IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 April 2023, accepted 30 May 2023, date of publication 7 June 2023, date of current version 14 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3283495

TOPICAL REVIEW

From a Visual Scene to a Virtual Representation: A Cross-Domain Review

AMÉRICO PEREIRA[1,2], PEDRO CARVALHO[1,3], (Senior Member, IEEE), NUNO PEREIRA[1,3], PAULA VIANA[1,3], (Senior Member, IEEE), AND LUÍS CÔRTE-REAL[1,2], (Member, IEEE)

[1]Centre for Telecommunications and Multimedia, Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), 4200-465 Porto, Portugal
[2]Faculty of Engineering, University of Porto, 4099-002 Porto, Portugal
[3]ISEP, Polytechnic of Porto, 4249-015 Porto, Portugal

Corresponding author: Américo Pereira (americo.j.pereira@inesctec.pt)

The work was supported by the European Union's Horizon Europe Research and Innovation Program (Project Converge—Telecommunications and Computer Vision Convergence Tools for Research Infrastructures) under Grant 101094831. Américo Pereira was supported by Fundação para a Ciência e a Tecnologia (FCT), under the Grant SFRH/BD/14600/2019.

ABSTRACT The widespread use of smartphones and other low-cost equipment as recording devices, the massive growth in bandwidth, and the ever-growing demand for new applications with enhanced capabilities, made visual data a must in several scenarios, including surveillance, sports, retail, entertainment, and intelligent vehicles. Despite significant advances in analyzing and extracting data from images and video, there is a lack of solutions able to analyze and semantically describe the information in the visual scene so that it can be efficiently used and repurposed. Scientific contributions have focused on individual aspects or addressing specific problems and application areas, and no cross-domain solution is available to implement a complete system that enables information passing between cross-cutting algorithms. This paper analyses the problem from an end-to-end perspective, i.e., from the visual scene analysis to the representation of information in a virtual environment, including how the extracted data can be described and stored. A simple processing pipeline is introduced to set up a structure for discussing challenges and opportunities in different steps of the entire process, allowing to identify current gaps in the literature. The work reviews various technologies specifically from the perspective of their applicability to an end-to-end pipeline for scene analysis and synthesis, along with an extensive analysis of datasets for relevant tasks.

INDEX TERMS Computer vision, datasets, scene analysis, scene reconstruction, visual scene understanding.

I. INTRODUCTION

Starting from a real-world scene and extracting its structure to obtain a virtual scene that accurately reproduces it is a long-standing computer vision challenge. This process, of scene understanding for 3D synthesis, can enable many exciting applications in sports, entertainment, and telepresence, to name a few: viewers of sports events can follow their favorite players by placing a virtual camera in the reconstructed scene; movie creators can obtain new camera views in post-production, or remote participants can visit a real-world location (for training purposes, for example) in virtual reality or even interact with other users in the location wearing augmented reality devices.

The pervasive nature of image and video-capture devices in our lives, coupled with increasing processing capabilities and algorithmic advances in machine learning, is fueling a renewed interest in performing automated scene understanding for 3D synthesis, with a vast amount of literature dedicated to its different sub-problems, such as analyzing and extracting data, and 3D content synthesis. To analyze and extract data from a scene, current approaches aim to efficiently detect visual structures and understand the inherent scene structure and the context of the information

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

extracted [1]. This might include extracting fine-grained geometry details as well as contextual information such as relative locations and object categories. For 3D content synthesis, we find methodologies often tailored for specific tasks. For instance, human parametric models can be used to model and generate human-like avatars that can be adapted to multiple situations while preserving human characteristics and anatomic coherence [2]. In addition, recent research on neural radiance fields has also enabled detailed and rich representations of natural scenes through multiple viewpoints of a scene [3], [4].

These different advances highlight the complexity of creating solutions for scene understanding and synthesis, and the extensive literature already available. Although there are advances in many research areas, which culminate in new forms of applications that leverage both visual analysis and scene representation/synthesis to create coherent virtual environments such as video games or even virtual meeting rooms, there is still a lack of unified pipelines that incorporate scene analysis with the interpretation of semantic information of the scene for a more detailed and accurate scene representation. As we move towards photo-realistic 3D scene understanding, semantic information allows expressing meaning and relationship of and between entities on the scene, thus improving perception by providing a more detailed analysis of the underlying scene. We argue that obtaining reliable and meaningful semantic information from the scene becomes essential, and representing this information to allow efficient storage, access, and interpretation is fundamental for handling the data extracted from a visual scene.

This work discusses visual scene understanding as a means to transition between the scene analysis and a posterior virtual synthesis, exploring the steps required for such transition while presenting an exploratory survey of recent works and their applicability to such a scenario. We structure the discussion around a cross-domain pipeline intended to analyze and extract human activity and interactions, with the ultimate goal of obtaining the underlying scene description that enables a richer posterior 3D synthesis of the scene. The pipeline (depicted in Figure 1) while high-level, introduces an end-to-end approach that enables information passing between cross-cutting algorithms. Advances in independent areas, for example, human body tracking, can be leveraged to correctly establish coherent spatio-temporal semantic representations in the scene, enhancing the representation of the underlying observed data. This cross-domain perspective potentiates the use of information from scene analysis (e.g., human position, facial data, gait), in the form of a compact and flexible description that can be stored or sent across the network, and enables the posterior synthesis and recreation for visualization under different points of view or with varying levels of detail. To this end, it is critical that the scene description is flexible to adequately capture the required information.

The pipeline depicted in Figure 1 shows the three main domains that guide the presented literature review: (1) Scene Analysis; (2) Scene Description; (3) Scene Synthesis. The first area consists of technologies responsible for obtaining semantic information from a visual scene for posterior visualization of the data. The second area is responsible for generating a scene description based on the outcome of the visual analysis. This description enables the structured organization of the semantic information extracted from the scene analysis and allows storing or conveying the data between algorithms. The final area is responsible for generating synthetic representations of the underlying scene based on the information extracted and presented by the second layer. Following this pipeline enables the usage and storage of higher-level information used to recreate the underlying observed data with varying levels of detail, and can be applied in many different application areas such as: semantic compression; surveillance and unauthorized access; sports analysis; data augmentation and synthetic data generation.

The contributions of the work are guided by the high level pipeline and are three-fold:

- a cross-domain survey of research required to establish a transition from scene analysis to a posterior virtual synthesis, with emphasis on recent research, applicable to an end-to-end system;
- a comprehensive exploration of applicable datasets, with a focus on providing an easy-to-access entry-point to obtain the data;
- a discussion of advantages and research opportunities emerging from structuring the problem around a cross-domain approach, highlighting the need for adequate data representations for interconnection.

This article is structured as follows: Section II explores areas related to visual scene analysis, based on processing low-level data, such as images or videos, to obtain semantic information from the scene; Section III explores methodologies for organizing semantic information and how high-level information extracted from the visual scene can be structured; Section IV targets methods and algorithms designed for scene synthesis and how to transition between visual data to virtual representations. Then Section V presents an exploratory overview of existing and relevant datasets encompassing this cross-domain analysis while also providing an easy-to-access resource to obtain more details of the given datasets and how to obtain the data. Finally, Section VI discusses the advantages that a cross-domain approach integrating the explored areas could potentiate, as well as emerging research opportunities due to the complex nature of such integration.

II. SCENE ANALYSIS AND DATA EXTRACTION

To understand a visual scene through an automatic or semi-automatic process, it is necessary first to analyze and interpret the visual information. Due to its importance, this has been an active research topic with many proposed works. Furthermore, as a wide variety of applications rely on visual data, it is also natural for multiple technologies to be derived and studied. In our exploration of different techniques for scene analysis, we take a particular interest in methodologies and algorithms that focus on humans and their interactions,

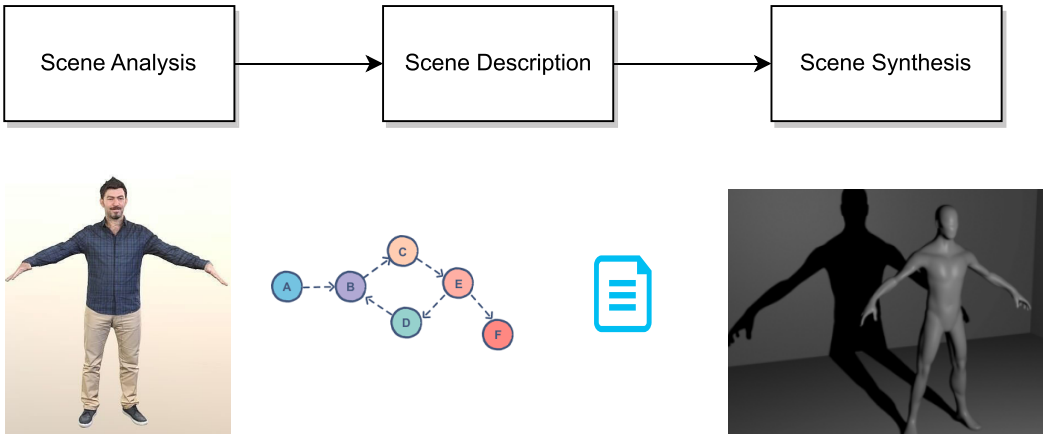

FIGURE 1. High level architecture of the proposed visual-virtual translation pipeline.

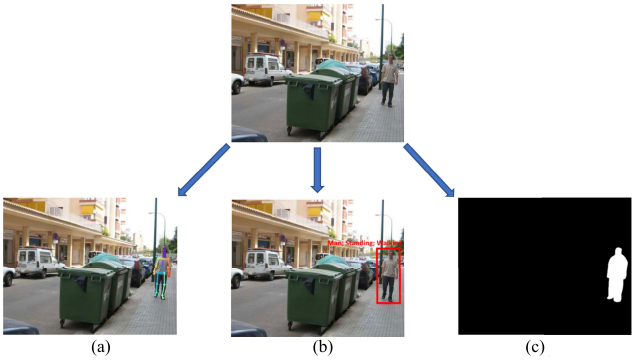

FIGURE 2. Example of different outputs of various algorithms applied to analyze a scene. Figure a) depicts the application of pose estimation; figure b) shows the detection of a person and some semantic knowledge obtained from the detection; figure c) illustrates the foreground of the scene.

and overview both classical and modern methods to illustrate the evolution that different areas within scene analysis have experienced. Figure 2 illustrates some of the modalities of scene analysis.

A. OBJECT AND PEOPLE DETECTION

Object detection and, in particular, people detection aims to determine if an object is present on an image and find all instances in the image. Detectors typically provide a bounding box around the detected object and the inherent problem of training an object detector is mostly seen as a supervised learning problem, which takes leverage of the vast amounts of labeled datasets available (see section V for an overview of multiple datasets). Traditional object detectors relied on the detection of specific hand-crafted image features like the Viola and Jones [5] or the HoG (Histogram of oriented Gradients) [6] algorithms. In 2012, AlexNet [7], marked the re-introduction of convolutional neural networks (CNN) and deep learning in computer vision, paving the way to more precise and sophisticated methodologies with increased accuracy.

Recent object detection approaches fall into three different types: single-stage; two-stage; and transformer-based. Single stage detectors perform the joint detection and classification of objects by using sets of anchor boxes of multiple scales and aspect ratios to detect all object instances. Two stage detectors have a specific module for generating candidate regions which are forwarded to a classifier that localizes and classifies the objects based on the candidates. Lastly, transformer-based detectors bring into the vision domain concepts used formerly in NLP (Natural Language Processing) and work by capturing the relationships between different regions of an image.

One of the most widely used single-stage detector is YOLO [8], which involves a single neural network that is trained end-to-end to predict object bounding boxes. A major drawback is the difficulty in detecting small or clustered objects, and it also lower recall when compared with the Faster-RCNN [9]. Single Shot MultiBox Detector (SSD) is also a single-stage detector that was built upon the VGG-16 network by appending progressively smaller convolutions to the end of the model and it also has difficulties to correctly detect small objects and requires a large training set. YOLOR (You Only Learn One Representation) [10] consists of a network that imitates how the human brain learns knowledge from normal learning and subconscious learning. Results show that this proposal is able to achieve accuracy comparable to state-of-the-art methods with fast inference speed. YOLOv7 [11] was proposed for real-time object detection. It derived from YOLOv4 [12], Scaled YOLOv4 [13] and YOLOR [10], with results surpassing the state-of-the-art detectors in both speed and accuracy.

Faster R-CNN [9] and Mask-RCNN [14] have several similarities and are two of the most widely used two-stage object detectors. Further improved methods have also been proposed for even better accuracy; Cascade R-CNN [15] uses a cascade of different specialized regressors and G-RCNN [16] includes the concept of spatio-temporal granulation within a deep convolutional neural network.

The introduction of Transformers [17] also impacted image and video processing; for example, Vision Transformer (ViT) [18] treated images as a sequence of patches to create an image classifier. Transformers were applied to

object detection with proposals like ViT-FRCNN [19], DeTR [20], Swin transformer [21]. More recently, the usage of non-hierarchical ViT has been explored [22]; Zhang et al. [23] pre-trained transformer encoder-decoders and Mask-DINO [24] used an extension of DINO [25] (DETR-like model), adding a mask prediction branch. For a deeper study of this topic the reader is referred to [26].[1]

Object detection is a cross-cutting concern in many research areas and is often a critical step in image processing pipelines. As a result, many algorithms and methods have been proposed. Object detection has an important role in the identification and understanding of the components of a visual scene; a statement supported by recent surveys [27], [28], [29]. Moreover, new methods such as YOLOv7 [11], DeTR [20] or DINO [25] have been demonstrating high potential, even in complicated scenarios.

B. OBJECT TRACKING

Object tracking is an essential and challenging task in computer vision, closely related to object detection, that has also grown significantly. The overall goal is to coherently establish correspondences between objects in consecutive frames and, in occlusion situations, infer the object's position to recover the tracking when it reapers [30]. Generally, the typical approach for Multiple Object Tracking (MOT) is first to apply object detection and then to apply target association. Many works that delve into these problems and other associated challenges can be found in the reviews [28], [31], [32].

As with many other areas, current tracking approaches rely on deep learning to provide effective and fast methods. In [33], an early method for combining a CNN for object detection with a Kalman filter for motion estimation and the Hungarian algorithm for tracking association was proposed for real-time tracking. Similar approaches intended to leverage the power of CNN-based object detectors with models such as Mask-RCNN [34] and SSD [35]. To overcome failures in detecting targets, Xiang et al. [36] proposed a combination of CNN+LSTM model that jointly combines target appearances and motion cues to reconstruct trajectories where gaps in detection exist.

Ma et al. [37] proposed two multi-tracking methods one for a single camera and another for multiple cameras. For a single camera, the method starts by generating tracklets for each target, followed by a Siamese bi-directional gated recurrent unit (SiaBiGRU) for trajectory post-processing. This approach cuts and reconnects tracklets to improve consistency and account for occlusions. The multi-camera method uses a Position Projection Network (PPN) that converts the trajectories from camera coordinates to world-coordinates to connect trackles from different cameras. In [38], the people detector Yolov5 is used with the DeepSORT [39] algorithm for data association to handle occlusions. Graph Convolution Neural Networks have also been applied for tracking purposes, namely for data association with promising results [40], [41]. Attention mechanisms from Transformers also enabled relevant improvements [42], [43], [44]. TrackFormer [45] used transformers for tracking, introducing the tracking-by-attention paradigm. Frame level features were extracted using a conventional CNN and self and encoder-decoder attention is applied through a transformer, achieving state-of-the-art accuracy in well-known benchmarks; HCAT [46] followed a similar strategy but with higher inference speed.

Recent works on object tracking have focused on improving the quality of the acquisition of tracklets and making the processing in real-time, even when applying very recent methodologies like transformers. Tracking is important not only to understand individual behaviors, but also how objects interact with the scene and with each other, thus assuming relevance in a scene understanding processing pipeline.

C. PEDESTRIAN ATTRIBUTE RECOGNITION

Pedestrian attribute recognition (PAR) is a sub-field of human attribute recognition (HAR) and focuses on complete human body data extracted from surveillance/monitoring scenarios. This topic targets less restrictive conditions making it essential when developing a scene understanding system that can adapt to different situations. In the following overview, we address different aspects, namely: global image-based methods, part-based methods, attention-based methods, and graph-based methods.

Abdulnabi et al. [47] proposed a global method where a multi-task learning methodology was employed so that CNN features are also used to estimate corresponding attributes in humans. Lin et al. [48] also proposed a multi-task network which simultaneously learns a re-ID embedding and predicts pedestrian attributes.

Part-based models use both local and global information to perform more accurate identifications of attributes; works such as LGNet [49] or PGDM [50] used different types of networks to jointly combine local and global information for attribute location and recognition. However, inaccuracies in the part detection procedure may result in erroneous input features to the classifiers, which may induce errors. Visual attention was applied to PAR, e.g., HydraPlus-Net [51] or DIAA [52], but with limited results. Recently, two types of visual attention consistency were enforced into a network that was capable of achieving state-of-the-art performance [53]. Graph-based approaches aim to leverage the connection between attributes and apply graph concepts. Park et al. [54] proposed an attribute and-or grammar (A-AOG) model where human body pose and attributes are inferred in a parse graph in which attributes are augmented to nodes in the hierarchical representation. In [55] and [56], HAR is viewed as a sequential attribute prediction problem and uses Graph Convolutional Network (GCN) on visual and semantic data.

Human attributes are fundamental for the characterization of humans in a scene. In scenarios of visual scene reconstruction, just placing virtual humans without corresponding

[1] A particularly useful repository for works related with vision transformers can be found in: https://github.com/IDEA-Research/awesome-detection-transformer

attributes may lead to a less than optimal reconstruction. Furthermore, knowing semantic and visual data from attributes can also help perceive the inherent semantics of the scene. Recent works on this topic that achieve high accuracy are mostly related to either using attention mechanisms or graphs. For additional information on PAR and HAR, the reader is referred to the surveys [57] and [58].

D. POSE ESTIMATION

Pose estimation is a computer vision task to predict and track the location of a person or object. This is typically done by identifying and tracking a number of keypoints on the given object or person. The keypoints are generally the major human body joints when applied to humans. Pose estimation has significant application in areas such as human activity monitoring, augmented reality, or animation. In the literature, there is a major distinction between 2D and 3D pose estimation. The first involves using visual inputs such as images or videos to predict the spatial location of 2D positions of human body key points, and the latter predicts the location of the human keypoints in a 3D space.

K-poselets [59] was proposed for 2D pose estimation and showed that CNN features could also be used for pose estimation. Deep pose [60] reinforced this by using a deep neural network that regressed the body joint locations using a cascade. Later, OpenPose [61] used spatial part affinity fields to represent the 2D orientation between limbs and achieved real-time performance, independently from the number of people in the images. OpenPose and AlphaPose [62] provide a complete API and have been used extensively. Recently, OpenPose's results were vastly improved by including data augmentation and refinements [63].

Although there has been extensive research in methods for 2D and 3D pose estimation, recent and more sophisticated approaches have been proposed for 3D pose estimation. Xu and Takano [64] used a GCN methodology to predict 3D human joint locations in the camera coordinate system from 2D inputs in the pixel domain. Similarly, in [65] a transformer-based approach was used to transform sequences of 2D joint locations to 3D poses. Li et al. [66] introduced a transformer-based method that proposes multiple pose hypothesis enabling the generation of plausible 3D human poses even with occluded body parts. For more details on this area the reader is referred to [67], [68], and [69].

E. ACTION RECOGNITION

Classifying human actions and activities is a challenging topic that benefited greatly from improvements in computational capabilities and neural networks. Action recognition and activity recognition are often used interchangeably [70]. This task is fundamental for scene understanding to capture object interactions. As actions are temporal events, it involves motion trajectory prediction and tracking.

As we have seen in other areas, deep learning has also vastly contributed to better action recognition systems. In [71], an RNN with LSTM was used to learn long-term temporal relationships to achieve spatio-temporal human action recognition in long videos with overlapping actions. In [72], a Spatial-Temporal Interaction Network was proposed to generalize action recognition regardless of the object's appearance in training. A self-supervised method named Temporal Contrastive Graph Learning (TCGL) was proposed in [73], with state-of-the-art performance. Other spatio-temporal graph-based approaches have also been proposed with competitive results [74], [75]. As temporal information is used for action recognition, Transformer-based methods have also been proposed [76], [77], [78]. A different interpretation of action recognition was made in [79]; since an event can be considered an interaction between objects and actions, actions can be decomposed as spatio-temporal scene graphs. A recent scene graph approach has also been proposed for action recognition and obtained state-of-the-art results in the AVA-Kinetics action localization task of ActivityNet Challenge 2020 [80].

Ren et al. [81] analyzed several works where action recognition algorithms are compared in multiple application scenarios. Pose estimation can also play an important role in classifying human activity [82], [83]. In [84], 2D skeleton-based action recognition methods that estimate the pose of humans from RGB images are compared and assessed. A similar study but for 3D skeleton-based action recognition was described in [85]. Further analysis on action recognition and prediction can be found in recent surveys [82], [86]. Recent works with scene graphs, CGNs and Transformers show that current proposals can provide exciting results, which hints that achieving a general scene understanding may be within reach.

F. DEPTH ESTIMATION

Depth estimation is an important task for scene understanding, allowing to determine the distance and spacial relations between elements of the scene. Advances in sensors and computational capabilities helped increase depth estimation's role in computer vision, particularly in areas such as autonomous driving or augmented reality. Traditional methods for depth estimation involved structure from motion or stereo vision matching, but this has been changing with the increased use of deep learning approaches; this is particularly noticeable in single image or monocular depth estimation, where a single RGB image is given as input to a system to estimate a depth map.

Recent works on depth estimation have focused on increasing accuracy and speed. Additionally, it's noticeable that different network structures and techniques used in other areas are also being adapted for depth estimation. In [87], wavelet decomposition was applied in an encoder-decoder approach. A multi-task learning approach where panoptic segmentation and depth estimation are jointly learned was presented in [88]; the system decomposes input images into segments, upon which an independent depth is predicted, for subsequent construction of a final complete depth map. In [89], object detection and their

associated depth are estimated jointly with an architecture based on YOLOv4. An efficient CNN that includes two shallow encoder-decoder style subnetworks was presented in [90], showing that it is possible to achieve state-of-the-art performance while requiring less computational power and at a faster speed. Work has also been proposed to achieve depth estimation using self-supervised methodologies. In [91], a ViT was used to achieve state-of-the-art performance on the well-known KITTI dataset. Similarly, in [92], a self-supervised CNN-GCN auto-encoder was used to estimate depth maps while presenting high prediction accuracy. For a more profound overview of past and recent methodologies for depth estimation, the reader is referred to [93], [94], and [95].

The usage of deep learning enabled significant accuracy increases in depth estimation. Even recent methods using self supervised learning have shown that depth estimation can be achieved even with small amounts of labeled data.

III. SCENE DESCRIPTION

In recent years, deep neural networks have been vastly used to achieve an understanding of a visual scene with respect to multiple tasks, such as Image Recognition [96], [97], Object Segmentation [98], [99], Object Recognition [100], [101], as we have seen previously. This resulted in the definition and development of multiple backbone networks that are able to extract valuable information from an image. Nonetheless, it is also essential for the information extracted to be readily available for storage and use.This section explores methodologies and ways for information associated with visual scenes to be represented and stored. It briefly explores knowledge bases, metadata models, and visual relationship extraction. Finally, it describes scene graphs, a useful primitive used in the context of computer vision which allows to define attributes and the relationships between objects in a scene.

A. ONTOLOGY-BASED KNOWLEDGE REPRESENTATION

Storing information related to a visual scene in a comparable way is crucial to many application scenarios, and work has been proposed toward this end. For instance, a graph-based approach that follows some of the OWL principles [102] to provide an information storage layer for the development of complex driving applications, is presented in [103].

Similarly, a Semantic Scene Graph is proposed in [104], where the model describes the dynamic elements of a traffic scene and their relationships in a graph format, which is then exported to the dot language [105]. An ontology-based approach is also presented in [106], where three ontologies are used to generate use cases for autonomous vehicles. Another graph-based approach is presented in [107], where the concept of a Road Scene Graph is proposed to provide a graph that represents traffic information, which is then used to provide a synthetic traffic scene generator. In this representation, the nodes correspond to actors and the edges to relationships between actors, which aid in the definition of actor's initial status and trajectories.

Work has also been done on providing representations for multimedia content using ontologies. For instance, Video Ontology (VidOnt) [108] provides an ontology defined in OWL for multimedia content that incorporates spatio-temporal annotations that integrate, in its vocabulary, elements from multimedia metadata standards such as MPEG-7, EBU CCDM, and Dublin Core. The need for introducing semantics in content descriptors has also been identified for parliamentary video content understanding [109] and television broadcasted content [110], where standard metadata schemas have been translated into ontologies. An ontology-based fuzzy video semantic content model for object, event, and concept extraction was also proposed in [111]. In [112], the problem of populating a multimedia ontology is addressed, and a multi-modality approach that combines textual and visual information obtained from CNNs is used to automatize the process. A semi-automatic NLP-guided framework for ontology generation for multimedia representation and information retrieval is presented in [113], where spatial, temporal, occurrence-based actions, descriptive verbs, and prepositions are represented in the generated ontology.

As described above, many works employ graph structures to store multimedia content. Within these proposals, there is a notion of a scene graph, which has been defined differently and applied in different forms in multiple works, which suggests that there is a problem of standardization of the terminologies and notions used. This way, it is possible to create an explicit way to describe image features associated with objects and their inherent relationships using Scene Graphs. As seen in Figure 3, a scene graph is able to express the semantic meaning of an image or part of it. Inherently, this means that a subgraph can also express portions of an image.

B. SCENE-GRAPH-BASED KNOWLEDGE REPRESENTATION

As a scene graph is a structured representation of information that can be extracted from images, the task of scene graph generation (SGG) can be defined as the process of generating a visually-grounded scene graph that accurately associates a scene graph to an image. With this notion, the nodes of a scene graph correspond to object instances with their associated bounding box and category, which is specific to the scene and to the algorithm used to detect it; and the edges represent the pair-wise relationships, with either other objects for intra-object relationships and with the object itself, representing attributes. Due to the inherent value of this semantic data association, scene graphs have been applied to multiple tasks, such as image captioning [115], [116], visual question answering [117], [118], image generation [119], [120]. This review starts by presenting a formal definition of a scene graph and its generation. It then explores different SGG methodologies and their associated problems for single images and videos, exploring current spatio-temporal scene graph methods.

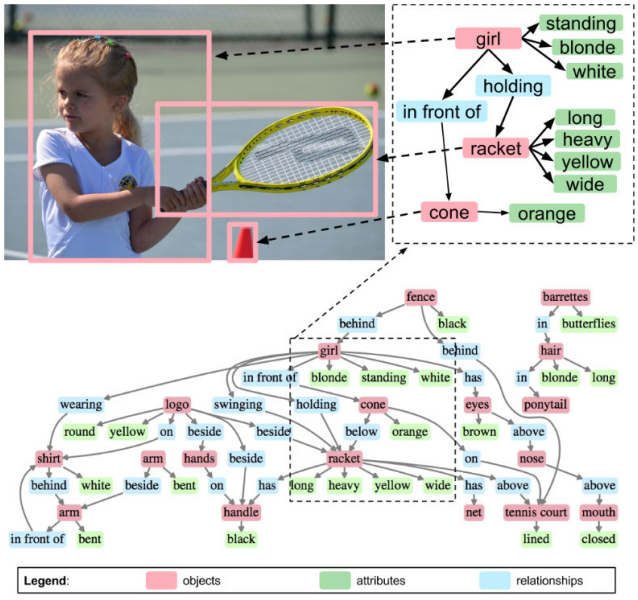

FIGURE 3. An example of a scene graph (taken from [114]). In the bottom there is a full scene graph of a scene, that contains a subgraph that expresses the semantic information of a portion of the scene, represented by the image. A scene graph is able to encode Objects ("girl") that can have attributes ("girl is standing") and relations ("girl holding racket").

Formally, a scene graph G is a data structure of the form of a directed graph, which can be represented by the tuple $G = (O, E, R)$, where $O = \{o_1 \dots o_n\}$ is a set of objects detected in the image, which can be people ("girl", "boy"), places ("street", "balcony"), parts of objects ("arm", "leg"), or things ("shirt", "bottle"). The objects are represented by $o_i = (b_i, c_i)$, where $b_i \in \mathbb{R}^4$ is the object bounding box detected on the image and $c_i \in C$ is the semantic label of the object given a pre-defined set of object classes C. R represents the relationships between pairs of object instances and can be represented by $r_{i \to j}, i, j \in \{1, 2, \dots n\}$. The edges of the graph are of the form $E \subseteq O \times R \times O$ and represent the connections between the object instances and the relationship nodes. Thus, the initial graph can have at most $n \times n$ edges. For a more detailed description of the scene graph generation process, we refer to [121].

Different methods have been proposed for SGG. On one side, we have two-stage detectors where objects are first detected using specific object detection networks and the relationships and the graph generation are made on another step atop of the detections. On another side, some methodologies jointly infer the object classes, localization and their relationships [122]. When dealing with video, image-based graph generation applied at frame level does not consider the temporal aspects. However, benefits from exploring the temporal information can be foreseen as it can contribute to correct some inconsistencies at frame level.

Regarding image-based approaches, iterative message passing was first proposed in [123] for SGG, where contextual information is used to improve object and relationship estimations. This methodology was revolutionary and is still used nowadays in many methods. Neural Motifs [124] were also important contributions, showing that leveraging reoccurring patterns in scene graphs helps increase performance. Contextual information was also used in [125] where a Relation Proposal Network (RePN) is proposed to deal with the dimensionality problem of object relations, drastically reducing the number of relations that actually need to be accounted for. As with many other areas, Transformers have also been successfully used in generating scene graphs with competent results [126], [127]. A fully convolutional SGG method was proposed in [128], showing that using a pre-trained detector is not necessary and good results can also be obtained, even high zero-shot recall. Another unified framework named Structured Sparse R-CNN was proposed in [129], obtaining state-of-the-art results in the well-known Visual Genome [130] and OpenImages V4/V6 [131] datasets.

One of the main problems of scene graph generation is the long tail dataset distribution, in which meaningful but rare relations are often not considered the most probable relations in trained models, resulting in biased scene graphs. There has been extensive research on this topic, and recent approaches have achieved important progress. For instance, in [132], a confidence-aware bipartite graph neural network with an adaptive message propagation mechanism is proposed for unbiased scene graph generation. In [133], a Dynamic Label Frequency Estimation (DLFE) is also proposed to treat the bias problem while training a network. Atom Correlation Based Graph Propagation (AC-GP) was also proposed in [134] to deal with complex and cluttered visual relationships, and results showed impressive improvements in detecting infrequent and missed relationships.

We believe that the proposal presented by Johnson et al. [135] for scene graphs in the context of still images can also be used for video content, as it defines a scene graph as a topological representation of a scene, where objects, attributes and the relationships between objects in a scene are established in a data structure such that the semantic information of the scene can be expressed by it.

C. SCENE-GRAPH-BASED KNOWLEDGE REPRESENTATION FOR VIDEO CONTENT

While, there has been extensive work done on SGG for images, the usage of video is still being incorporated nowadays. For instance, in [136] a transformer-based approach is proposed, where information from videos with exocentric and egocentric views, as well as individual frames, is used in a 3d CNN model to improve SGG performance. Cong et al. [137] proposed a Spatial-temporal Transformer (STTran) for dynamic scene graph generation that is able to process videos of varying lengths without clipping them. Teng et al. [138] targets the task of video scene graph generation (VidSGG) and proposes a framework for frame-level VidSGG that can also be applied for video-level VidSGG by incorporating a temporal association strategy. Gao et al. [139] also addresses VidSGG by proposing temporal bipartite graphs which take a classification-then-grounding framework instead of the traditional proposal-based framework and

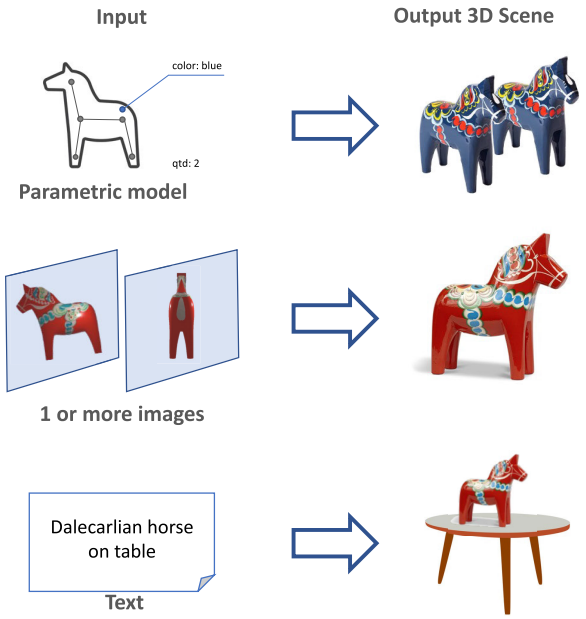

FIGURE 4. Example synthesis modalities. Among others, we overview methods to synthesize a 3D scene from parametric models, one or more images, and text.

obtains competitive results on multiple datasets. Other approaches using anticipatory pre-training [140], transformers [141] and meta training [142] have also been proposed for VidSGG. A particularly interesting approach is proposed in [143], where a tracking-based approach is used to explicitly associate spatio-temporal contexts thus identifying spatio-temporal human-object interactions while simultaneously localizing humans and objects. Additionally, the authors propose a new simulated dataset that contains consistent temporal annotation of relationships.

Current approaches for scene graph generation, for both images and videos are able to achieve competent results in most datasets. However, they are still not a popular tool that is applied in many application scenarios. We believe that the usage of scene graphs as an actual format or schema to store, maintain and distribute visual and semantic information extracted from a scene is not being explored and this could potentiate a step forward on achieving a scene understanding system or environment that could be adapted to many application scenarios.

IV. SCENE SYNTHESIS

3D virtual assets are commonly used across various industries for special effects in movies, computer games, and product advertisements. However, producing realistic and believable 3D content is challenging and requires a thorough understanding of computer graphics. To streamline and automate this process, researchers have studied algorithmic approaches to creating 3D virtual content. In this section we overview several modalities for creating virtual content, from different types of input data. We depict an example of currently researched modalities for 3D generation in Figure 4.

Procedural generation was one of the first techniques to create virtual content. It initially focused on particular graphic elements such as plants [144], and was applied to create larges scenes like cities [145], terrains [146] or realistic buildings [147].

With the growth of machine learning techniques there has been an increased use of neural networks to automatically create 3D content. One area of particular interest is the use of images to generate 3D scenes, with research exploring the reconstruction of both individual objects and entire scenes from single or multiple images [148], [149], [150], [151], [152], [153], [154]. Many of these approaches focused on category-specific object-level reconstruction from a single image [148], [149]. However, there have also been efforts towards category-agnostic methods [150], as well as the reconstruction of 3D object models from a limited number of images [151], [152]. In scene-level reconstruction, researchers have demonstrated the ability to generate 3D indoor scenes from a single image [153], [154]. These techniques can produce different 3D representations, including voxels, meshes, and depth maps, and can be useful according to the specific use cases.

Recent research has focused on the creation of neural representations of scenes that can predict novel views from a limited set of images or depth data. The work in [155] represents the scene geometry and appearance as continuous functions that map world coordinates to a feature representation of local scene properties. In [156], authors use a set of images with known cameras to create a neural network representation of a scene that outputs the volume density and view-dependent emitted radiance at a spatial location and uses it to synthesize new views. A neural network was also used to learn the surface light fields from an input image and predict unknown views and scene lighting in [157]. Other recent methods encode signed distance functions (SDFs) with a large, fixed-size neural network to approximate complex shapes with implicit surfaces [158], [159].

The generation of digital humans has been the target of attention for a long time. In particular, human parametric models are useful because they allow for the creation of human representations that are highly detailed and accurate, while also being flexible and customized. These have been used for multiple scenarios such as 3D human pose and shape estimation [84], controllable 3D human synthesis [85] or virtual try-ons of clothing [160]. Graph Convolutional Networks have also been researched to generate 3D human shapes with better resolution [161].

A different perspective for generating 3D content is the ability to create 3D shapes from text. Recent works [162], [163], [164] explored this idea and used natural language as input to a network that generates high-quality 3D textured meshes, ranging from cars, animals, and human characters to buildings. Opposite to these approaches that focus on finding better modeling assumptions for training on a fixed dataset, DALL-E [165] proposed a transformer-based approach that showed good one-shot generalization results

on tasks the model had not been specifically trained on. Another approach, combining high-level descriptions and 2D images [166], presented a memory-efficient methodology based on octrees.

This overview demonstrates the pervasive use of machine learning and neural networks to generate 3D content automatically, showcasing multiple application scenarios where 3D content generation is based on text, images, or models. These imply a dependency on the techniques applied with the type of content to be recreated. See [167], [168], [169], [170] to support the selection of techniques and for further details on methods and methodologies for 3D object reconstruction.

V. DATASETS

As described in the previous sections, machine learning-based methods have been assuming increased importance in different domains. However, applying these methodologies, particularly the ones based on deep-learning, is accompanied by a dependency on the datasets used for training, which must capture scenario-specific characteristics. As a result, considerable efforts have been made toward preparing and making datasets available. The following subsections identify and briefly describe relevant datasets for scene analysis and scene reconstruction.[2]

A. DATASETS FOR SCENE ANALYSIS

The richness of visual scenes led to the preparation of many datasets from different research areas relevant to scene understanding. This section explores datasets tailored for many tasks associated with scene analysis to present the reader with an entry point to help train models for different application scenarios. This overview addresses datasets and challenges related to object detection; visual relationship and scene graph generation; action recognition; human attribute recognition; 2D and 3D human pose estimation; gait recognition.

1) OBJECT DETECTION AND TRACKING DATASETS

Object detection and tracking have been a research topic for many years, with early datasets still being used. These include the MIT pedestrian dataset [171] for human detection in images and the INRIA Pedestrian Dataset [6] for detecting pedestrians in images and videos. Both are composed of small amounts of images that are considered too small to train deep-learning models. The Caltech Pedestrian Dataset [172] is also an image dataset tailored for pedestrian detection and consists of roughly 10 hours of video taken from a driving car in a regular urban area.

The PASCAL Visual Object Classes (VOC) Challenges were a series of competitions running from 2005 to 2012 that enabled the evaluation and comparison of different computer vision algorithms, providing both a standardized image dataset for object class recognition and a common set of tools for accessing the data sets and annotations. From the proposed datasets, the two most used datasets for object detection are the VOC07 [173] and VOC12 [174], with amounts of data fairly larger than recent datasets. Another challenge was the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [175] (2010-2017) for evaluation of algorithms for object detection and image classification at a large scale; the images used for training and testing came from the manually annotated ImageNet [176] dataset, organized according to the WordNet [177] hierarchy, using only the nouns. Recently, the Microsoft COCO (MS-COCO) [178], and Open Images [131] datasets represented a major shift in the usage of datasets for training models; these are large datasets that contain millions of labeled instances, thus providing a great starting point for training algorithms. MS-COCO contains annotations for detection, segmentation, and captioning, with 2.5 million labeled instances. The Open Images dataset contains image-level labels, object bounding boxes, object segmentation masks, visual relationships, and localized narratives, with 16 million bounding boxes for 600 object classes on 1.9 million images and 3.3 million annotations from 1466 distinct relationship triplets used for visual relationship estimation. An overview of the data format and size of datasets is provided in Table 1.

TABLE 1. Object detection and tracking datasets.

Dataset	Year	Format	Citations
MIT [171]	2000	709 images	1000+
Caltech [172]	2009	250K frames	1000+
Inria [6]	2005	842 images	10k+
Pascal Voc 2007 [173]	2007	10K images	10k+
Pascal Voc 2012 [174]	2012	11K images	10k+
ILSVRC [175]	2015	1.2M images	10k+
MS-COCO [178]	2014	328K images	10k+
Open Images [131]	2017	1.9M images	615

2) VISUAL RELATIONSHIP AND SCENE GRAPH GENERATION

The Real-World Scene Graphs Dataset(RWSG) [135] was the first dataset explicitly created for scene graph generation. It consists of 5000 images extracted from MS-COCO [178] and processed by Amazon's Mechanical Turk [179] to produce human-generated scene graphs. The Visual Relationship Dataset (VRD) [180] was constructed for visual relationship prediction; it has 100 object classes and 37993 relationships. The Visual Genome Dataset (VGD) [130] dataset and knowledge base connects structured image concepts to English language terms, with millions of labeled attributes and relationships. Associated with this dataset is the VG150 [123] and VrR-VG [181], where the former filters out rare instances and the latter removes redundant and visually irrelevant relationships. UnRel Dataset (UnRel-D) [182] is a small dataset for visual relationships with unusual language triplets and is considered a good benchmark for testing the generalization of models. Similarly, HCVRD [183] is a benchmark for large-scale human-centered visual relationship detection. A problem with most of these datasets is the long tail distribution problem, which essentially translates to common relationships having a large amount of instances while rare

[2]A searchable list with more information on the mentioned datasets is available at: http://mct.inesctec.pt/americo-2

TABLE 2. Visual relationship and scene graph generation.

Dataset	Year	Format	Citations
RWSG [135]	2015	5000 images	895
VRD [180]	2016	5000 images	982
VGD [130]	2017	108K images	3k+
HCVRD [183]	2018	52K images	30
VidVRD [184]	2017	800 videos	89
VidOR [185]	2019	10K videos	81
ActionGenome [79]	2020	9848 videos	183

and sometimes more meaningful relationships have minimal amounts of instances.

ImageNet-VidVRD [184] is a video visual relationship detection dataset composed of 1000 videos from ILSVRC2016-VID; the VidOR [185] dataset consists of 10,000 videos (98.6 hours) from YFCC100M [186] with dense annotations for 80 categories of objects and 50 categories of relation predicates. ActionGenome [79] builds upon the Charades dataset [187] and is a large-scale video dataset that provides human-object relationships in multiple videos.

The aforementioned datasets are summarized in in Table 2.

3) ACTION RECOGNITION

Human action recognition requires good quality and large amounts of training data to obtain reliable models that can adapt to multiple situations. Moreover, a small number of actors may introduce bias during training [83]. HMDB51 [188] is an interesting dataset that contains videos obtained from multiple sources with different camera settings and lighting conditions. Sports1M [189] was introduced in 2014 and is considered the first large-scale video dataset for annotated actions. YouTube8M [190] is a very large-scale video dataset containing 8 million videos from YouTube; however, only a portion of the data has human-verified labels. The Charades [187] dataset was tailored for video activity recognition and commonsense reasoning for daily human activities. Available online videos have been used for many action recognition datasets, namely: the Kinetics series [191], [192], [193], [194], [195]; AVA Actions [194]; HACS [196]; HVU (Holistic Video Understanding) [197]. Moments in Time(MiT) [198] has been widely used for recognizing and understanding action in videos; it contains 1 million labeled 3-second videos involving people, animals, objects, or natural phenomena. BABEL [199] was recently proposed with language labels describing actions performed in MOCAP sequences. These datasets contain large amounts of annotated data, some with human interaction and others obtained by automatic processes, as the sheer amount of videos demanded more automated methodologies. To better understand the magnitude of these datasets, an overview is presented in Table 3.

4) HUMAN ATTRIBUTE RECOGNITION

Human Attribute Recognition (HAR) datasets focus primarily on cropped human images where the different parts of the human body are delimited and annotated. Early datasets such as HAT [200] or Berkeley-Attributes of People (BAP) [201]

TABLE 3. Action recognition datasets.

Dataset	Year	Format	Citations
HMDB51 [188]	2011	6849 videos	3k+
Sports1M [189]	2014	1.13m videos	7k+
YouTube8M [190]	2016	8m videos	1k+
Charades [187]	2016	9848 videos	951
Kinetics [191]–[193]	2017	650k videos	2k+
AVA [194]	2018	430 videos	748
MiT [198]	2019	1m videos	433
HACS [196]	2019	50k videos	172
HVU [197]	2020	577k videos	60
BABEL [199]	2021	63k frames	38

TABLE 4. Human attribute recognition.

Dataset	Year	Format	Citations
PETA [202]	2014	19k images	379
Parse 27K [203]	2015	8 videos	181
RAP-2.0 [204]	2018	85k images	109
HAT [200]	2011	9344 images	100
BAP [201]	2011	8k images	474
PA-100K [51]	2017	100k images	482

had a small number of attributes (9 and 27, respectively) annotated and were composed of less than 10000 images. The PEdesTrian Attribute dataset (PETA) [202] was proposed in 2014 and contained more data than the previous datasets and 65 annotated attributes. PARSE-27K [203] addresses the HAR problem with a different type of data, as it is composed of 7 video sequences taken from a moving camera in a city. PA-100K [51] also addressed outdoor environments and is a large-scale dataset composed of 100k images with varying resolutions. Recently, RAP-2.0 [204] was proposed with an even larger attribute count of 72 and was captured by 25 cameras in a surveillance network on an indoor shopping mall. Table 4 shows an overview of these datasets.

5) 2D AND 3D HUMAN POSE ESTIMATION

The estimation of human pose is a challenging topic that can be divided into different sub-problems, namely: 2D human pose estimation; 3D human pose estimation; single person or multi person pose estimation. Naturally, since the nature of these problems differs, several datasets became available for human pose estimation. This section explores a relevant set of the most known datasets for the above-mentioned topics.

Starting with 2D pose estimation, the Leeds Sports Pose (LSP) Dataset [205] was proposed in 2010 for single person pose estimation. MPII Dataset [206] is a large-scale dataset made by using Amazon Mechanical Turk and contains 2D locations of 16 keypoints, full 3D torso and head orientations, occlusion labels for keypoints and activity labels. More recently, CrowdPose [207] was designed for human pose estimation in crowded scenarios. Joint-annotated HMDB (J-HMDB) [208], a subset of the HMDB51 database, as proposed for 2D video pose estimation. More recent datasets, such as the PoseTrack [209] and Human-in-Events (HiEve) [210], are large-scale datasets that include multiple scenarios of varying difficulties.

The topic of 2D pose estimation is currently closely followed by 3D pose estimation that not only estimates the

TABLE 5. Pose estimation.

Dataset	Year	Format	Citations
LSP [205]	2010	2k images	893
MPII [206]	2014	25k images	2282
CrowdPose [207]	2019	28k images	316
J-HMDB [208]	2013	928 videos	805
PoseTrack [209]	2018	550 videos	351
HiEve [210]	2020	49k frames	76
Human3.6M [211]	2013	3.6m images	2266
CMU Panoptic [212]	2015	65 videos	583
3DPW [213]	2018	60 videos	522
AMASS [214]	2019	40h of videos	452
MoVi [215]	2020	1068 videos	36
SURREAL [216]	2017	6m frames	812
JTA [217]	2018	512 videos	129

location of human keypoints, but also estimates their location in 3D. Human3.6M [211] and CMU Panoptic [212] are examples of early datasets designed for 3D pose estimation. The dataset 3DPW [213] includes videos captured in multiple scenarios, including taken from moving phones. Annotations include 2D and 3D pose information, 3D body scannings and SMPL parameters. More recently, AMASS [214] was proposed as a large-scale motion capture (MoCap) dataset comprised of a unification of 15 MoCap datasets by converting them to the SMPL parameters. MoVi [215] contains synchronized pose with body mesh and video recordings. There has also been work done on creating synthetic datasets with precise 3D pose annotations of humans with datasets such as SURREAL [216] and Joint Track Auto (JTA) [217]. An overview of pose estimation datasets is summarised in Table 5.

6) GAIT RECOGNITION

Gait recognition aims at analyzing how people move and has been an important research topic in vision-based systems, in areas such as sports or rehabilitation. There have been proposals of datasets and benchmarks since to 2006 with CASIA [218], [219]. Ever since, there have been proposals for datasets with more details and resolution, as well as greater quantities of labeled data. GAID (TUM Gait from Audio, Image, and Depth) [220] is a multi-modal gait dataset that includes RGB, depth, and audio data. As for large-scale datasets, there have been various proposals. OU-ISIR [221] includes images captured in indoor halls using four cameras and OU-MVLP [222] encompasses multi-view data obtained from a seven-network camera system and its extension; OUMVLP-Pose [223] was built upon the OU-MVLP dataset, adding the pose estimations obtained for each of the subjects using the OpenPose [61] and AlphaPose [62] algorithms. More recently, GREW [224] was constructed from natural videos, containing multiple hours of content in open systems. An overview of gait recognition datasets is present in Table 6.

B. DATASETS FOR SCENE RECONSTRUCTION

With regards to 3D reconstruction there are multiple different types of datasets, depending on the scope; Some are focused on humans and others on objects. The provided information

TABLE 6. Gait estimation.

Dataset	Year	Format	Citations
CASIA [218], [219]	2006	19139 images	928
TUM GAID [220]	2012	305 subjects	219
OU-ISIR [221]	2012	4k subjects	222
OU-MVLP [222]	2018	10k subjects	212
OUMVLP-Pose [223]	2020	10k subjects	64
GREW [224]	2021	26k subjects	31

TABLE 7. Scene reconstruction.

Dataset	Year	Format	Citations
SUN RGB-D [225]	2015	10k images	1458
ScanNet [226]	2017	2.5m views	2060
UP-3D [228]	2017	59k images	444
SURREAL [216]	2017	6m frames	812
Scan2CAD [230]	2019	14k CAD models	153
KITTI-360 [231]	2021	320k images	83
RELLIS-3D [232]	2020	6k images	59
Hypersim [233]	2021	77.4k images	69

is also different and may include depth information (RGB-D), synthetic data or CAD models. SUN RGB-D [225] is a dataset for scene understanding providing RGB-D data, with 3D bounding boxes with object orientation, and 3D room layout and category for the scenes. ScanNet [226] provides reconstructed surface mesh files and was recently updated to provide data for tasks such as, 3D object classification and segmentation; semantic voxel labeling or CAD model retrieval. The ShapeNet [227] is an ongoing large-scale dataset of 3D shapes that has richly annotated 3D data using the WordNet hierarchy. Work was also done in generating 3D body models for other publicly available datasets, as with the UP-3D [228], which uses the SMPLify [229] method to generate the 3D models. Surreal [216] also applies a similar method; however, in this case, entirety of the data was generated providing a large-scale synthetic dataset with synthetic humans rendered photo-realistically under large variations of shape, texture, viewpoint and pose. For CAD models, Scan2CAD [230] is a scan to CAD alignment dataset that pairs CAD models from ShapeNet and their corresponding objects in ScanNet scans. More recently, KITTI-360 [231] contains dense semantic and instance annotations for both 3D point clouds and 2D images captured from multiple sensors in an urban area from a moving vehicle. RELLIS-3D [232] is a similar dataset captured in an off-road environment. Hypersim [233] is a recent dataset that provides photorealistic synthetic data for holistic indoor scene understanding. Table 7 summarizes some details about these datasets.

VI. DISCUSSION

Since the early days of computer vision, we have seen a constant influx of research to allow more information to be extracted from visual scenes and improve the accuracy and quality of the information extracted. Recent advances in machine learning enabled rapid progress in automatic human detection and tracking, pose estimation, depth estimation, action recognition, and many others. Coupled with data representation and 3D scene reconstruction improvements,

a complete automatic scene understanding approach for virtual synthesis is becoming possible. To assess this possibility, we present a cross-domain analysis of different techniques and methodologies that can allow a transition between scene analysis and a posterior 3D virtual synthesis.

We start this discussion with a brief overview of the main areas presented in this survey: (i) scene analysis; (ii) scene representation and (iii) scene synthesis (presented in the following paragraphs). To finalize, we will describe the challenges toward a unified scene understanding framework, discussing the advantages and research opportunities a cross-domain approach can promote.

1) SCENE ANALYSIS

Early scene analysis algorithms targeting object detection and tracking relied on hand-crafted features that often produced erroneous results when applied in unconstrained scenarios, and a significant amount of work naturally focused on humans and their behavior. The introduction of machine learning and deep learning promoted a resurgence of interest, resulting in new proposals that improved the quality of human and object detection, tracking, and other areas of scene analysis. These advances also pushed the definition and creation of large-scale datasets so that algorithms could better generalize and produce more accurate results, along with many benchmarking efforts. Finally, recent research focused on deeper and more complex network architectures, such as Transformers or GCNs, and continued to offer steady improvements.

2) SCENE REPRESENTATION

In terms of information representation, there has also been a notorious shift toward using neural networks to aid processes. However, we notice that no particular representation standard is available to address the needs of a generic scenario. On the one hand, we have ontologies with vocabulary definitions and well structured models for representing data. On another, we have graph-based approaches that allow complex semantic data to also be structured, but following less rules and standards. This divergence in what's the best practice and what should be applied hinders the uniformization and unification of methods on more complex or general application scenarios. We observed that there has also been a particular interest in the research community for defining and automatically obtaining scene graphs for both frame-level and video-level content and that research primarily focused on how to solve problems related to the definition and generation of scene graphs rather than investigating their applicability in different scenarios.

3) SCENE SYNTHESIS

Scene synthesis has also seen a revival of interest, with current approaches dedicated to recreating both objects or a complete 3D scene from one or multiple images. In addition, recent work creates neural representations, such as radiance fields or SDFs of scenes, that can predict novel views from a limited set of images or depth data. Finally, we also see advances in generating 3D human shapes using deep learning techniques and reconstructing entire 3D scenes from a text description.

4) A CROSS-DOMAIN APPROACH

The advances in scene analysis, scene description, and synthesis described above, suggest that a cross-domain approach is feasible. A framework that integrates these areas can provide a unified scene understanding framework that takes advantage of the continuous algorithmic improvements we observe in each area, and deliver different information extracted from scene analysis algorithms in a structured way so that that information can be interchanged between algorithms seamlessly. Having data stored and represented in a well-established structure can also enable the integration of this framework with other external applications. For instance, in scenarios where only the semantic knowledge is meaningful, transmission requirements could be drastically reduced by only transmitting a compact semantic summary of the scene and using it on other endpoints to reconstruct the underlying scene semantics in a virtual environment. The definition of a well-established data structure also provides elasticity to the framework by increasing or decreasing its complexity depending on the detail levels required for applications that integrate this framework. Data augmentation and synthetic data generation adaptable to different scenarios are also possible by processing a visual scene and editing the underlying semantics.

Although a framework incorporating these concepts could enable many applications, it would require precise definitions of the entire framework and the articulation and integration of many algorithms that required different types of data and have varying complexity constraints. We believe this framework could promote multiple research opportunities such as, but not limited to:

- definition of a unified and standard structure for the output of algorithms encompassing the different scene analysis domains to correctly interpret data provided by different algorithms, ensuring its plasticity and ability to plug and unplug algorithms;
- complementary data definitions and structures to provide an adaptable and well-structured data management and storage strategy;
- automatic virtual synthesis methodologies that are easily adaptable to such framework;
- methods to scene semantics to correct errors;
- definition of data standards by the scientific community for better integration of algorithms.

Addressing these challenges enables cross-domain methodologies applied to different use cases, and allows taking advantage of the continuous algorithmic improvements we observe in scene analysis, scene description, and synthesis described in this survey.

REFERENCES

[1] C. Yan, X. Chang, M. Luo, H. Liu, X. Zhang, and Q. Zheng, "Semantics-guided contrastive network for zero-shot object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 4, 2022, doi: 10.1109/TPAMI.2021.3140070.

[2] Y. Tian, H. Zhang, Y. Liu, and L. Wang, "Recovering 3D human mesh from monocular images: A survey," 2022, *arXiv:2203.01923*.

[3] V. Lazova, V. Guzov, K. Olszewski, S. Tulyakov, and G. Pons-Moll, "Control-NeRF: Editable feature volumes for scene rendering and manipulation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4329–4339.

[4] X. Gao, J. Yang, J. Kim, S. Peng, Z. Liu, and X. Tong, "MPS-NeRF: Generalizable 3D human rendering from multiview images," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 12, 2022, doi: 10.1109/TPAMI.2022.3205910.

[5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, 2001, p. I.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 1. Red Hook, NY, USA: Curran, 2012, pp. 1097–1105.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[10] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "You only learn one representation: Unified network for multiple tasks," 2021, *arXiv:2105.04206*.

[11] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[12] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[13] C. Wang, A. Bochkovskiy, and H. M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13024–13033.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[15] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[16] A. Pramanik, S. K. Pal, J. Maiti, and P. Mitra, "Granulated RCNN and multi-class deep SORT for multi-object detection and tracking," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 1, pp. 171–181, Feb. 2022.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[19] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, "Toward transformer-based object detection," 2020, *arXiv:2012.09958*.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[22] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," 2022, *arXiv:2203.16527*.

[23] F. Liu, X. Zhang, Z. Peng, Z. Guo, F. Wan, X. Ji, and Q. Ye, "Integrally migrating pre-trained transformer encoder–decoders for visual object detection," 2022, *arXiv:2205.09613*.

[24] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask DINO: Towards a unified transformer-based framework for object detection and segmentation," 2022, *arXiv:2206.02777*.

[25] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.

[26] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surveys*, vol. 54, no. 10s, pp. 1–41, Jan. 2022.

[27] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*.

[28] M. Bashar, S. Islam, K. K. Hussain, M. B. Hasan, A. B. M. A. Rahman, and M. H. Kabir, "Multiple object tracking in recent times: A literature review," 2022, *arXiv:2209.04796*.

[29] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *Digit. Signal Process.*, vol. 126, Jun. 2022, Art. no. 103514.

[30] P. Carvalho, T. Oliveira, L. Ciobanu, F. Gaspar, L. F. Teixeira, R. Bastos, J. S. Cardoso, M. S. Dias, and L. Corte-Real, "Analysis of object description methods in a video object tracking environment," *Mach. Vis. Appl.*, vol. 24, no. 6, pp. 1149–1165, Aug. 2013. [Online]. Available: http://www.inescporto.pt/~jsc/publications/journals/2013PCarvalhoMVAP.pdf

[31] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61–88, Mar. 2020.

[32] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T.-K. Kim, "Multiple object tracking: A literature review," *Artif. Intell.*, vol. 293, Apr. 2021, Art. no. 103448.

[33] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3464–3468.

[34] Z. Zhou, J. Xing, M. Zhang, and W. Hu, "Online multi-target tracking with tensor-based high-order graph matching," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1809–1814.

[35] D. Zhao, H. Fu, L. Xiao, T. Wu, and B. Dai, "Multi-object tracking with correlation filter for autonomous vehicle," *Sensors*, vol. 18, no. 7, p. 2004, Jun. 2018.

[36] J. Xiang, G. Zhang, and J. Hou, "Online multi-object tracking based on feature representation and Bayesian filtering within a deep learning architecture," *IEEE Access*, vol. 7, pp. 27923–27935, 2019.

[37] C. Ma, F. Yang, Y. Li, H. Jia, X. Xie, and W. Gao, "Deep trajectory post-processing and position projection for single & multiple camera multiple object tracking," *Int. J. Comput. Vis.*, vol. 129, no. 12, pp. 3255–3278, Dec. 2021.

[38] Y. Wang and H. Yang, "Multi-target pedestrian tracking based on YOLOv5 and DeepSORT," in *Proc. IEEE Asia–Pacific Conf. Image Process., Electron. Comput. (IPEC)*, Apr. 2022, pp. 508–514.

[39] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.

[40] J. Lee, M. Jeong, and B. C. Ko, "Graph convolution neural network-based data association for online multi-object tracking," *IEEE Access*, vol. 9, pp. 114535–114546, 2021.

[41] G. Wang, R. Gu, Z. Liu, W. Hu, M. Song, and J. Hwang, "Track without appearance: Learn box and tracklet embedding with local and global motion patterns for vehicle tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9856–9866.

[42] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "MOTR: End-to-end multiple-object tracking with transformer," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 659–675.

[43] S. Gao, C. Zhou, C. Ma, X. Wang, and J. Yuan, "AiATrack: Attention in attention for transformer visual tracking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 146–164.

[44] F. Ma, M. Z. Shou, L. Zhu, H. Fan, Y. Xu, Y. Yang, and Z. Yan, "Unified transformer tracker for object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8771–8780.

[45] T. Meinhardt, A. Kirillov, L. Leal-Taixé, and C. Feichtenhofer, "TrackFormer: Multi-object tracking with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8834–8844.

[46] X. Chen, B. Kang, D. Wang, D. Li, and H. Lu, "Efficient visual tracking via hierarchical cross-attention transformer," 2022, *arXiv:2203.13537*.

[47] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task CNN model for attribute prediction," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1949–1959, Nov. 2015.

[48] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, Nov. 2019.

[49] P. Liu, X. Liu, J. Yan, and J. Shao, "Localization guided learning for pedestrian attribute recognition," 2018, *arXiv:1808.09102*.

[50] D. Li, X. Chen, Z. Zhang, and K. Huang, "Pose guided deep model for pedestrian attribute recognition in surveillance scenarios," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[51] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "HydraPlus-Net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 350–359.

[52] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 680–697.

[53] H. Guo, X. Fan, and S. Wang, "Visual attention consistency for human attribute recognition," *Int. J. Comput. Vis.*, vol. 130, no. 4, pp. 1088–1106, Apr. 2022.

[54] S. Park, B. X. Nie, and S. Zhu, "Attribute and-or grammar for joint parsing of human pose, parts and attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1555–1569, Jul. 2018.

[55] Q. Li, X. Zhao, R. He, and K. Huang, "Visual-semantic graph reasoning for pedestrian attribute recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, 2019, pp. 8634–8641.

[56] Z. Tan, Y. Yang, J. Wan, G. Guo, and S. Li, "Relation-aware pedestrian attribute recognition with graph convolutional networks," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, vol. 34, no. 7, pp. 12055–12062.

[57] E. Yaghoubi, F. Khezeli, D. Borza, S. A. Kumar, J. Neves, and H. Proença, "Human attribute recognition—A comprehensive survey," *Appl. Sci.*, vol. 10, no. 16, p. 5608, Aug. 2020.

[58] X. Wang, S. Zheng, R. Yang, A. Zheng, Z. Chen, J. Tang, and B. Luo, "Pedestrian attribute recognition: A survey," *Pattern Recognit.*, vol. 121, Jan. 2022, Art. no. 108220.

[59] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik, "Using k-poselets for detecting people and localizing their keypoints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3582–3589.

[60] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.

[61] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.

[62] H. Fang, S. Xie, Y. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2353–2362.

[63] T. Kitamura, H. Teshima, D. Thomas, and H. Kawasaki, "Refining OpenPose with a new sports dataset for robust 2D pose estimation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2022, pp. 672–681.

[64] T. Xu and W. Takano, "Graph stacked hourglass networks for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16100–16109.

[65] W. Li, H. Liu, R. Ding, M. Liu, P. Wang, and W. Yang, "Exploiting temporal contexts with Strided Transformer for 3D human pose estimation," *IEEE Trans. Multimedia*, vol. 25, pp. 1282–1293, 2023.

[66] W. Li, H. Liu, H. Tang, P. Wang, and L. Van Gool, "MHFormer: Multi-hypothesis transformer for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13137–13146.

[67] E. S. Dos Reis, L. A. Seewald, R. S. Antunes, V. F. Rodrigues, R. D. R. Righi, C. A. D. Costa, L. G. D. Silveira Jr., B. Eskofier, A. Maier, T. Horz, and R. Fahrig, "Monocular multi-person pose estimation: A survey," *Pattern Recognit.*, vol. 118, Oct. 2021, Art. no. 108046.

[68] W. Liu, Q. Bao, Y. Sun, and T. Mei, "Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective," *ACM Comput. Surv.*, vol. 55, no. 4, pp. 1–41, Apr. 2023.

[69] Z.-U.-D. Muhammad, Z. Huang, and R. Khan, "A review of 3D human body pose estimation and mesh recovery," *Digit. Signal Process.*, vol. 128, Aug. 2022, Art. no. 103628.

[70] J. R. Pinto, T. Gonçalves, C. Pinto, L. Sanhudo, J. Fonseca, F. Gonçalves, P. Carvalho, and J. S. Cardoso, "Audiovisual classification of group emotion valence using activity recognition networks," in *Proc. IEEE 4th Int. Conf. Image Process., Appl. Syst. (IPAS)*, Dec. 2020, pp. 114–119.

[71] R. D. Brehar, M. P. Muresan, T. Marita, C. Vancea, M. Negru, and S. Nedevschi, "Pedestrian street-cross action recognition in monocular far infrared sequences," *IEEE Access*, vol. 9, pp. 74302–74324, 2021.

[72] J. Materzynska, T. Xiao, R. Herzig, H. Xu, X. Wang, and T. Darrell, "Something-else: Compositional action recognition with spatial–temporal interaction networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1046–1056.

[73] Y. Liu, K. Wang, H. Lan, and L. Lin, "Temporal contrastive graph learning for video action recognition and retrieval," 2021, *arXiv:2101.00820*.

[74] D. Yang, M. Mengqi Li, H. Fu, J. Fan, Z. Zhang, and H. Leung, "Unifying graph embedding features with graph convolutional networks for skeleton-based action recognition," 2020, *arXiv:2003.03007*.

[75] P. Ghosh, Y. Yao, L. S. Davis, and A. Divakaran, "Stacked spatio-temporal graph convolutional networks for action segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 565–574.

[76] R. Girdhar, J. J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 244–253.

[77] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Comput. Vis. Image Understand.*, vols. 208–209, Jul. 2021, Art. no. 103219.

[78] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge, "Action transformer: A self-attention model for short-time pose-based human action recognition," *Pattern Recognit.*, vol. 124, Apr. 2022, Art. no. 108487.

[79] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, "Action genome: Actions as compositions of spatio-temporal scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10233–10244.

[80] J. Pan, S. Chen, M. Z. Shou, Y. Liu, J. Shao, and H. Li, "Actor-context-actor relation network for spatio-temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 464–474.

[81] B. Ren, M. Liu, R. Ding, and H. Liu, "A survey on 3D skeleton-based action recognition using learning method," 2020, *arXiv:2002.05907*.

[82] L. Song, G. Yu, J. Yuan, and Z. Liu, "Human pose estimation and its application to action recognition: A survey," *J. Vis. Commun. Image Represent.*, vol. 76, Apr. 2021, Art. no. 103055.

[83] L. Capozzi, V. Barbosa, C. Pinto, J. R. Pinto, A. Pereira, P. M. Carvalho, and J. S. Cardoso, "Toward vehicle occupant-invariant models for activity characterization," *IEEE Access*, vol. 10, pp. 104215–104225, 2022.

[84] K. Wang, G. Zhang, and J. Yang, "3D human pose and shape estimation with dense correspondence from a single depth image," *Vis. Comput.*, vol. 39, pp. 429–441, Jan. 2022.

[85] T. Xu, Y. Fujita, and E. Matsumoto, "Surface-aligned neural radiance fields for controllable 3D human synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15862–15871.

[86] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *Int. J. Comput. Vis.*, vol. 130, no. 5, pp. 1366–1401, May 2022.

[87] M. Ramamonjisoa, M. Firman, J. Watson, V. Lepetit, and D. Turmukhambetov, "Single image depth prediction with wavelet decomposition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11084–11093.

[88] L. Wang, J. Zhang, O. Wang, Z. Lin, and H. Lu, "SDC-depth: Semantic divide-and-conquer network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 538–547.

[89] J. Yu and H. Choi, "YOLO MDE: Object detection with monocular depth estimation," *Electronics*, vol. 11, no. 1, p. 76, Dec. 2021.

[90] X. Dong, M. A. Garratt, S. G. Anavatti, and H. A. Abbass, "MobileXNet: An efficient convolutional neural network for monocular depth estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 20134–20147, Nov. 2022.

[91] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia, "MonoViT: Self-supervised monocular depth estimation with a vision transformer," 2022, *arXiv:2208.03543*.

[92] A. Masoumian, H. A. Rashwan, S. Abdulwahab, J. Cristiano, M. S. Asif, and D. Puig, "GCNDepth: Self-supervised monocular depth estimation based on graph convolutional network," *Neurocomputing*, vol. 517, pp. 81–92, Jan. 2023.

[93] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian, "Monocular depth estimation based on deep learning: An overview," *Sci. China Technol. Sci.*, vol. 63, no. 9, pp. 1612–1627, Sep. 2020.
[94] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: A review," *Neurocomputing*, vol. 438, pp. 14–33, May 2021.
[95] A. Mertan, D. J. Duff, and G. Unal, "Single image depth estimation: An overview," *Digit. Signal Process.*, vol. 123, Apr. 2022, Art. no. 103441.
[96] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[97] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[98] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[99] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
[100] A. Kortylewski, Q. Liu, A. Wang, Y. Sun, and A. Yuille, "Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion," *Int. J. Comput. Vis.*, vol. 129, no. 3, pp. 736–760, Mar. 2021.
[101] A. Salari, A. Djavadifar, X. R. Liu, and H. Najjaran, "Object recognition datasets and challenges: A review," *Neurocomputing*, vol. 495, pp. 129–152, Jul. 2022.
[102] G. Antoniou and F. van Harmelen, "Web ontology language: OWL," Berlin, Germany: Springer, 2004, pp. 67–92, doi: 10.1007/978-3-540-24750-0_4.
[103] S. Ulbrich, T. Nothdurft, M. Maurer, and P. Hecker, "Graph-based context representation, environment modeling and information aggregation for automated driving," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2014, pp. 541–547.
[104] M. Zipfl and J. M. Zollner, "Towards traffic scene description: The semantic scene graph," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 3748–3755.
[105] AT&T Research and Lucent Bell Labs. (2022). *DOT Language*. Accessed: Jan. 4, 2023. [Online]. Available: https://graphviz.org/doc/info/lang.html
[106] W. Chen and L. Kloul, "An ontology-based approach to generate the advanced driver assistance use cases of highway traffic," in *Proc. 10th Int. Joint Conf. Knowl. Discovery, Knowl. Eng. Knowl. Manag.*, 2018, pp. 1–10.
[107] Y. Tian, A. Carballo, R. Li, and K. Takeda, "Real-to-synthetic: Generating simulator friendly traffic scenes from graph representation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2022, pp. 1615–1622.
[108] L. F. Sikos, "VidOnt: A core reference ontology for reasoning over video scenes," *J. Inf. Telecommun.*, vol. 2, no. 2, pp. 192–204, Apr. 2018.
[109] E. Sánchez-Nielsen, F. Chávez-Gutiérrez, and J. Lorenzo-Navarro, "A semantic parliamentary multimedia approach for retrieval of video clips with content understanding," *Multimedia Syst.*, vol. 25, no. 4, pp. 337–354, Aug. 2019.
[110] P. Viana and A. P. Alves, "A semantic management model to enable the integrated management of media and devices," *Multimedia Tools Appl.*, vol. 49, no. 1, pp. 37–62, Aug. 2010.
[111] Y. Yildirim, A. Yazici, and T. Yilmaz, "Automatic semantic content extraction in videos using a fuzzy ontology and rule-based model," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 47–61, Jan. 2013.
[112] M. Muscetti, A. M. Rinaldi, C. Russo, and C. Tommasino, "Multimedia ontology population through semantic analysis and hierarchical deep features extraction techniques," *Knowl. Inf. Syst.*, vol. 64, no. 5, pp. 1283–1303, May 2022.
[113] A. S. Patel, G. Merlino, A. Puliafito, R. Vyas, O. P. Vyas, M. Ojha, and V. Tiwari, "An NLP-guided ontology development and refinement approach to represent and query visual information," *Exp. Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118998.
[114] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1219–1228.
[115] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10677–10686.
[116] M. Mozes, M. Schmitt, V. Golkov, H. Schutze, and D. Cremers, "Scene graph generation for better image captioning?" 2021, *arXiv:2109.11398*.
[117] R. Koner, H. Li, M. Hildebrandt, D. Das, V. Tresp, and S. Gunnemann, "Graphhopper: Multi-hop scene graph reasoning for visual question answering," in *Proc. Int. Semantic Web Conf.* Cham, Switzerland: Springer, 2021, pp. 111–127.
[118] V. Damodaran, S. Chakravarthy, A. Kumar, A. Umapathy, T. Mitamura, Y. Nakashima, N. Garcia, and C. Chu, "Understanding the role of scene graphs in visual question answering," 2021, *arXiv:2101.05479*.
[119] X. Zhao, L. Wu, X. Chen, and B. Gong, "High-quality image generation from scene graphs with transformer," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.
[120] R. Sortino, S. Palazzo, and C. Spampinato, "Transforming image generation from scene graphs," 2022, *arXiv:2207.00545*.
[121] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, "A comprehensive survey of scene graphs: Generation and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1–26, Jan. 2023.
[122] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, and X. Wang, "Factorizable Net: An efficient subgraph-based framework for scene graph generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 335–351.
[123] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3097–3106.
[124] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5831–5840.
[125] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 670–685.
[126] M. Andrews, Y. K. Chia, and S. Witteveen, "Scene graph parsing by attention graph," 2019, *arXiv:1909.06273*.
[127] R. Li, S. Zhang, and X. He, "SGTR: End-to-end scene graph generation with transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19464–19474.
[128] H. Liu, N. Yan, M. Mortazavi, and B. Bhanu, "Fully convolutional scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11541–11551.
[129] Y. Teng and L. Wang, "Structured sparse R-CNN for direct scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19415–19424.
[130] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.
[131] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, and A. Veit, "OpenImages: A public dataset for large-scale multi-label and multi-class image classification," *Dataset*, vol. 2, no. 3, p. 18, 2017.
[132] R. Li, S. Zhang, B. Wan, and X. He, "Bipartite graph network with adaptive message passing for unbiased scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11104–11114.
[133] M.-J. Chiou, H. Ding, H. Yan, C. Wang, R. Zimmermann, and J. Feng, "Recovering the unbiased scene graphs from the biased ones," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1581–1590.
[134] B. Lin, Y. Zhu, and X. Liang, "Atom correlation based graph propagation for scene graph generation," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108300.
[135] J. Johnson, R. Krishna, M. Stark, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3668–3678.
[136] Y. Lu, C. Chang, H. Rai, G. Yu, and M. Volkovs, "Multi-view scene graph generation in videos," in *Proc. Int. Challenge Activity Recognit.*, vol. 3, 2021, p. 2.
[137] Y. Cong, W. Liao, H. Ackermann, B. Rosenhahn, and M. Y. Yang, "Spatial–temporal transformer for dynamic scene graph generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16352–16362.
[138] Y. Teng, L. Wang, Z. Li, and G. Wu, "Target adaptive context aggregation for video scene graph generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13668–13677.

[139] K. Gao, L. Chen, Y. Niu, J. Shao, and J. Xiao, "Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19475–19484.

[140] Y. Li, X. Yang, and C. Xu, "Dynamic scene graph generation via anticipatory pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13864–13873.

[141] Y. Cong, M. Ying Yang, and B. Rosenhahn, "RelTR: Relation transformer for scene graph generation," 2022, *arXiv:2201.11460*.

[142] L. Xu, H. Qu, J. Kuen, J. Gu, and J. Liu, "Meta spatio-temporal debiasing for video scene graph generation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 374–390.

[143] Y. Qiu, Y. Nagasaki, K. Hara, H. Kataoka, R. Suzuki, K. Iwata, and Y. Satoh, "VirtualHome action genome: A simulated spatio-temporal scene graph dataset with consistent relationship labels," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3340–3349.

[144] P. Prusinkiewicz and A. Lindenmayer, *The Algorithmic Beauty of Plants*. Berlin, Germany: Springer-Verlag, 1990. [Online]. Available: https://archive.org/details/algorithmicbeaut0000prus/page/101

[145] Y. I. H. Parish and P. Muller, "Procedural modeling of cities," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn.*, New York, NY, USA, Aug. 2001, pp. 301–308, doi: 10.1145/383259.383292.

[146] J. Dormans, "Adventures in level design: Generating missions and spaces for action adventure games," in *Proc. Workshop Procedural Content Gener. Games*, Jun. 2010, pp. 1–8.

[147] D. Salpisti, M. de Clerk, S. Hinz, F. Henkies, and G. Klinker, "A procedural building generator based on real-world data enabling designers to create context for XR automotive design experiences," in *Virtual Reality and Mixed Reality*, G. Zachmann, M. A. Raya, P. Bourdot, M. Marchal, J. Stefanucci, and X. Yang, Eds. Cham, Switzerland: Springer, 2022, pp. 149–170.

[148] H. Xie, H. Yao, X. Sun, S. Zhou, and S. Zhang, "Pix2Vox: Context-aware 3D reconstruction from single and multi-view images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2690–2698.

[149] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3D-R2N2: A unified approach for single and multi-view 3D object reconstruction," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 628–644.

[150] Z. Xing, Y. Chen, Z. Ling, X. Zhou, and Y. Xiang, "Few-shot single-view 3D reconstruction with memory prior contrastive network," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 55–70.

[151] Z. Yang, Z. Ren, M. A. Bautista, Z. Zhang, Q. Shan, and Q. Huang, "FvOR: Robust joint shape and pose optimization for few-view object reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2487–2497.

[152] L. C. O. Tiong, D. Sigmund, and A. B. J. Teoh, "3D-C2FT: Coarse-to-fine transformer for multi-view 3D reconstruction," 2022, *arXiv:2205.14575*.

[153] Y. Nie, X. Han, S. Guo, Y. Zheng, J. Chang, and J. J. Zhang, "Total3DUnderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 55–64.

[154] C. Zhang, Z. Cui, Y. Zhang, B. Zeng, M. Pollefeys, and S. Liu, "Holistic 3D scene understanding from a single image with implicit representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8829–8838.

[155] V. Sitzmann, M. Zollhofer, and G. Wetzstein, "Scene representation networks: Continuous 3D-structure-aware neural scene representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–12.

[156] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, p. 99–106, Dec. 2021, doi: 10.1145/3503250.

[157] M. Oechsle, M. Niemeyer, C. Reiser, L. Mescheder, T. Strauss, and A. Geiger, "Learning implicit surface light fields," in *Proc. Int. Conf. 3D Vis. (3DV)*, Los Alamitos, CA, USA, Nov. 2020, pp. 452–462, doi: 10.1109/3DV50981.2020.00055.

[158] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning continuous signed distance functions for shape representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 165–174.

[159] T. Takikawa, J. Litalien, K. Yin, K. Kreis, C. Loop, D. Nowrouzezahrai, A. Jacobson, M. McGuire, and S. Fidler, "Neural geometric level of detail: Real-time rendering with implicit 3D shapes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11353–11362.

[160] R. Vidaurre, I. Santesteban, E. Garces, and D. Casas, "Fully convolutional graph neural networks for parametric virtual try-on," *Comput. Graph. Forum*, vol. 39, no. 8, pp. 145–156, 2020.

[161] H. Yu, C. Cheang, Y. Fu, and X. Xue, "Multi-view shape generation for a 3D human-like body," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 1, pp. 1–22, Jan. 2023.

[162] K. Chen, C. B. Choy, M. Savva, A. X. Chang, T. Funkhouser, and S. Savarese, "Text2Shape: Generating shapes from natural language by learning joint embeddings," 2018, *arXiv:1803.08495*.

[163] Z. Liu, Y. Wang, X. Qi, and C. Fu, "Towards implicit text-guided 3D shape generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17875–17885.

[164] J. Gao, T. Shen, Z. Wang, W. Chen, K. Yin, D. Li, O. Litany, Z. Gojcic, and S. Fidler, "GET3D: A generative model of high quality 3D textured shapes learned from images," 2022, *arXiv:2209.11163*.

[165] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.

[166] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2107–2115.

[167] K. Fu, J. Peng, Q. He, and H. Zhang, "Single image 3D object reconstruction based on deep learning: A review," *Multimedia Tools Appl.*, vol. 80, no. 1, pp. 463–498, Jan. 2021.

[168] Z. Kang, J. Yang, Z. Yang, and S. Cheng, "A review of techniques for 3D reconstruction of indoor environments," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 5, p. 330, May 2020.

[169] G. Fahim, K. Amin, and S. Zarif, "Single-view 3D reconstruction: A survey of deep learning methods," *Comput. Graph.*, vol. 94, pp. 164–190, Feb. 2021.

[170] A. Morales, G. Piella, and F. M. Sukno, "Survey on 3D face reconstruction from uncalibrated images," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100400.

[171] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comput. Vis.*, vol. 38, no. 1, pp. 15–33, 2000.

[172] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 304–311.

[173] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. (2007). *The PASCAL Visual Object Classes Challenge*. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

[174] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. (2012). *The PASCAL Visual Object Classes Challenge*. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

[175] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[176] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[177] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[178] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[179] M. Buhrmester, T. Kwang, and S. D. Gosling, *Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?* (Methodological Issues and Strategies in Clinical Research), 4th ed. Washington, DC, USA: American Psychological Association, 2016, doi: 10.1037/14805-009.

[180] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 852–869.

[181] Y. Liang, Y. Bai, W. Zhang, X. Qian, L. Zhu, and T. Mei, "VrR-VG: Refocusing visually-relevant relationships," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10402–10411.

[182] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Weakly-supervised learning of visual relations," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2017, pp. 5179–5188.

[183] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. Van Den Hengel, "HCVRD: A benchmark for large-scale human-centered visual relationship detection," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[184] X. Shang, T. Ren, J. Guo, H. Zhang, and T.-S. Chua, "Video visual relation detection," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1300–1308.

[185] X. Shang, D. Di, J. Xiao, Y. Cao, X. Yang, and T.-S. Chua, "Annotating objects and relations in user-generated videos," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2019, pp. 279–287.

[186] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100M: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, Jan. 2016.

[187] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 510–526.

[188] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[189] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[190] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8M: A large-scale video classification benchmark," 2016, *arXiv:1609.08675*.

[191] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[192] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," 2018, *arXiv:1808.01340*.

[193] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," 2019, *arXiv:1907.06987*.

[194] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6047–6056.

[195] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu, and A. Zisserman, "A short note on the Kinetics-700–2020 human action dataset," 2020, *arXiv:2010.10864*.

[196] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "HACS: Human action clips and segments dataset for recognition and temporal localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8667–8677.

[197] A. Diba, M. Fayyaz, V. Sharma, M. Paluri, J. Gall, R. Stiefelhagen, and L. V. Gool, "Large scale holistic video understanding," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 593–610.

[198] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick, and A. Oliva, "Moments in time dataset: One million videos for event understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 502–508, Feb. 2020.

[199] A. R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quirós-Ramírez, and M. J. Black, "BABEL: Bodies, action and behavior with English labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 722–731.

[200] G. Sharma and F. Jurie, "Learning discriminative spatial representation for image classification," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 1–11.

[201] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1543–1550.

[202] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 789–792.

[203] P. Sudowe, H. Spitzer, and B. Leibe, "Person attribute recognition with a jointly-trained holistic CNN model," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 329–337.

[204] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1575–1590, Apr. 2019.

[205] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. Brit. Mach. Vis. Conf.*, 2010, p. 5.

[206] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.

[207] J. Li, C. Wang, H. Zhu, Y. Mao, H. Fang, and C. Lu, "CrowdPose: Efficient crowded scenes pose estimation and a new benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10855–10864.

[208] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2013, pp. 3192–3199.

[209] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, "PoseTrack: A benchmark for human pose estimation and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5167–5176.

[210] W. Lin, H. Liu, S. Liu, Y. Li, R. Qian, T. Wang, N. Xu, H. Xiong, G.-J. Qi, and N. Sebe, "Human in events: A large-scale benchmark for human-centric video analysis in complex events," 2020, *arXiv:2005.04490*.

[211] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

[212] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3334–3342.

[213] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using IMUs and a moving camera," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 601–617.

[214] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. Black, "AMASS: Archive of motion capture as surface shapes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5441–5450.

[215] S. Ghorbani, K. Mahdaviani, A. Thaler, K. Kording, D. J. Cook, G. Blohm, and N. F. Troje, "MoVi: A large multipurpose motion and video dataset," 2020, *arXiv:2003.01888*.

[216] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proc. CVPR*, 2017, pp. 109–117.

[217] M. Fabbri, F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, and R. Cucchiara, "Learning to detect and track visible and occluded body joints in a virtual world," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 430–446.

[218] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 441–444.

[219] D. Tan, K. Huang, S. Yu, and T. Tan, "Efficient night gait recognition based on template matching," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 1000–1003.

[220] M. Hofmann, S. Bachmann, and G. Rigoll, "2.5D gait biometrics using the depth gradient histogram energy image," in *Proc. IEEE 4th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2012, pp. 399–403, doi: 10.1109/BTAS.2012.6374606.

[221] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1511–1521, Oct. 2012, doi: 10.1109/TIFS.2012.2204253.

[222] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 1, pp. 1–14, Dec. 2018.

[223] W. An, S. Yu, Y. Makihara, X. Wu, C. Xu, Y. Yu, R. Liao, and Y. Yagi, "Performance evaluation of model-based gait on multi-view very large population database with pose sequences," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 4, pp. 421–430, Oct. 2020.

[224] Z. Zhu, X. Guo, T. Yang, J. Huang, J. Deng, G. Huang, D. Du, J. Lu, and J. Zhou, "Gait recognition in the wild: A benchmark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 14789–14799.
[225] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 567–576.
[226] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2432–2443.
[227] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*.
[228] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3D and 2D human representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4704–4713. [Online]. Available: http://up.is.tuebingen.mpg.de
[229] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 561–578.
[230] A. Avetisyan, M. Dahnert, A. Dai, M. Savva, A. X. Chang, and M. Nießner, "Scan2CAD: Learning CAD model alignment in RGB-D scans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2609–2618.
[231] Y. Liao, J. Xie, and A. Geiger, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D," 2021, *arXiv:2109.13410*.
[232] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "RELLIS-3D dataset: Data, benchmarks and analysis," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2020, pp. 1110–1116.
[233] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind, "Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10892–10902.

AMÉRICO PEREIRA received the bachelor's and M.Sc. degrees in computer science from the Faculty of Science, University of Porto, Portugal, in 2011 and 2013, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the Faculty of Engineering, University of Porto. He joined INESC TEC, in 2014, where he is currently a Researcher with the Center of Telecommunications and Multimedia. His research interests include computer vision and image/video processing, with an emphasis on machine learning.

PEDRO CARVALHO (Senior Member, IEEE) received the degree in electrical and computer engineering, the M.Sc. degree in network and communication services, and the Ph.D. degree in electrical and computers engineering from the Faculty of Engineering, University of Porto, Portugal, in 2001, 2004, 2012, respectively. He joined INESC TEC, in 2001, where he is currently a Senior Researcher with the Center of Telecommunications and Multimedia. He has been an invited Adjunct Professor with the School of Engineering, Polytechnic Institute of Porto, Porto, since 2014. His current research interests include image/video processing and computer vision.

NUNO PEREIRA was a Visiting Scholar with Carnegie Mellon University and the Executive Director of the CONIX Research Center, from 2019 to 2022. He is currently a Professor (an Adjunct Professor) with the School of Engineering, Polytechnic Institute of Porto (ISEP), where he is also working on mixed reality and the infrastructure needed to create distributed applications that seamlessly span the cloud and edge. He has developed research in distributed embedded systems with more than 50 technical articles in peer-reviewed scientific venues. He was involved in numerous national and international research projects. He is the co-inventor of several patents.

PAULA VIANA (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Porto, in 2008. She is currently a Coordinator Professor with the School of Engineering, Polytechnic Institute of Porto, and the Head of multimedia communication technologies with INESC TEC. She has more than 30 years of experience in the area of multimedia content analysis and management, computer vision, and multimedia metadata. She has been coordinating the participation of INESC TEC in several national and European projects. She is the author of several publications, an active reviewer for several journals and conferences, and European and Portuguese research projects. She has been involved in the organization of several scientific events, including the Immersive Media Experiences Workshop Series at ACM Multimedia, from 2013 to 2015.

LUÍS CÔRTE-REAL (Member, IEEE) was born in Vila do Conde, Portugal, in 1958. He received the degree in electrical engineering from the Faculty of Engineering, University of Porto, Portugal, in 1981, the M.Sc. degree in electrical and computers engineering from Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisbon, Portugal, in 1986, and the Ph.D. degree from the Faculty of Engineering, University of Porto, in 1994. In 1984, he joined the University of Porto as a Lecturer of telecommunications. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Faculty of Engineering, University of Porto. He has been a Researcher with INESC TEC, since 1985. His current research interests include image/video processing and coding.