaze-driven spatial tasking of an autonomous mav. *IEEE Robotics and Automation Letters*, 4(2):1343–1350, 2019.

[103] Anam Ahmad Khan, Joshua Newn, James Bailey, and Eduardo Velloso. Integrating gaze and speech for enabling implicit interactions. In *CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2022.

[104] Kyle Reinholt, Darren Guinness, and Shaun K Kane. Eyedescribe: Combining eye gaze and speech to automatically create accessible touch screen artwork. In *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces*, pages 101–112, 2019.

[105] Hongzhi Zhu, Septimiu E Salcudean, and Robert N Rohling. A novel gaze-supported multimodal human–computer interaction for ultrasound machines. *International journal of computer assisted radiology and surgery*, 14(7):1107–1115, 2019.

[106] Chris Creed, Maite Frutos-Pascual, and Ian Williams. Multimodal gaze interaction for creative design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[107] Atsushi Fukayama, Takehiko Ohno, Naoki Mukawa, Minako Sawaki, and Norihiro Hagita. Messages embedded in gaze of interface agents—impression management with agent's gaze. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 41–48, 2002.

[108] Michael Nauge, Mohamed-Chaker Larabi, and Christine Fernandez-Maloigne. A statistical study of the correlation between interest points and gaze points. In *Human Vision and Electronic Imaging XVII*, volume 8291, pages 308–322. SPIE, 2012.

[109] Fatemeh Koochaki and Laleh Najafizadeh. Predicting intention through eye gaze patterns. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4. IEEE, 2018.

[110] Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(2):1–30, 2013.

[111] Kalin Stefanov, Jonas Beskow, and Giampiero Salvi. Vision-based active speaker detection in multiparty interaction. In *Grounding Language Understanding GLU2017 August 25, 2017, KTH Royal Institute of Technology, Stockholm, Sweden*, 2017.

[112] Ruth B Grossman, Erin Steinhart, Teresa Mitchell, and William McIlvane. "look who's talking!" gaze patterns for implicit and explicit audio-visual speech synchrony detection in children with high-functioning autism. *Autism Research*, 8(3):307–316, 2015.

[113] Onur Ferhat, Fernando Vilarino, and Francisco Javier Sánchez. A cheap portable eye-tracker solution for common setups. *Journal of eye movement research*, 7(3), 2014.

[114] Edoardo Ardizzone, Alessandro Bruno, and Giuseppe Mazzola. Saliency based image cropping. In Alfredo Petrosino, editor, *Image Analysis and Processing – ICIAP 2013*, pages 773–782, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-41181-6.

[115] Ziaur Rahman, Yi-Fei Pu, Muhammad Aamir, and Farhan Ullah. A framework for fast automatic image cropping based on deep saliency map detection and gaussian filter. *International Journal of Computers and Applications*, pages 1–11, 2018.

[116] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1105–1108, 2014.

[117] Wenguan Wang and Jianbing Shen. Deep cropping via attention box prediction and aesthetics assessment. In *Proceedings of the IEEE international conference on computer vision*, pages 2186–2194, 2017.

[118] Wenguan Wang, Jianbing Shen, and Haibin Ling. A deep network solution for attention and aesthetics aware photo cropping. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1531–1544, 2018.

[119] Simon Ho, Tom Foulsham, and Alan Kingstone. Speaking and listening with the eyes: Gaze signaling during dyadic interactions. *PloS one*, 10(8):e0136905, 2015.

[120] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014.

[121] Preethi Vaidyanathan, Emily Prud'hommeaux, Cecilia Ovesdotter Alm, Jeff B Pelz, and Anne Haake. Alignment of eye movements and spoken language for semantic image understanding. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 76–81, 2015.

[122] Nathan J Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews*, 24(6):581–604, 2000.

[123] Huiyu Duan, Xiongkuo Min, Yi Fang, Lei Fan, Xiaokang Yang, and Guangtao Zhai. Visual attention analysis and prediction on human faces for children with autism spectrum disorder. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(3s):1–23, 2019.

[124] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.

[125] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018.

[126] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7479–7489, 2019.

[127] Meir Cohen, Ilan Shimshoni, Ehud Rivlin, and Amit Adam. Detecting mutual awareness events. *IEEE transactions on pattern analysis and machine intelligence*, 34(12): 2327–2340, 2012.

[128] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3477–3485, 2019.

[129] Manuel Jesús Marin-Jimenez, Andrew Zisserman, Marcin Eichner, and Vittorio Ferrari. Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3):282–296, 2014.

[130] Michael Tomasello, Brian Hare, Hagen Lehmann, and Josep Call. Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *Journal of human evolution*, 52(3):314–320, 2007.

[131] Chris D Frith and Uta Frith. Social cognition in humans. *Current biology*, 17(16): R724–R732, 2007.

[132] Lisa J Stephenson, S Gareth Edwards, and Andrew P Bayliss. From gaze perception to social cognition: The shared-attention system. *Perspectives on Psychological Science*, 16(3):553–576, 2021.

[133] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6460–6468, 2018.

[134] Siavash Gorji and James J Clark. Attentional push: A deep convolutional network for augmenting image salience with shared attention modeling in social scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2510–2519, 2017.

[135] Adria Recasens*, Aditya Khosla*, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*, 2015. * indicates equal contribution.

[136] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 383–398, 2018.

[137] Omer Sumer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Attention flow: End-to-end joint attention estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3327–3336, 2020.

[138] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. October 2019.

[139] Adria Recasens Continente, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? *Neural Information Processing Systems Foundation*, 2015.

[140] Jian Guan, Liming Yin, Jianguo Sun, Shuhan Qi, Xuan Wang, and Qing Liao. Enhanced gaze following via object detection and human pose estimation. In *International Conference on Multimedia Modeling*, pages 502–513. Springer, 2020.

[141] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer, 2018.

[142] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1435–1443, 2017.

[143] Jun Bao, Buyu Liu, and Jun Yu. Escnet: Gaze target detection with the understanding of 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14126–14135, 2022.

[144] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[145] Felix A Gers and Jürgen Schmidhuber. Recurrent nets that time and count. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 3, pages 189–194. IEEE, 2000.

[146] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modelling. *arXiv preprint arXiv:1412.3555*, 2014.

[147] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European conference on computer vision*, pages 47–54. Springer, 2016.

[148] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[149] Kaiyu Yue, Ming Sun, Yuchen Yuan, Feng Zhou, Errui Ding, and Fuxin Xu. Compact generalized non-local network. *Advances in Neural Information Processing Systems*, 31:6510–6519, 2018.

[150] S. Zhuoran, Z. Mingyuan, Z. Haiyu, Y. Shuai, and L. Hongsheng. Efficient attention: Attention with linear complexities. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3530–3538, Los Alamitos, CA, USA, jan 2021. IEEE Computer Society. doi: 10.1109/WACV48630.2021.00357. URL https://doi.ieeecomputersociety.org/10.1109/WACV48630.2021.00357.

[151] Woncheol Jang, Joon-Ho Shin, Mingyu Kim, and Kwanguk Kenny Kim. Human field of regard, field of view, and attention bias. *Computer methods and programs in biomedicine*, 135:115–123, 2016.

[152] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921, 2019.

[153] Jae-Han Lee and Chang-Su Kim. Monocular depth estimation using relative depth maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2019.

[154] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[155] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. End-to-end human-gaze-target detection with transformers. *arXiv preprint arXiv:2203.10433*, 2022.

[156] Moreno I Coco and Frank Keller. Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive science*, 36(7):1204–1223, 2012.

[157] Lila R Gleitman, David January, Rebecca Nappa, and John C Trueswell. On the give and take between event apprehension and utterance formulation. *Journal of memory and language*, 57(4):544–569, 2007.

[158] Alfred L Yarbus. Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer, 1967.

[159] Michaël Fonder, Damien Ernst, and Marc Van Droogenbroeck. Parallax inference for robust temporal monocular depth estimation in unstructured environments. *Sensors*, 22(23):9374, 2022.

[160] Seokju Lee, Sunghoon Im, Stephen Lin, and In So Kweon. Learning monocular depth in dynamic scenes via instance-aware projection consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1863–1872, 2021.

Appendix A

Caption and Aesthetic-Guided Image Cropping

A.1 Additional qualitative highlights

We provide more qualitative results that were omitted due to spatial constraints. We report the same variants reported in Section 2.4.3 where we compare our results to the eight baseline methods: A2-RL [44], VPN [68], Anchor [86], GradCam [87], GradCam+A2-RL, GradCam+VPN, GradCAM+Anchor, and MAttNet [47]. Similarly to the figures in Section 2.4.3, we show multiple ground truth bounding boxes on the original images generated by eight users based on the captions.

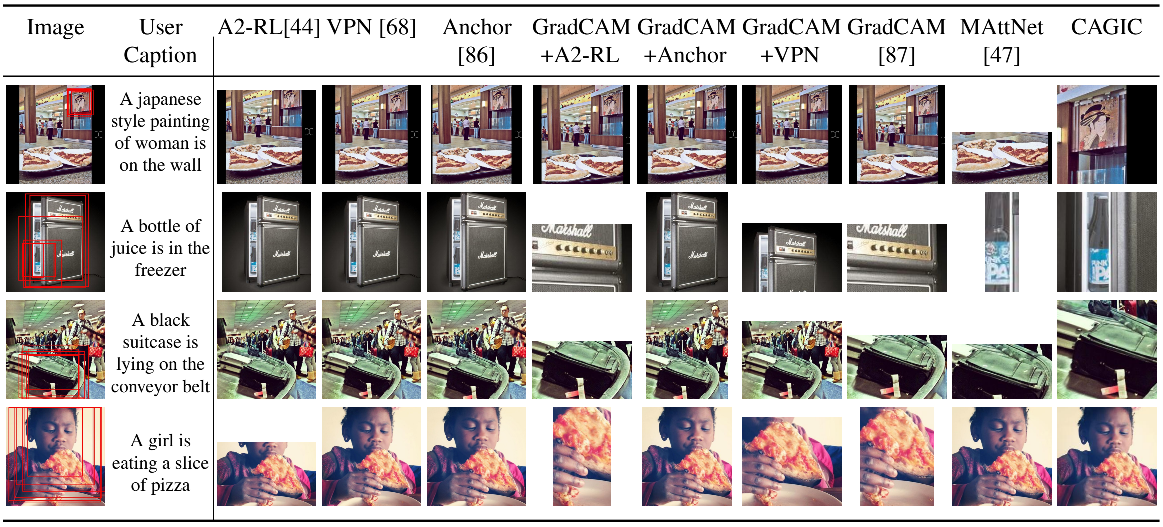


Image
User Caption
A2-RL[44]
VPN [68]
Anchor [86]
GradCAM +A2-RL
GradCAM +Anchor
GradCAM +VPN
GradCAM [87]
MAttNet [47]
CAGIC
A japanese style painting of woman is on the wall
A bottle of juice is in the freezer
A black suitcase is lying on the conveyor belt
A girl is eating a slice of pizza

Figure A.1: Qualitative results on our dataset.

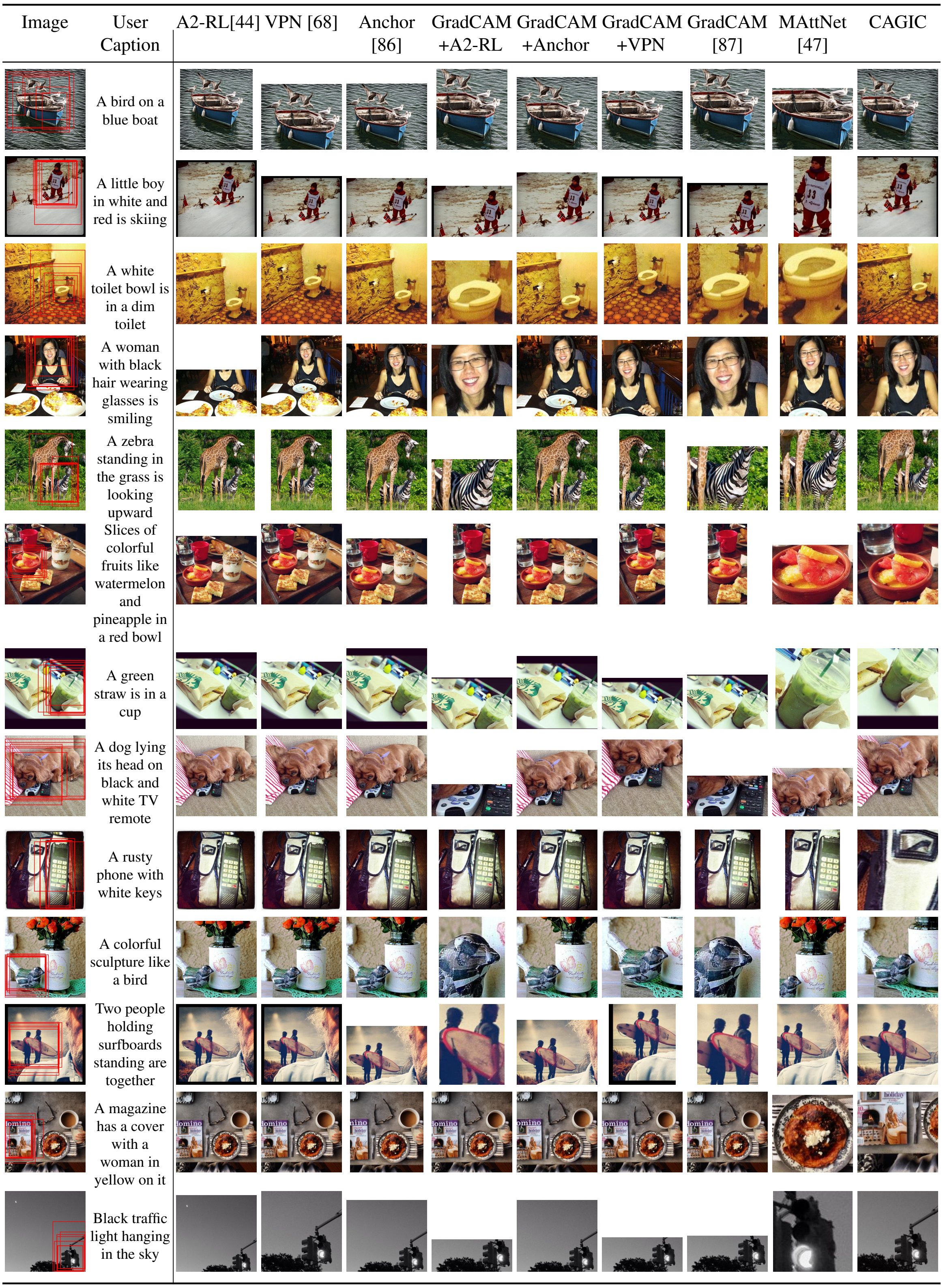


Image
User Caption
A2-RL[44]
VPN [68]
Anchor [86]
GradCAM +A2-RL
GradCAM +Anchor
GradCAM +VPN
GradCAM [87]
MAttNet [47]
CAGIC
A bird on a blue boat
A little boy in white and red is skiing
A white toilet bowl is in a dim toilet
A woman with black hair wearing glasses is smiling
A zebra standing in the grass is looking upward
Slices of colorful fruits like watermelon and pineapple in a red bowl
A green straw is in a cup
A dog lying its head on black and white TV remote
A rusty phone with white keys
A colorful sculpture like a bird
Two people holding surfboards standing are together
A magazine has a cover with a woman in yellow on it
Black traffic light hanging in the sky

Figure A.2: Qualitative results on our dataset (continued).

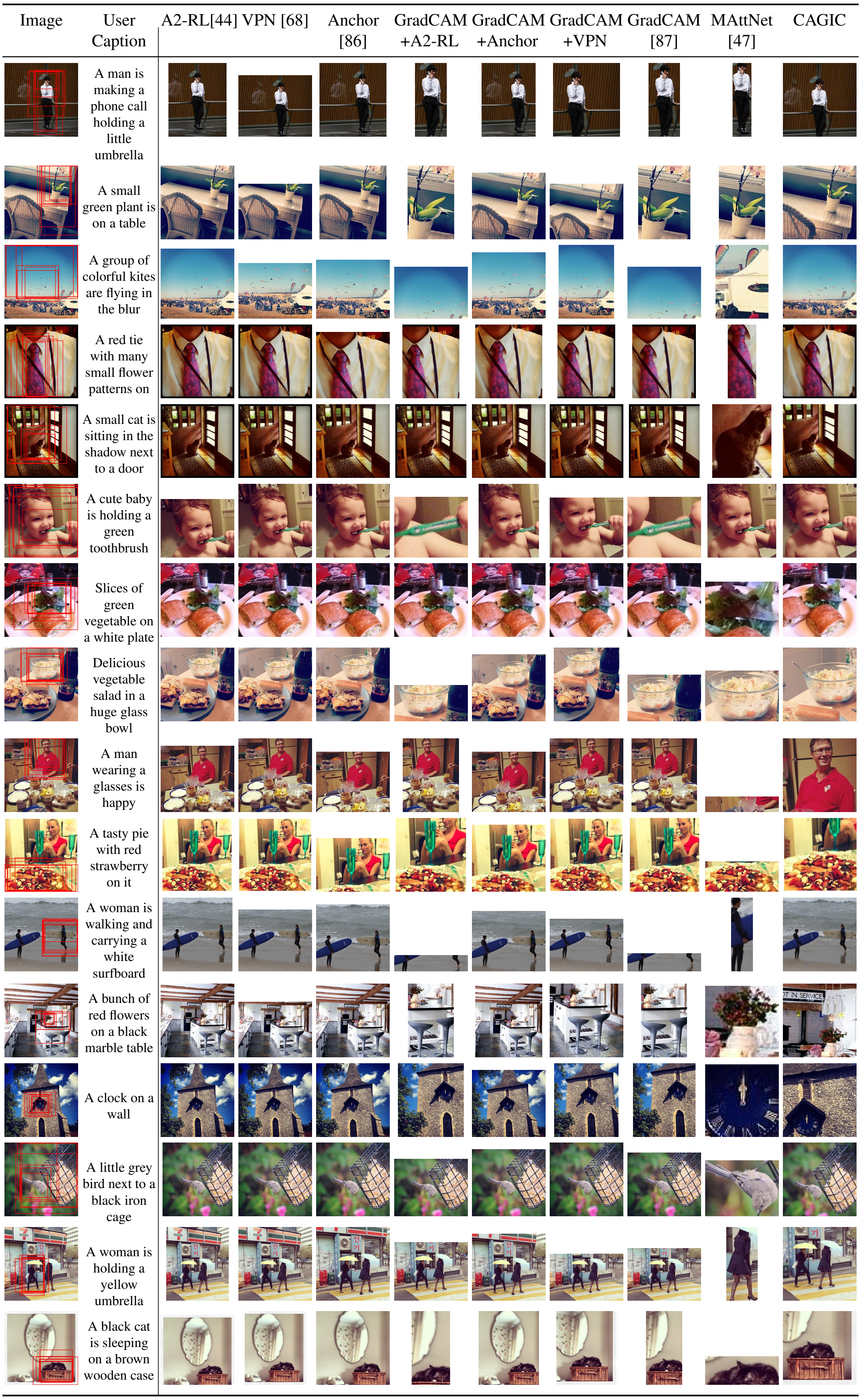


Image
User Caption
A2-RL[44]
VPN [68]
Anchor [86]
GradCAM +A2-RL
GradCAM +Anchor
GradCAM +VPN
GradCAM [87]
MAttNet [47]
CAGIC
A man is making a phone call holding a little umbrella
A small green plant is on a table
A group of colorful kites are flying in the blur
A red tie with many small flower patterns on
A small cat is sitting in the shadow next to a door
A cute baby is holding a green toothbrush
Slices of green vegetable on a white plate
Delicious vegetable salad in a huge glass bowl
A man wearing a glasses is happy
A tasty pie with red strawberry on it
A woman is walking and carrying a white surfboard
A bunch of red flowers on a black marble table
A clock on a wall
A little grey bird next to a black iron cage
A woman is holding a yellow umbrella
A black cat is sleeping on a brown wooden case

Figure A.3: Qualitative results on our dataset (continued).

A.2 Dataset documentation

The proposed, description-based image cropping dataset introduced in Section 2.4.1 is available online [1]. The dataset is structured as follows:

- **Images:** A folder containing 100 images (001-100.jpg).
- **Ground truth annotations:**
  - **captions.xlsx:** Excel file containing the collected user captions corresponding to the images.
  - **caption_gt.xlsx:** Ground truth annotation information of the proposed dataset organized as [img name][MS-COCO val2014 Ids][Image caption][7 bbox annotations], where we first list the image filename and the corresponding MS-COCO dataset ID of the randomly selected image. Following, we provide the user caption and the 7 ground truth image crop bounding box annotations collected from human annotators as $[x_{min}, y_{min}, x_{max}, y_{max}]$.
  - **annotation.csv:** CSV equivalent of the caption_gt.xlsx file.

Figure A.4 shows an example image and its annotation below:
[087.jpg][469936][a magazine has a cover with a woman in yellow on it][[2,222,301,301] [5,226,206,297] [3,232,216,297] [3,232,216,297] [1,199,224,345] [2,174,239,393] [3,227,206,297] [38,259,195,249]]

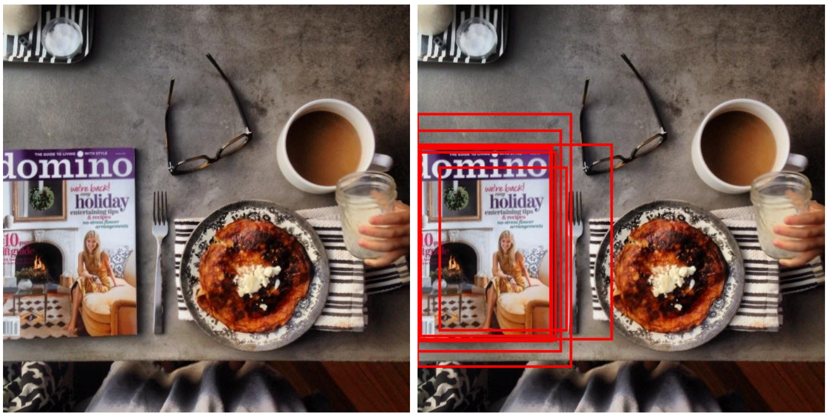


Figure A.4: **Dataset example visualisation.** Image #87 of the proposed dataset and the visualisation of its ground truth bounding box annotations as red.

[1] https://horanyinora.github.io/publication/Horanyi_PR_data.zip

Appendix B

Gaze and Description-Based Image Cropping

B.1 Performance evaluation of the available eye tracking devices

B.1.1 Aim and objective

In Section 3.4.1 we introduced a novel dataset for the G-DAIC task. This dataset is the extension of the CAGIC dataset introduced in Section 2.4.1 proposed in [33]. The experimental setting of the gaze data collection is discussed in detail in Section 3.4.1. This section summarises our preliminary experiments before gaze data collection. We aimed to study and compare the behaviour of the eye tracker devices available in our laboratory and identify the most suitable one for our data collection task.

B.1.2 Existing eye trackers

Many devices provided by various companies in various price ranges are available for eye-tracking purposes. Low-end eye trackers, like The Eye Tribe eye tracker, are not recommended for advanced research but are useful to better understand the technology and data collection. Middle-end eye trackers such as Mirametrix S2 are adequate for certain research questions and provide good value for those with a limited budget. Finally, the Tobii Pro Fusion and Pro Glasses 2 eye trackers are high-end. High-end trackers are normally used by organisations with more advanced research objectives that rely heavily on high accuracy, precision, stability, and usability. See Table B.1. for more details.

We planned to use monitor-mounted or free-standing eye-tracking bars for our screen-based study. Currently, our laboratory owns an Eye Tribe and a Tobii Pro Fusion screen-based remote eye tracking bars; therefore, in our experiments, we compared the performance of these trackers.

Table B.1: **Comparison of eye trackers accuracy, frequency and operating distance from the different price ranges.**

	Price range	Accuracy	Frequency	Operating distance
The Eye Tribe Tracker	\$	0.5 – 1°	30 Hz	45 – 75 cm
Mirametrix S2 Tracker	\$\$	0.5 – 1°	60 Hz	65 cm
Tobii Pro Fusion	\$\$\$	Approx. 0.4°	60 & 120 Hz	50 – 80 cm
Tobii Pro Glasses 2	\$\$\$	Requested	50 & 100 Hz	Not screen based

B.1.3 Experiments

Our experiments compare the eye-tracking devices in terms of accuracy and precision. Based on the specifications of the eye trackers, our expectation was that the cheaper eye tracker would perform worse than the more expensive screen-based tracker. We anticipated that the cheap eye tracker with lower frequency would fail to accurately track the faster eye movements, it would take longer to recover the tracking after blinking, and occasionally would output incomplete trajectories. Meanwhile, we expected the high-resolution tracker to provide almost complete and precise eye movement trajectories. To compare the performance of the available devices, we had to ensure simultaneous tracking in our experimental setup. The most challenging part of using trackers simultaneously is the synchronisation problem.

Experimental setting

We used the Tobii Pro Lab software (no Linux operating system support) to record and process the data collected by the Tobii eye tracker device. This software was specifically developed to provide a simple but efficient user interface from experimental design and recording to analysis. We used this software to calibrate the Tobii Pro Fusion device through a VMware Windows virtual machine on a Lambda TensorBook (Linux operating system) and the Psychology toolbox for the Eye Tribe device. As the toolbox does not support virtual machines, we had to set up a bridge between the host Ubuntu machine and the guest to stream data from the tracker devices. Running the two tracker devices simultaneously required a graphics card (NVIDIA GeForce RTX 2080), and the power settings of the TensorBook had to be adjusted.

The screen-based eye tracker was stuck to the bottom of the screen, while for the Eye Tribe tracker, we used a tripod and carefully positioned it such that the trackers did not block each other. In Figure B.1, we show the experimental setting. On the screen, we show the simultaneous output of both trackers using the Eye Tribe SDK and the Tobii Pro Lab software for visualisation. Both trackers were calibrated in this position before every measurement. The calibration results of the measurements are shown in Figure B.2.

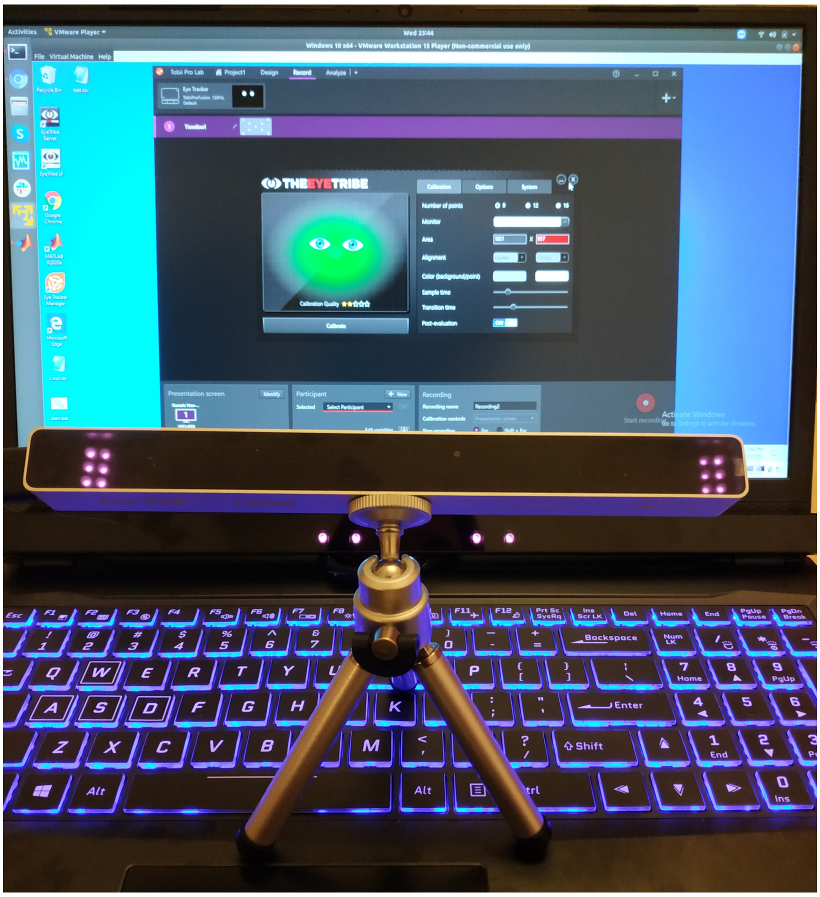


Figure B.1: **Experimental setting for simultaneous eye tracking.** In the front an Eye Tribe eye tracker is standing on a tripod, behind the Tobii eye tracker bar is attached to the bottom of the screen.

During the measurements, we asked the subjects to fixate on the middle of a predefined marker circle for as long as it was visible. To perform this experiment, we generated a video sequence with GT target locations for the recordings. The generated video was similar to the calibration, where the user had to follow a floating marker on the screen and fixate on different locations when the marker stopped moving. In our experiment, we distributed nine

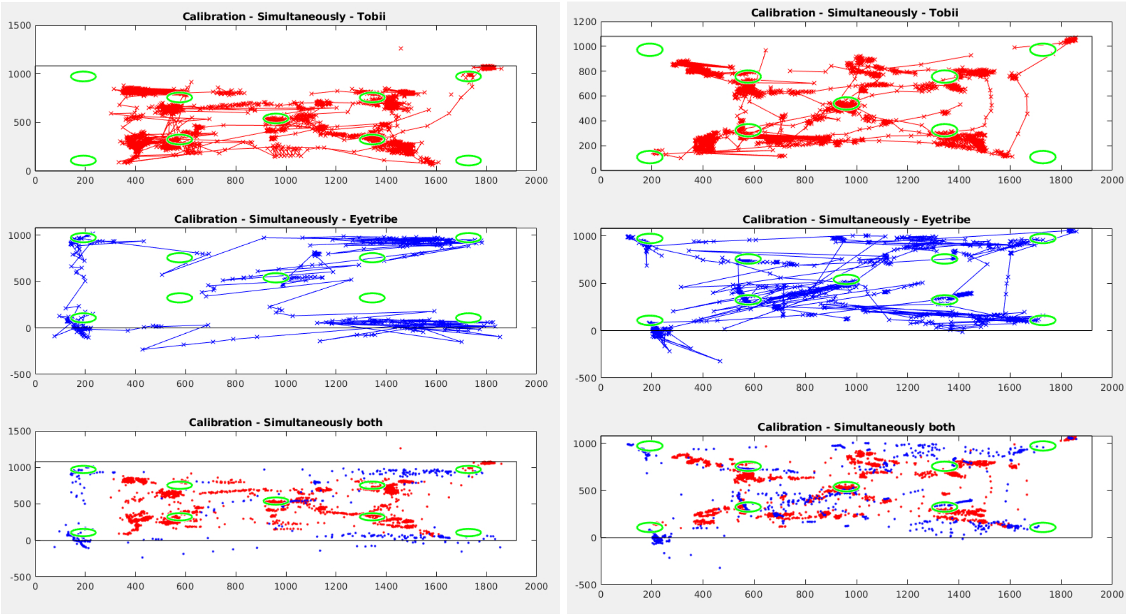


Figure B.2: **Illustration of the eye-tracker calibration results.**

marker points throughout the scene. For the exact marker locations, refer to Figure B.3. We ran several experiments to confirm that the experimental setting was suitable for the recording and that the calibration was robust and reliable.

B.1.4 Results

In Figure B.4. we present the qualitative results of our experiment. At the top left, we show the fixation points collected from the Eye Tribe eye tracker device and on the right, the points collected by the Tobii tracker. Note that the points were collected simultaneously during our experiment and that we only visualise them separately for better visibility. At the bottom of Figure B.4, we visualise the fixation points of both devices. We evaluated the performance of the eye trackers by counting the number of points within the ground truth marker locations and measuring the distance of the fixation points from the centre of the marker.

During the measurement, the ground truth markers appeared on the scene sequentially. Each marker was displayed on the screen for 3.5 seconds; after this, the marker disappeared immediately, and 0.25 seconds later, a new one appeared on the screen. Therefore, at any given time, a maximum of one marker was visible on the screen. As we do not have information on when exactly the subject begins the fixation on the target points, we divided the recording into nine sections and the recorded gaze points accordingly to determine which points belong to a specific marker. Note that the Tobii eye tracker device has a higher frequency (60-120 Hz) compared to the Eye Tribe cheaper tracker (30 Hz), which can be seen

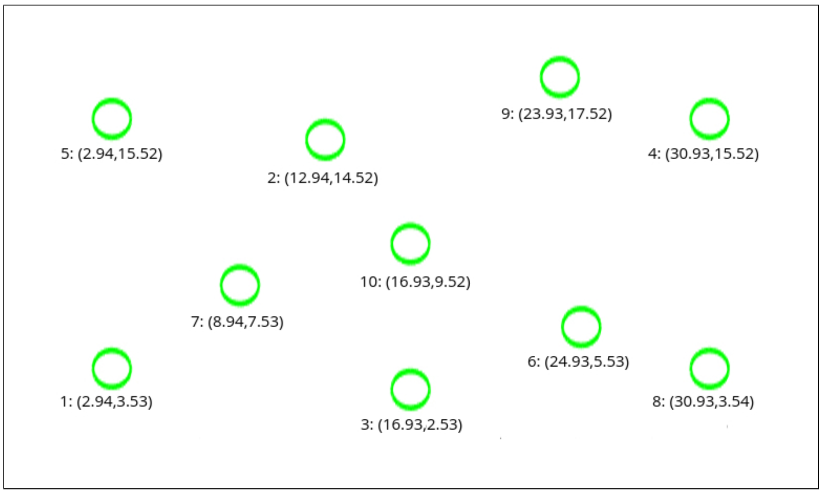


Figure B.3: **Ground truth locations (x,y coordinates) of the nine fixation points during our experiment.** The floating dot is moving between the predefined fixation points in random order.

in Figure B.4 and that this way of splitting the data puts the Tobii eye tracker in disadvantage due to its recording frequency.

The quantitative results are shown in Table B.2. We can see in Figure B.4 that the Tobii tracker nearly did not have any fixation points within Marker 7, which is reflected in the table as well.

Table B.2: **Error calculated *w.r.t.*nine ground truth marker positions for both trackers.** Count (%) indicates the percentage of gaze points inside the marker disk.

GT marker	Tobii Pro Fusion							Eye Tribe						
	mean			std			count	mean			std			count
	d_x	d_y	euc	d_x	d_y	euc	%	d_x	d_y	euc	d_x	d_y	euc	%
1	109.65	24.65	125.03	200.88	28.63	212.91	73.16	0.13	0.35	41.15	217.82	33.68	128.57	46.49
2	107.44	176.92	232.04	189.16	209.41	293.86	5.33	134.69	177.65	103.22	227.93	265.79	236.18	42.20
3	60.59	115.66	145.82	93.90	202.20	233.75	74	79.33	166.08	154.63	104.05	217.63	261.27	25.89
4	128.11	145.35	215.70	266.88	252.89	389.43	78.22	175.81	172.47	206.97	279.22	298.19	322.02	64.60
5	318.38	72.45	384.94	597.54	153.73	642.62	73.39	384.48	65.95	276.66	638.44	148.30	447.22	69.03
6	211.47	128.49	284.67	405.08	164.02	457.29	31.11	290.34	119.78	324.55	486.32	194.27	508.43	53.10
7	196.33	95.70	267.81	343.93	51.21	348.84	0.67	246.80	71.25	307.14	362.22	62.84	457.95	52.68
8	215.82	57.26	252.94	409.59	76.07	438.96	76.89	269.62	66.74	307.07	476.64	82.68	462.53	65.49
9	85.31	183.65	206.94	119.95	264.43	287.21	20.26	104.54	177.34	258.38	138.41	302.85	426.51	60
Total	178.13	114.29	235.10	308.00	163.15	367.21	**48.13**	223.54	129.73	219.97	339.98	185.37	361.19	**53.34**

B.1.5 Conclusion

The quantitative results show that despite the trackers' factory specifications (See Table B.1.) suggest that the Tobii Pro Fusion tracker's accuracy is higher than the cheaper Eye Tribe tracker's, in our experiments, we found that the cheaper tracker had lower average

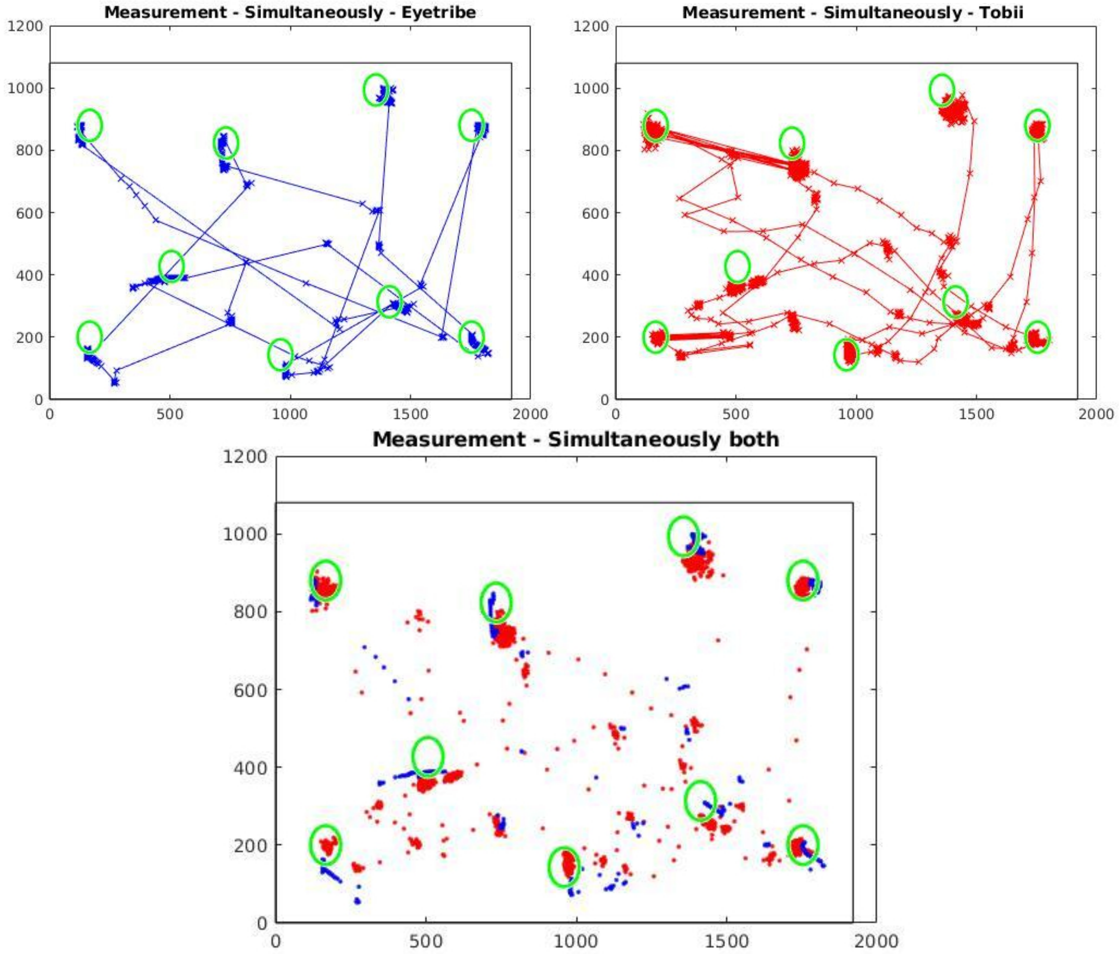


Figure B.4: **Qualitative comparison of the eye tracker performances.** Visualisation of the simultaneously recorded fixation points of the two trackers separately and together. The rectangles within the plots represent the screen and the green circles are the ground truth predefined fixation locations.

Euclidean distance error as well as higher in marker fixation percentage. Note that the recording frequencies of the devices are different. Therefore, we recorded more points with the Tobii eye tracker during the dynamic eye movements, which explains why we have more out-of-marker gaze points. Furthermore, despite the quantitative results, the points recorded by the Tobii tracker seem more compactly clustered. Finally, we decided to use the Tobii Pro Fusion eye tracker device for the data collection as our task required higher frequency.

B.2 Dataset documentation

The extended description-based image cropping dataset, including user fixation information introduced in Section 3.4.1, is available online[1]. The dataset is structured as follows:

- **CAGIC_annotations_extended.xlsx:** We provide the image captions, their bounding box annotations and the length of each word of the captions in this excel file. The data is organised as: [img name][MS-COCO val2014 Ids][Image caption][7 bbox annotations][length (sec) of every word in the caption]

- **data/:** Folder containing the image files, the annotations and the collected gaze data for each participant.
  - **data/[001..100]/:** we collected the original image and visualised the ground truth bounding box annotations of the dataset and the generated caption recordings.
  - **data/[001...100]/gaze/:** we organised the collected gaze data under fixation/[u01...u14] free/[u01...u14] and stimuli/[u01...u14]. Each folder contains 14 CSV files generated by the Tobii Pro Lab eye-tracking software.

The directory tree of the dataset is as follows:

```
CAGIC extended dataset
├─ CAGIC_annotations_extended.xlsx
└─ data
   └─ [001...100/] image folder
      ├─ RGB image
      ├─ Caption mp3 recording
      ├─ Ground truth annotation visualisation image
      └─ gaze/
         ├─ fixation/
         │  └─ [u01..014] user fixation CSV file
         ├─ free/
         │  └─ [u01..014] user free-viewing CSV file
         └─ stimuli/
            └─ [u01..014] user stimuli CSV file
```

Figure A.4 shows an example image and its annotation below:
[086.jpg][241934][a pancake with black and dark chocolate on it][[400,172,212,212] [413,183,184,193] [403,171,209,216] [386,156,226,250] [382,179,227,217] [414,184,196,189] [385,173,227,212]][0 0.204 0.919 1.266 1.752 2.057 2.406 3.131 3.781 4.0228333333]

[1] https://horanyinora.github.io/publication/Horanyi_ETRA_data.zip

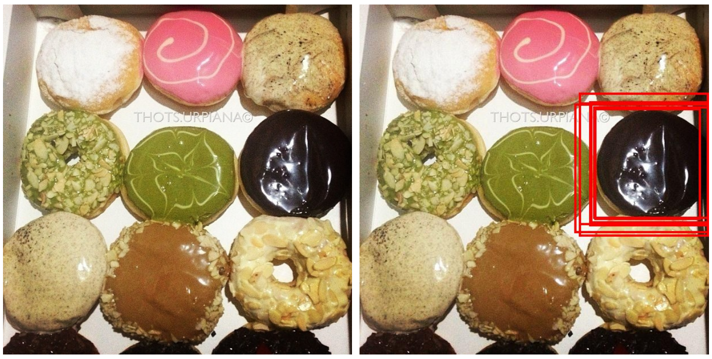


Figure B.5: **Image #86 of the proposed dataset and the visualisation of its ground truth bounding box annotations as red.**

The gaze data (Free-viewing/Stimuli/Fixation) can be visualised as a heatmap as a collection of gaze points or overlapped over the original image. In Figure B.6, we show the gaze data visualisation results of Image #86.

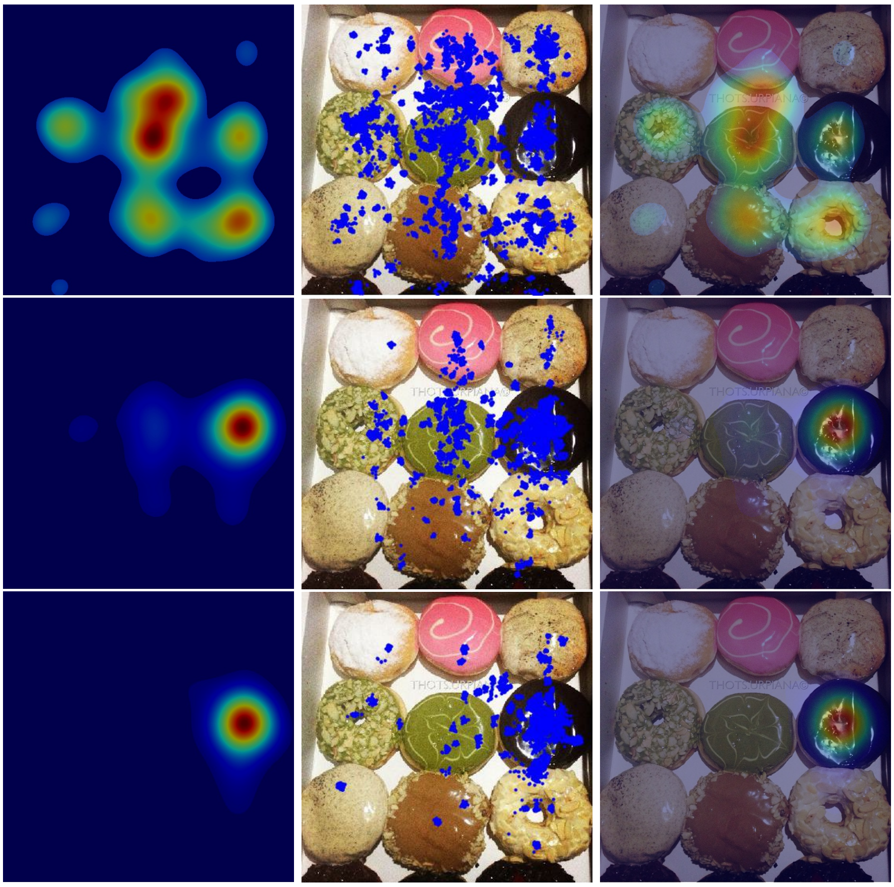


Figure B.6: **Collected gaze data visualisation example on Image #86 of the extended dataset.** First free-viewing, second stimuli and last, the fixation data visualised in the form of a heatmap, gaze points, and overlapped with the original image.