香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Structured Representation for Visual Data and Its Applications

ZHANG Yong
章勇

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Computer and Information Engineering

The Chinese University of Hong Kong, Shenzhen
June 9, 2023

Thesis Assessment Committee

Professor LI Zhen (Chair)
Professor HUANG Rui (Thesis Supervisor)
Professor MEI Tao (Thesis Co-supervisor)
Professor CHEN Chang-Wen (Thesis Co-supervisor)
Professor XU Jie (Committee Member)
Professor YU Bei (Examiner from CUHK)
Professor HE Zhihai (External Examiner)

Abstract

The world has stepped into a multimedia big data era. An explosively huge amount of visual data (e.g., images, videos) have brought opportunities for mining the underlying great value, yet tremendous computational challenges emerge since raw visual data are unstructured. In this thesis, we focus on designing efficient models/algorithms to generate semantic structured representations for visual data and utilizing such representations to facilitate high-level applications. Particularly, we adopt the *scene graph* structured representation, which describes visual semantics by encoding object entities as nodes and their relations as edges in a symbolic graph data structure.

Though great effort has been made, there are still various challenges in generating useful scene graphs for practical applications: 1) existing methods for scene graph generation (SGG) bias to predict frequent and uninformative relations between objects; 2) training SGG models requires time-consuming ground-truth annotations; 3) the closed-set object categories make the SGG models limited in their ability to recognize novel objects outside of training corpora. To address the first issue, in Chapter 3, we introduce visual relation saliency to align scene graph representation with human visual perception. Specifically, we propose a novel model dubbed Saliency-guided Message Passing (SMP), which jointly learns visual relation saliency estimation and scene graph generation. The proposed SMP can generate informative scene graphs, which have also been validated to effectively support downstream applications like cross-modal retrieval and image captioning. In Chapter 4, we address the latter two challenges in a "two birds with one stone" fashion. Particularly, we novelly exploit a powerful visual-semantic space (VSS) via large-scale language-image pre-training to trigger language-supervised and open-vocabulary SGG in a simple yet effective manner.

Moreover, we pursue high-quality visual structure parsing from the perspective of model architecture design. In Chapter 5, we devise the Structure-aware Transformer over Interaction Proposals (STIP) model, particularly for extracting human-centric visual structure (i.e., humans, objects and their interactions) in SGG. STIP upgrades vanilla Transformer by additionally encoding the holistic semantic structure among interaction proposals and the local spatial structure of human/object within each interaction proposal, so as to strengthen predictions. Taking a step further, in Chapter 6, we tackle the challenging video SGG task by proposing a novel end-to-end Transformer-based architecture. Such design jointly models the three sub-tasks in video SGG (i.e., object detection, instance temporal association and relation recognition) through a monolithic spatio-temporal Transformer, making it fully exploit the temporal dynamics and consistency across video frames to strengthen relation recognition. In summary, our proposed solutions and empirical results set new state-of-the-arts for semantic representation, comprehension and applications of visual data in a structural way.

摘要

世界已经步入了多媒体大数据时代。爆炸式增长的海量视觉数据（如图像、视频）蕴藏着巨大的价值，然而由于原始视觉数据是非结构化的，因此在计算与信息挖掘方面也面临巨大挑战。此研究论文专注于设计高效的模型与算法，为视觉数据生成语义结构化表示，并利用这种表示来支撑高层级的应用。特别地，我们采用了场景图（scene graph）结构化表示形式，它通过将对象实体抽象为节点、将对象实体之间的关系抽象为边来描述视觉数据的语义信息。

尽管前人已经付出了诸多努力，生成实用的场景图仍然面临多方面的挑战，比如：1）现有的场景图生成方法偏向于预测高频而信息量低的目标间关系；2）训练场景图生成模型依赖昂贵耗时的监督标签；3）训练时封闭的目标类别限制了场景图生成模型对于语料库以外新颖类别的识别能力。 针对第一个挑战，在 Chapter 3 中，我们引入视觉关系显著度，将场景图表示和人类视觉感知对齐。具体来说，我们设计了新颖的显著度引导的消息传递模型，它将视觉关系显著度估计和场景图生成进行联合学习。该模型不仅获得了高质量的的场景图，同时在支撑跨模态检索和以图生文等下游应用中，也被验证具有很好的泛化能力。 在 Chapter 4 中，我们提出了“一石二鸟”的解决方案来应对后两个挑战。具体而言，我们创新性地采用通过大规模语言-图像预训练构建的视觉-语义空间，使得场景图生成任务只需要文本数据来监督训练，且能够识别开放的目标类别。

此外，我们也从模型架构设计的角度探索高质量的视觉结构解析效果。在 Chapter 5 中，我们设计了一种新颖的 Transformer 结构的模型，即基于交互提案的结构敏感 Transformer 模型，专注于解析场景图生成任务中以人为中心的结构（人、物以及它们之间的关系/交互）。该模型对原始的 Transformer 网络进行了升级，增加了交互提案间的整体语义结构编码，以及每个交互提案内的局部空间结构编码，以增强预测能力。 更进一步地，在 Chapter 6 中，我们设计了一种全新的端到端的基于 Transformer 的方法，用来处理极具挑战性的视频场景图生成任务。该方法利用一个完全统一的时空 Transformer 架构，联合建模视频场景图生成过程中的三个子任务（即目标检测、目标时序关联和关系识别），这使其充分利用视频的动态信息及上下文强化对视觉关系的识别。 总体而言，我们提出的一系列解决方案和所取得的实验结果，为视觉数据结构化表示、理解与应用开辟了新局面。

Acknowledgements

I would like to express my deepest gratitude and appreciation to all those who have supported, encouraged, and contributed to the successful completion of this thesis. It is truly a privilege to have had the opportunity to learn from and collaborate with such an extraordinary group of individuals.

First and foremost, I would like to thank my supervisors, Prof. Rui Huang, Prof. Tao Mei and Prof. Chang-Wen Chen, for their endless support, guidance, and mentorship throughout my doctoral journey. Their expertise, patience, and enthusiasm have been invaluable to my academic and personal growth. I am truly grateful for the great patience they have dedicated to helping me refine my ideas, challenge my assumptions, and navigate the complexities of structured visual representation and applications.

I am very fortunate to have had several mentors, Dr. Yingwei Pan and Dr. Ting Yao, during my internship at JD AI Research. They have provided a stimulating and supportive research environment. I am grateful for their valuable insights and collaboration. Our many discussions and brainstorming sessions have enriched my research experience and helped improve my professionalism and critical thinking. They also offered me a great deal of advice on career development. I would also like to acknowledge JD.com for providing the financial support that made this research possible.

I want to thank my friends, Yang Chen, Jianjie Luo, Rui Zhu, Yan Shu, Qi Cai, Zhongwei Zhang and Panwen Hu etc., for their support, encouragement, and companionship during this challenging period, especially overcoming the COVID-19 pandemic. The free and easy brainstorming discussions we have are really helpful. Their friendship has made the entire process more enjoyable and fulfilling.

I would also like to extend my sincere gratitude to thesis committee members, Prof. Zhen Li, Prof. Jie Xu, Prof. Bei Yu, and Prof. Zhihai He, for their invaluable feedback, constructive criticism, and insightful suggestions. Their unique perspectives and rigorous standards have greatly enhanced the quality of my work and the clarity of my thinking.

On a personal note, I would like to express my profound appreciation to my family, especially my wife, for their unlimited love, encouragement, and understanding throughout this journey. Their unwavering belief in my abilities and constant support are my sources of strength and motivation.

Contents

Chapter 1

Introduction

1.1 Motivation

The world has stepped into a multimedia big data era [1]. Over the years, due to the proliferation of ubiquitous cameras (e.g., smartphone cameras, surveillance cameras), the easy accessibility of communication networks (e.g., mobile networks, internet of things, cloud computing) and the flourishing of online platforms (e.g., TikTok, Twitter, Instagram), visual data (e.g., images, videos) increase explosively. Such a tremendous amount and growing speed impose significant challenges for various management and analytic aspects, such as storage, indexing, mining, and retrieval. Despite these challenges, the great value of such big visual data should be acknowledged, since these data depict how the physical world runs, capture people's behaviors, record events, and even reflect public sentiments and opinions.

However, raw visual data are **unstructured** and do not readily provide any high-level semantics. Note that when people say the word "structure", they usually mean *the arrangement of and relations between the parts or elements of something complex*. Raw visual data are considered unstructured because they consist of a large array of pixel values that lack explicit organization, relationships, or meaning about the real-world scene. To make such data useful for knowledge extraction and further decision-making, we envision a system to first transform them into structured representations that are expressed symbolically and aligned with human cognition [2, 3]. In particular, such structured representations should involve organizing the raw visual data in a way that captures high-level semantics including the objects and their relationships in the visual scene, making it easier for algorithms or humans to interpret, analyze, and utilize.

Using graph data structure is a natural choice for structured visual representation. Specifically, the **scene graph** representation is proposed to describe visual semantics [4], which encodes object entities and their relations in the visual scene to graph nodes and edges. As such, visual data in the form of pixel values are transformed into symbolic semantic graphs. This offers great advantages in many aspects:

1. Such structured representation aligns with human visual perception as we tend to perceive the world in terms of objects and their interactions. We humans often

abstract the raw visual perceptual input into semantic concepts like objects and understand their compositions as relations [5]. Similarly, scene graph representations naturally model visual scenes as a collection of objects (nodes) and their relationships (edges). By capturing object relationships and their properties as graph edges, such graph-based representation enables a high-level understanding of scenes, which is essential for tasks such as visual reasoning and scene understanding like humans.

2. Scene graphs can serve as a shared semantic structure to bridge multi-modal and multi-source information. Both visual information and natural language can be represented as graphs, which facilitate connections between vision and language. For example, objects and relationships in the graph can be linked to words or phrases in a language, making it easier to integrate information from both modalities.

3. Extracted knowledge in scene graphs can be easily inspected and interpreted. This makes such representation and its downstream applications more explainable and trustworthy, which is more desirable than black-box deep learning representations.

Actually, in recent years, there have been a bunch of works that demonstrate the great potential of scene graphs in supporting various downstream applications, such as image retrieval [4, 6, 7, 8], image captioning [9, 10, 11, 12, 13], visual question answering [14, 15, 16, 17] and cross-media knowledge graph [18].

Though great effort has been made, the current state of AI is not capable of generating structured visual representations for visual data with practical accuracy and efficiency. In this thesis, we focus on designing effective algorithms to extract high-quality scene graph representations from visual data and exploring their usage to facilitate high-level applications.

1.2 Problem Statement

There is rich modern literature investigating the generation and application of structured visual representations. This is studied comprehensively in Chapter 2, which particularly includes a summarization of the main limitations of existing works. Here we specify the key directions that we have pursued to address these limitations, finally achieving new state-of-the-arts for semantic representation, comprehension and applications of visual data in a structural way.

- One prominent issue in existing scene graph generation (SGG) methods is that they bias to predict frequent relations between objects. This often leads to uninformative extracted scene graphs, in which salient objects and relationships that sketch key image content are not prioritized. In Chapter 3, we introduce visual relation saliency to align scene graph representation with human visual perception. Specifically, we propose a novel model dubbed Saliency-guided Message

Passing (SMP), which jointly learns visual relation saliency estimation and scene graph generation. The proposed SMP can generate informative scene graphs, which have also been validated to effectively support downstream applications like cross-modal retrieval and image captioning. This work has been published on ACM TOMM (2022) [1].

- Another two knotty obstacles limit the feasibility of current SGG methods in real-world scenarios: 1) training SGG models requires time-consuming ground-truth annotations, and 2) the closed-set object categories make the SGG models limited in their ability to recognize novel objects outside of training corpora. To address these issues, we present a "two birds with one stone" solution in Chapter 4. Particularly, we novelly exploit a powerful visual-semantic space (VSS) via large-scale language-image pre-training to trigger language-supervised and open-vocabulary SGG in a simple yet effective manner. We validate our proposed approach with extensive experiments on the Visual Genome benchmark across various SGG scenarios (i.e., supervised / language-supervised, closed-set / open-vocabulary). Consistent superior performances are achieved compared with existing methods, demonstrating the potential of exploiting pre-trained VSS for SGG in more practical scenarios. This work is accepted by CVPR′23 [2].

- The architecture design of deep learning models is also an important research direction for pursuing high-quality visual structure parsing. In Chapter 5, we devise a novel Transformer-style model, i.e., Structure-aware Transformer over Interaction Proposals (STIP). Particularly we evaluate its performance in detecting human-centric visual structure including humans, objects and their relations in images, which are the most noteworthy targets and relations in structured visual representation. STIP upgrades vanilla Transformer by additionally encoding the holistic semantic structure among interaction proposals as well as the local spatial structure of human/object within each interaction proposal, so as to strengthen predictions. Extensive experiments conducted on V-COCO and HICO-DET benchmarks have demonstrated the effectiveness of STIP, and significantly superior results are reported when compared with the state-of-the-art HOI detectors. This work has been published on CVPR′22 [3].

- We finally extend the effective Transformer-style model to tackle video scene graph generation, which has been an emerging research topic. Existing approaches predominantly follow a multi-step scheme, including frame-level object detection, relation recognition and temporal association. Although effective,

[1] ***Yong Zhang**, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. "Boosting Scene Graph Generation with Visual Relation Saliency." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) (2022).*

[2] ***Yong Zhang**, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. "Learning to Generate Language-supervised and Open-vocabulary Scene Graph using Pre-trained Visual-Semantic Space." CVPR 2023.*

[3] ***Yong Zhang**, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. "Exploring Structure-aware Transformer over Interaction Proposals for Human-Object Interaction Detection." CVPR 2022.*

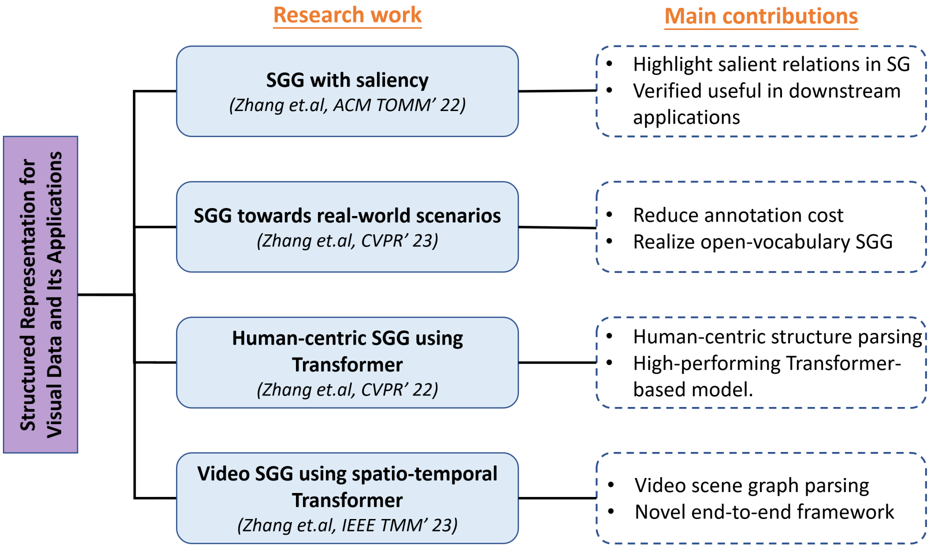

Research work
Main contributions
Structured Representation for Visual Data and Its Applications
SGG with saliency
(Zhang et.al, ACM TOMM' 22)
Highlight salient relations in SG
Verified useful in downstream applications
SGG towards real-world scenarios
(Zhang et.al, CVPR' 23)
Reduce annotation cost
Realize open-vocabulary SGG
Human-centric SGG using Transformer
(Zhang et.al, CVPR' 22)
Human-centric structure parsing
High-performing Transformer-based model.
Video SGG using spatio-temporal Transformer
(Zhang et.al, IEEE TMM' 23)
Video scene graph parsing
Novel end-to-end framework

Figure 1.1: An overview of our research works and main contributions in this thesis.

these approaches neglect the mutual interactions between independent steps, resulting in a sub-optimal solution. In Chapter 6, we tackle the challenging video SGG task by proposing a novel end-to-end Transformer-based architecture. Such design jointly models the three sub-tasks in video SGG (i.e., object detection, instance temporal association and relation recognition) through a monolithic spatio-temporal Transformer, making it fully exploit the temporal dynamics and consistency across video frames to strengthen relation recognition. Extensive experiments conducted on VidHOI and Action Genome benchmarks demonstrate the superior performance of the proposed TDT over the state-of-the-art methods. This work has been submitted to IEEE TMM (2023) [4].

We have summarized the above-mentioned research works and our main contributions in Figure 1.1, which constitute the main content of this thesis.

1.3 Thesis Overview

In the rest of this thesis, we begin with a comprehensive survey in Chapter 2, which includes a review of visual representations, the deep learning technical background, rich literature on scene graph generation and applications, and a summarization of key challenges in previous works that motivate our contributions toward addressing those challenges. After that, Chapters 3-6 present a collection of approaches we developed for parsing visual data into structured representations, also including our efforts to exploit parsed structured representations to facilitate downstream tasks. Finally, in Chapter 7, we conclude this thesis with remarks on open problems and future research opportunities.

[4] ***Yong Zhang**, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. "End-to-End Video Scene Graph Generation with Temporal Propagation Transformer.", IEEE Transactions on Multimedia (TMM), 2023.*

Chapter 2

Background and Related Work

In this chapter, we briefly introduce the technical background when we conduct the research in Section 2.1, especially about deep learning techniques. Next, in Section 2.2, we review different ways of visual representation. In the following Section 2.3-2.5, we review the literature in detail for scene graph generation, and its applications and challenges. A comprehensive literature review of existing works on scene graphs is summarized in Table 2.1.

2.1 Technical Background

Since the success of AlexNet [19] for ImageNet classification in 2012, deep learning has enjoyed immense popularity in visual data understanding. Briefly speaking, deep learning models jointly learn representations of the raw input data and a predictive model for a particular task. These models are generally implemented by stacking multiple 'layers' of differentiable non-linear transformations and trained to fit large-scale labeled data using gradient descent techniques. Their rising popularity comes from their incredible effectiveness in tackling various computer vision tasks, from image classification [19, 20], object detection [21, 22, 23, 24, 25, 26], segmentation [27, 28, 29], to high-level tasks like image captioning [30, 31, 32, 33, 34, 35], visual question answering [14, 17, 33, 36, 37, 38] etc. Such models have not only achieved superior performance than traditional machine learning techniques, and are even surpassing human brain performances in some specific computer vision tasks. One famous example is that Microsoft beat humans in the image classification task [20] as early as 2016. It is no exaggeration to say that deep learning has become the most essential tool for computer vision research.

Technically, the advancement of deep learning research comes from many aspects: the design of model architecture, the training objective, the optimization procedure, the scale of training data, and even the way how training data are presented to the model (e.g., data augmentation) can have a profound impact for the model's generalization ability. Specifically, in terms of network architectures, the family of convolutional neural networks (CNN) is widely employed in the computer vision community. The parameters for convolutional operations in CNN are shared across image locations, which exploits the fact that visual features are usually transition invariant. ResNet [39] is one of the most famous CNN, which innovates the effective design of skip connections. More recently, the community has shown great interest to incorporate the Transformer [40] architecture to tackle various vision tasks, which mainly takes advantage of attention mechanisms to gradually learn context-aware visual representations.

Though more sophisticated deep learning techniques are being developed and great achievements have been obtained, there is still a long way to achieve the goal of making intelligent agents understand the visual world like humans. To this end, here are some promising directions that have attracted the great interest of researchers but are still in their infancy. Firstly, understanding a visual scene goes far beyond merely recognizing individual objects in isolation. Hence, researchers take a step further to examine the detailed structure of a visual scene, especially in the form of understanding object relationships [41, 42, 43]. Secondly, the success in visual perception tasks (e.g., recognition, detection) using deep learning models has surged great interest in pushing forward visual cognition. This requires the perceived visual information to be integrated with human knowledge for reasoning and decision-making [18]. In this thesis, our basic idea is to employ powerful deep learning techniques to push the limit of structured visual understanding and application.

2.2 Visual Representations

"An image is worth a thousand words". Visual data contain rich and versatile information, and how to describe/represent the embedded semantics has always attracted people's interest, either for communication purposes or automatic analysis in computer systems. We briefly review existing ways for visual representation and illustrate *why using scene graph representation?*

Language description. For humans, nothing is more straightforward than describing visual data with natural language. One way is to tag an image or video with some keywords to describe its content, which is useful for content-based indexing and retrieval systems. However, to comprehensively describe a visual scene, we usually need a very long paragraph or many tags. One limitation of using language description is that the co-references (e.g., whether two mentions are coreferent) between language and vision concepts or inside a sentence are hard to align. Moreover, language descriptions rely on human labor for annotating. Due to these limitations, it is infeasible to represent oceans of visual data with human language descriptions in computer systems.

Hand-crafted features. Representing visual data as hand-crafted features has a long research history. It aims to enable computers to analyze visual data automatically instead of relying on human efforts. In computer vision, researchers have developed numerous methods to represent images/videos from various aspects, such as feature points (e.g., HOG [44], SIFT [45]), edges, shapes, colors, textures, motions (e.g., optical flow) [46]. In addition, the "bag of words“ (BoW) [47] model is popular to describe a visual scene. BoW regards an image/video as a "text document" that contains multiple "visual words", then the distribution histogram of these words is used as the representation. These classic methods lie in the low-level vision domain and are still far from describing visual semantics. This means the computer system still does not understand how pixels are organized to objects, not to mention the arrangement or relationships of objects.

Learned representation. In the current deep learning era, the paradigm of learning representation from raw pixel inputs becomes popular. Generally, a deep neural network takes raw images/videos as inputs. The first layers in the network are considered as the feature extractor, while the following layers are considered as the classifier for a particular task, e.g., image classification. This paradigm is validated by visualizing the intermediate outputs of the trained network. For example, as shown in [48], the CNN trained for classifying images has actually learned to firstly detect low-level features such as oriented edges, and then gradually combine them to more complex shapes, and finally recognize category concepts. Therefore, a network is usually trained on a pretext task, then its intermediate outputs are used as features for other tasks. Here, the pretext task can be supervised, e.g., the backbone part of the network trained for ImageNet classification is applied to extract image features for object detection or image retrieval. It can also be unsupervised, e.g., He et. al. in [49] propose Momentum Contrast (MoCo) framework for unsupervised visual representation learning. Experiments of MoCo suggest that the gap between unsupervised and supervised representation learning has been largely closed in many vision tasks. Recently, pre-training on large-scale datasets using the Transformer architecture has shown very promising results for learning visual representation [50, 51, 52]. Despite the effectiveness, a common concern is that deep networks are still black boxes, hence the learned representations (e.g., feature vectors or feature maps) have implicit meaning and are not interpretable and controllable for humans.

Scene graph representation. Scene graph representation is a promising way to depict visual semantics explicitly for images [4, 53], videos [54] or 3D data [55, 56]. It abstracts object entities in a visual scene as graph nodes, and relations between objects as graph edges connecting nodes. In this way, scene graphs depict not only what entities exist in a scene, but also how these entities are organized or interacted – the structure of the visual scene. Compared with the aforementioned visual representations, scene graphs have various advantages: 1) scene graphs naturally align with human visual perception; 2) scene graphs can serve as a shared semantic structure to bridge multimodal and multi-source information; 3) extracted knowledge in scene graphs can be easily inspected and interpreted.

2.3 Scene Graph Generation

The task of scene graph generation (SGG) aims to generate a scene graph representation from input visual data. We mainly focus on SGG from an image input, since it is the foundation. A more comprehensive literature summarization is shown in Table 2.1. Here, we formally formulate this task, and summarize existing SGG methods by categorizing them into two groups same as in [117]: the bottom-up methods and the top-down methods.

Task formulation. For image SGG, the input is a raw image, and the outputs are objects and their relationships that compose a target scene graph. Formally, denote the image as I and target scene graph as $G = \{O, R\}$, where $O = \{o_1, ..., o_M\}$ represents the set of object entities and $R = \{r_1, ..., r_N\}$ is the relationship set, then the SGG task it to predict $P(G|I) = P(O, R|I)$. Specifically, $o_i = (< x_i^1, y_i^1, x_i^2, y_i^2 >, l_i) \in O$ represent the object's bounding box coordinates and its category label; $r_j =< sub_j, pred_j, obj_j > \in R$ stands for a ⟨*subject-predicate-object*⟩ relation triplet, where $sub_j, obj_j \in O$ and $pred_j$ is the predicate label, which can be positional (e.g., 'on', 'in'), interaction (e.g., 'riding', 'sit on'), possessive (e.g., 'of', 'has') etc. Video or 3D SGG tasks are extensions of image SGG with different forms of input visual data, e.g., video frames and 3D point clouds. The output scene graph may also require adaptations. For example, video scene graphs can be represented as a sequence of temporally evolving frame-level scene graph [54].

Bottom-up SGG methods. As shown in Figure 2.1 (a), bottom-up SGG methods sequentially perform object detection and visual relationship detection, i.e., $P(SG = O, R|I) = P(O|I) \cdot P(R|I, O)$, then individual results of two modules are composed to the scene graph. For the *object detection* stage, it aims to predict both location and category information for appeared objects. Deep learning models have revolutionized this task by developing various efficient object detectors, which can be roughly categorized into two types: two-stage approach and one-stage approach. The former performs localization and category classification sequentially, while the latter predicts the location and category in a single CNN network. The R-CNN family (i.e., R-CNN [26],

Input	Datasets	SGG Methods	Challenges	Applications
Image	• Visual Genome (VG) [53] • VG150 [42] • VG-KR [43]	• Bottom-up methods [57, 58, 59, 60, 61] • Top-down methods [42, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77]	• Biased/long-tailed distribution [78, 79, 80, 81, 82, 83, 84] • Expensive annotation cost [85, 86, 87] • Saliency [43, 88, 89, 90, 91, 92] • Close-set vocabulary [93, 94]	• Image retrieval [4, 6, 7, 8] • Image captioning [9, 10, 11, 12, 13] • Visual question answering [14, 15, 16, 17] • Image generation or manipulation [95, 96, 97, 98] • Cross-media knowledge graph [18, 99]
Video	• Action Genome [54] • VidHOI [100]	• Trajectory-based [100, 101] • Spatio-temporal contextualization [102, 103, 104, 104, 105] • End-to-end [106]	• Biased/long-tailed distribution [107, 108] • Reasoning over temporally evolving scene graphs	• Video question answering [109] • Action recognition [54, 110] • Anomaly detection [111] • Robot navigation or planning [112]
3D data	• 3DSSG [55]	• E.g., 3DSSG [55], SceneGraphFusion [113]	• Efficient models [56, 114]	• 3D scene generation or manipulation [115] • E.g., operation room modeling [116]

Table 2.1: A comprehensive survey on scene graph, including typical datasets, models, challenges and applications for SGG from three kinds of visual data, i.e., image, video, 3D data (such as multi-view images, point clouds).

SPPNet [118], Fast-RCNN [119], Faster-RCNN [21]) is probably the most famous two-stage approach, which employs a CNN for feature extraction, then generates several regions of interest (RoI) proposals with selective search or Region Proposal Network (RPN). Finally, features of each RoI proposal are extracted for category classification. In contrast, one-stage object detection approaches target for end-to-end model training and testing. YOLO [25], object detection is formulated as a single regression problem that directly predicts the bounding boxes and categories from full images. SSD [23] is another efficient one-stage object detection architecture, which conveys the concept of end-to-end regression in YOLO and anchor mechanism in the Faster R-CNN. The major challenge for one-stage is the extreme foreground-background class imbalance, which can be alleviated by hard negative sampling used by YOLO and SSD, or using focal loss [120] to down-weight well-classified examples and enhance the importance of misclassified ones. The performance of object detection is affected by several factors, such as backbone networks and input size, suitable design of one-stage architecture would have comparable performance with the complicated two-stage networks. However, as reported in much of the literature, one-stage approaches exhibit a bit lower mean average precision (mAP) on small objects [121]. More recently, Transformer-based object detectors [22, 24] are also developed for effective end-to-end object detection. In SGG, most existing works adopt off-the-shelf detectors and focus on the key challenge of reasoning the visual relationship.

For the *visual relationship detection* (VRD) stage, the goal is to find out a set of relationship triplets, describing where and what are the subject and object entities, and how they interact, i.e., recognizing predicates. It originates from a pioneering task - visual phrase detection [122], e.g., detecting "person riding a bike", of which the output is the whole image region described by the phrase without locating the "person" or "bike". VRD may also be independently studied as a single task [57, 58, 59, 60, 61, 88, 123, 124]. Lu et.al [57] propose to recognize subject, object and predicates individually, then compose results to a relationship triplet. This paradigm liberates visual relationship understanding from only concentrating on a handful of relationships, as the number of possible combinations of $\langle$*subject-predicate-object*$\rangle$ triplet can be highly diverse. Li et.al [58] show that visual phrase detection can efficiently guide relationship detection. Especially, they present a Phrase-guided Message Passing Structure (PMPS) framework over CNN, in order to refine features of relationship components and finally enhance the evidence for prediction. The relationship proposal network is presented in [59], to get rid of classifying each pair of detected object entities. As the appearance of the same predicate can vary significantly, instead of learning an appearance-based model for each predicate, the predicate in [125] is novelly modeled as a visual attention shift operation from the one entity to the other. Some other VRD works focus on building comprehensive feature representations for visual relations. For example, DR-Net [60] integrates a variety of cues for VRD, such as appearance, spatial configurations, as well as the statistical relations between objects and relationship predicates. Particularly, it is found that incorporating statistical relation modeling can substantially improve relationship prediction. Semantic features from language prior are further incorporated for relationship representation in [57, 61]. Especially, RelDN [61] proposes to use contrastive loss, enabling the model to disambiguate related and unrelated visual entities explicitly. We have also observed that there are some works addressing particular challenges in VRD like extending to large-scale relationship understanding [93], incomplete relationship annotation [123, 124], discovering salient relationships [88, 89]. The human-object interaction detection [126, 127, 128, 129, 130, 131, 132] task is a special case of VRD. It also attracts particular interest and has stimulated the advancement in relationship understanding, since humans are usually the focus of visual understanding.

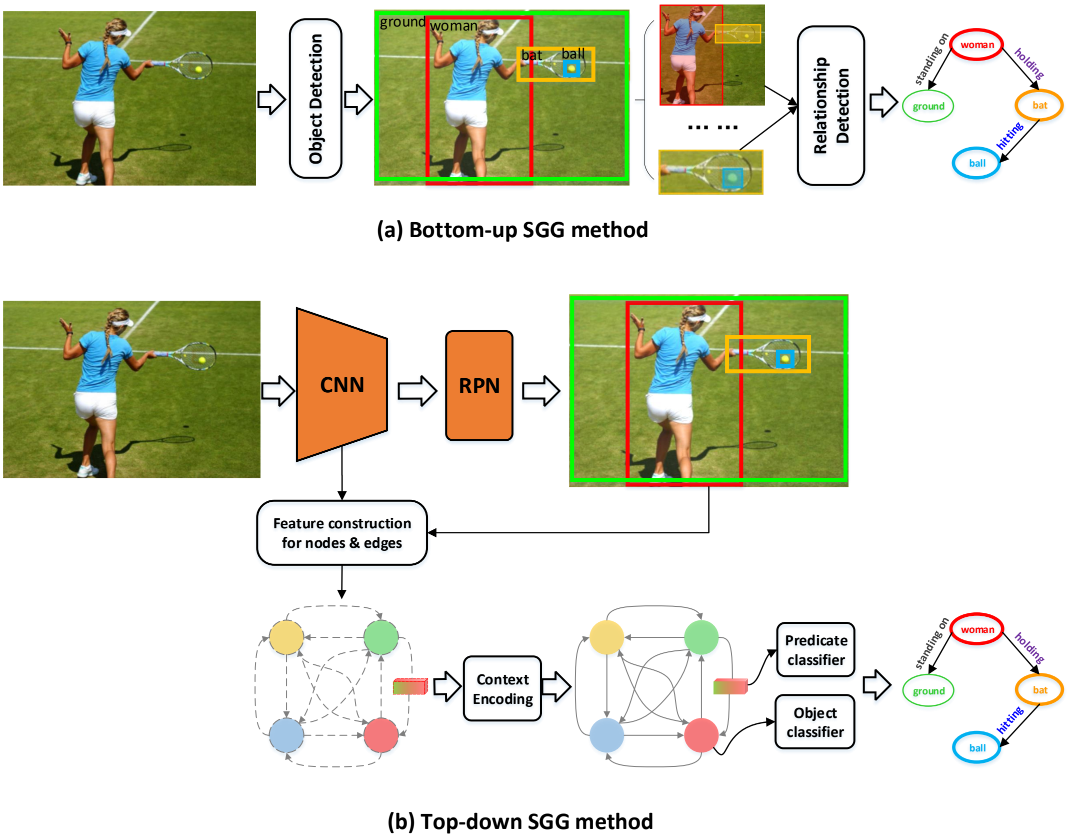


Figure 2.1: We categorize existing methods for scene graph generation (SGG) into two types: the bottom-up methods sequentially detect image objects and relations; while the top-down SGG methods regard all object entities and relationships as a holistic structure, and predict them together. (RPN: region proposal network.)

Top-down SGG methods. Top-down SGG methods regard all object entities and relationships as a holistic structure, and predict them together, i.e., $P(SG = O, R|I)$. Most methods of this group exploit the contextual information of the whole scene to enhance predictions for object and relation labels (usually using an RPN network to obtain bounding boxes at first), as shown in Figure 2.1 (b). For example, IMP [42] is one of the early attempts to jointly reason objects and relationships as a whole, employing a message propagation mechanism over graph topology to refine local features with context information. Zellers et.al [62] leverage bidirectional LSTM to implement message propagation, which can still effectively incorporate context information to boost SGG, even if the graph structure is simplified to a flat chain during message passing.

This work also finds the importance of using frequency statistics, which significantly improves SGG performance. The frequency statistics are further leveraged by [133] to regularize message propagation. In [43, 79, 134], researchers propose to organize objects into hierarchical trees, then employ TreeLSTM to produce context-ware node and edge features. More works [63, 64, 65, 66, 67, 68, 69, 70, 71] adopt graph neural network (GNN) for SGG. Especially, Yang et.al [63] propose to prune the densely connected graph to a sparse one with a relationship proposal network, avoiding predicting all pairwise relationships. Li et.al [66] cluster object pairs referring to similar interacting regions to sub-graphs, over which message propagation is performed for efficient contextualization. In [67, 68, 70], the authors take more graph properties for consideration, such as node priority, edge directions, graph density etc. In addition, researchers also employ CNN [73, 135] or Transformer [74, 77] to directly generate scene graphs from raw pixels in a fully end-to-end fashion. Some other works focus on building more comprehensive relationship features, incorporating global context [72] and appearance from the entities' intersection box [136], deriving better loss function [61], introducing multi-task learning [137], and importing external knowledge to SGG [138, 139, 140, 141], to improve the robustness of scene understanding and generate more semantically plausible scene graphs.

2.4 Scene Graph Applications

We summarize the applications of scene graphs mainly into three categories: 1) facilitating downstream vision-language tasks; 2) visual content generation/manipulation; 3) bridging cross-media data.

Firstly, scene graph representation has facilitated numerous high-level vision-language tasks. Compared with existing works that exploit a global visual feature or a set of region features [33], a scene graph has the advantage of holistically capturing the inner structure of a scene (e.g., semantic and spatial relationships between objects). We list several typical applications as follows:

- **Image-text retrieval.** This task aims to retrieve a matched image with a text query, or vice versa. Using scene graphs for content-based image retrieval is firstly seen in [4, 142, 143]. Rather than querying with unstructured text, scene graphs parsed from texts that could better express the semantics of what we are searching, are exploited for retrieval. The retrieval problem is transformed as grounding a scene graph on an image, which can be modeled with Conditional Random Fields (CRF) as in [4]. Other works [6, 7, 8] represent both image and text into scene-graph-like representations, and formulate image retrieval as a scene graph matching problem. Another way of leveraging visual structure for retrieval is to enhance image features by considering how image regions are related as in [144].

- **Image captioning.** This task aims to predict a descriptive sentence for an input image. Some pioneer works validate that introducing fine-grained visual structures (e.g., relations, attributes) can effectively boost image captioning [32, 145]. Yang et.al [9] incorporate language inductive bias into visual scene graphs by sharing an encoding dictionary with the text scene graph, in order to boost performance for image captioning. Given scene graphs, [10] can extract both object and relationship features as input for caption generating. In [43], Wang et.al show that some key relationship triplets contribute mostly to the resulting caption. Gu et.al [11] present a framework to train an image captioning model in an unsupervised manner without using any paired image-sentence data, by using scene graphs as an intermediate to align vision and language. Chen et.al [12] utilize scene graph as a control signal to generate controllable and diverse captions. Milewski et.al in [13] have made a thorough empirical study on whether using scene graphs can lead to better captioning performance. The conclusion is yes but it depends on the quality of the scene graph, while current automatically generated scene graphs are still too noisy to achieve significant improvement. More interestingly, Wang et.al [146] have demonstrated a scenario of applying to generate a coherent story from a sequence of images, i.e., visual storytelling.

- **Visual question answering.** This task aims to produce an answer to a question according to a given image. Previous approaches for visual question answering rely on implicit image features and black-box deep networks. DIfferently, scene graph representation is promising to achieve interpretable visual reasoning, and therefore it has been very attractive [14, 15, 16, 17]. Especially, Zhang et.al [14] naturally employ GNN on scene graphs to encode image structure information, and empirically demonstrate its potential to outperform state-of-the-art models but with a cleaner architecture. Shi et.al [16] parse a question to an executable program, and its execution, i.e., the reasoning procedure, is simulated as attention transitions on scene graphs. Similarly, Hudson et.al [15] treat the extracted scene graph as a state machine, and translate the natural language question as a series of implicit instructions according to which reasoning is seemed as traversing on states. These works have shown the great potential of symbolical scene graphs to enable explicit and explainable visual reasoning, and are also good practices for implementing neural-symbolic AI that exploits both the merits of deep neural networks and classic symbolical reasoning [147, 148].

Secondly, in reverse, scene graphs are used for visual content generation [95, 115] or manipulation [96]. Compared with other visual representations such as language description or feature vectors, scene graphs provide a more explicit description of the visual scene, hence may provide clear instructions for accurately generating the desired visual data. Also, such a structured representation is easy to edit, allowing easy user interaction. This makes the generation process controllable and can facilitate easy semantically manipulating of visual content. More recently, Dhamo et.al in [115] employ scene graphs as interfaces for semantically synthesizing and manipulating 3D scenes, which is promising to help the work of designers through automatically generated intermediate results.

Thirdly, since the graph structure is a very general way to represent knowledge, scene graphs can naturally serve as the bridge to connect vision with language or knowledge graphs. Actually, in natural language processing, scene graphs can be parsed from language descriptions. This can be achieved by explicitly encoding the objects, attributes and relationships found in texts, and abstracting away most of the lexical and syntactic idiosyncrasies. Such technique is well studied in [142, 149]. And the generated text scene graph can be aligned and merged with the visual scene graph, resulting in a cross-media scene graph [18]. Furthermore, scene graphs are linked with knowledge graphs by entities in [99], which is essential for multimedia knowledge extraction [18, 150] and recommender systems [151].

In addition, video scene graphs in the form of a sequence of temporally-evolving graphs for depicting dynamic scenes, have been shown promising in applications such as video action recognition [54, 110, 152], anomaly detection [111, 153] and robot navigation [112].

2.5 Challenges

For the SGG task, there are still various problems that limit the generation quality. We summarize the challenges as follows:

- **Biased/long-tailed distribution.** Existing SGG methods bias to predict frequent but usually uninformative predicates, such as "on" and "of". This issue comes from the extremely unbalanced relationship category distribution among current SGG datasets (e.g., Visual Genome [53]). However, as pointed out by Tang et.al [78], we should not entirely blame the imperfect data collection. In contrast, we should admit that human description of the real world is biased. Hence, the problem becomes: how to generate unbiased scene graphs with biased training data? In [78], Tang et.al introduce counterfactual causality theory to cut off the effect of non-visual factors for de-biasing. Alternatively, researchers in [79, 80, 81, 154] derive special loss functions to balance different categories. For example, Yan et.al [154] innovates loss re-weighting schemes according to semantic correlations between predicates. Li et.al [65] employ resampling for de-bias. He et.al [155] propose a head-to-tail knowledge transfer module to preserve rich knowledge learned from the head into the tail to deal with long-tail category distribution.

- **Incomplete annotation.** Some object pairs have obvious relationships but are not annotated in available SGG datasets such as Visual Genome [53, 54]. This issue could be mitigated by improving annotation quality, but annotating relations for every pair of objects is too demanding for crowdsourcing workers. Wan et.al [156] modify translation-based models in knowledge graphs scene graph completion. Chen et.al [124] propose to generate missing relationship labels from a model trained on a small labeled dataset. Wang et.al [157] tackle the unannotated problem by capitalizing on self-learned knowledge. And Yao et.al in [158] propose visual distant supervision, which generates relation labels from a commonsense knowledge base.

- **Expensive annotation.** Scene graphs are also very expensive to annotate. To address this issue, weakly supervised SGG [85, 86, 87] is proposed by parsing a text scene graph from language description as the supervision. Unlike human-annotated scene graphs, scene graphs parsed from language description are not grounded to image regions, and they may have more serious incompleteness problems. The biggest advantage is low-cost to collect large-scale training data, e.g., by crawling from the web or utilizing image captioning datasets.

- **Saliency.** Almost all existing SGG methods treat all visual relationships as equal, ignoring the fact that only a fraction of these relationships attracts more human attention. Hence, researchers in [43, 88, 89] contend that we should highlight salient/key relations that sketch the main content of images. Especially, [43, 89] have constructed datasets to support this new trend of research.

- **Closed-set vocabulary.** Almost all existing SGG methods [42, 62, 65, 75, 76, 78, 134, 159, 160] involve a pre-defined closed set of object and relationship categories, making them limited in recognizing novel objects outside of training corpora. The most widely-adopted VG-150 dataset [42] for SGG is a processed version of Visual Genome [53], which covers 150 object categories and 50 predicate categories. Obviously, the categories are limited and in a closed set, which means the resulting scene graph might not generalize well for downstream tasks like image captioning which involves a much larger vocabulary. [93] is one of the few attempts for addressing this issue, which embeds visual entities and language concepts into a shared continuous space to improve generalization.

Though many methods are proposed to mitigate the above issues, they are not capable of generating structured visual representations for visual data with practical accuracy and efficiency. This motivates our research in this thesis. Furthermore, in terms of scene graph applications, what are the "killer applications" or other application scenarios of scene graphs? How to effectively leverage scene graphs in these applications? These questions also need further research efforts.

Chapter 3

Scene Graph Generation with Saliency

In this chapter[1], we introduce visual relation saliency to align scene graph representation with human visual perception. Also, we have proposed a novel architecture for generating scene graphs with relation saliency.

As we know, a scene graph is a symbolic data structure that comprehensively describes the objects and visual relations in a visual scene, while ignoring the inherent perceptual saliency of each visual relation (i.e., relation saliency). However, humans often quickly allocate attention to important/salient visual relations in a scene. To align with such human perception of a scene, we explicitly model the perceptual saliency of visual relations in a scene graph by upgrading each graph edge (i.e., visual relation) with an attribute of relation saliency. We present a new design, named Saliency-guided Message Passing (SMP), that boosts the generation of such scene graph structure with guidance from the visual relation saliency. Technically, an object interaction encoder is first utilized to strengthen object relation representations by jointly exploiting the appearance, semantic, and spatial relations in between. A branch is further leveraged to estimate the relation saliency of each visual relation by ordinal regression. Next, conditioned on the object and relation features (coupled with the estimated relation saliency), our SMP enhances scene graph generation by performing message passing over the objects and the most salient relations. Extensive experiments on VG-KR and VG150 datasets demonstrate the superiority of SMP for scene graph generation. Moreover, we empirically validate the compelling generalizability of the learned scene graphs via SMP on downstream tasks like cross-modal retrieval and image captioning.

3.1 Introduction

Visual perception of the real-world scene is one fundamental capability of human intelligence. To formalize such kind of scene understanding, a symbolic data structure named scene graph [4] is abstracted from natural images via Scene Graph Generation (SGG) techniques, which describes the object instances in a scene and the visual rela-

[1] ***Yong Zhang**, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. "Boosting Scene Graph Generation with Visual Relation Saliency." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) (2022).*

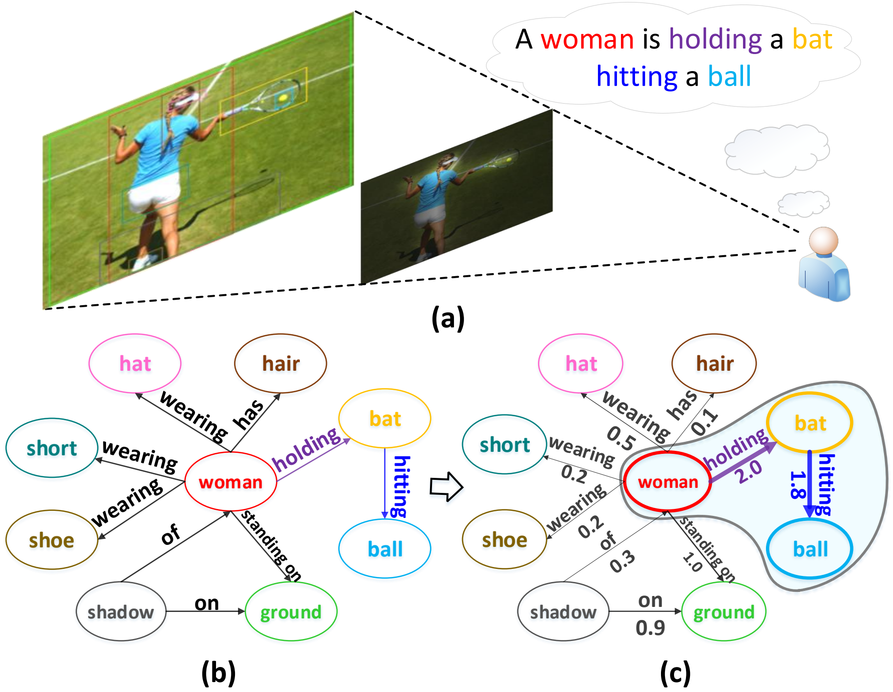


Figure 3.1: An illustration of human perception of a visual scene (a), where humans often allocate attention to the salient visual relations that are worthy of mention in a natural-language utterance. The typical scene graph (b) fails to identify such salient relations, while the scene graph with key relations (c) better aligns with human perception by upgrading each edge with an attribute of relation saliency.

tions (predicates) in between. It is well recognized that reasoning over scene graph is crucial to a richer semantic understanding of the visual scene in an image and benefits a wide range of vision-language downstream tasks, e.g., visual question answering [37], cross-modal retrieval [4], and image captioning [137, 145]. Though promising results are reported on these downstream tasks, the typical design of scene graph aims to capture all visual relations between objects as completely as possible, and thus inevitably fails to focus on the most important/salient visual relations. This seems contradictory to the human perception of a scene, where humans often quickly allocate attention to the visually salient objects/visual relations in an image. Take the image in Figure 3.1 as an example, the generated typical scene graph (Figure 3.1(b)) is unable to identify the salient visual relations (e.g., ⟨*woman-holding-bat*⟩ and ⟨*bat-hitting-ball*⟩), which highlight the key gists that are worthy of mention in a natural-language utterance (Figure 3.1(a)) by humans. As a result, in analogy to salient object detection [161, 162], the specific task of SGG with key relations is introduced recently in [43], along with a well-established benchmark VG-KR, which offers a fertile ground for key/salient relation detection over scene graph. Here we explicitly define the perceptual saliency of each visual relation in a scene graph as the weight of each graph edge (i.e., relation saliency), leading to a weighted directed graph (Figure 3.1(c)). The ultimate goal of this task is to jointly generate the scene graph and predict the relation saliency for each visual relation existing in the image.

The difficulty of SGG with key relations originates from two aspects: 1) how to accurately detect each object and the visual relations in between for scene graph generation? 2) how to discriminate salient visual relations on scene graphs? In the literature, there have been several techniques, including scene graph generation [42, 62, 134] and salient relation detection [88, 89], being proposed for each individual aspect. Nevertheless, simply solving the problem of SGG with key relations via two separate branches as in [43] (i.e., one for scene graph generation and the other for relation ranking according to relation saliency) may destroy the interaction between visual relation detection and relation saliency estimation, resulting in a sub-optimal solution.

In this work, we propose to mitigate this issue by unifying both relation saliency estimation and visual relation detection for SGG with key relations. Technically, we devise a novel Saliency-guided Message Passing (SMP) architecture to facilitate scene graph generation with additional guidance from the estimated visual relation saliency. Specifically, Faster R-CNN is first leveraged to produce a set of image regions that depict the detected objects. After that, for each pair of objects, we capitalize on an object interaction encoder to enrich the relation representation with the object relation information mined from appearance, semantic, and spatial perspectives. Each enriched relation representation is further fed into the relation saliency estimation branch for estimating the corresponding visual relation saliency via ordinal regression. After filtering out the less salient relations, message passing is executed on the objects and the salient visual relations to iteratively improve the predictions of objects and their visual relations, and thus boost scene graph generation with key relations. The whole SMP framework is jointly optimized in an end-to-end fashion.

3.2 Related Work

Scene Graph Generation. With the prevalence of deep learning and the availability of large-scale image datasets annotated with scene graphs (Visual Genome [53]), a series of innovations have been proposed to generate scene graphs from images [42, 61, 62, 63, 66, 67, 68, 78, 133, 134, 136]. Unlike the visual relationship detection task [57] which performs isolated relationship classification, Scene Graph Generation (SGG) holistically associates all objects and their relationships in a scene, and exploits the contextual information in between to strengthen the predictions of objects and relationships. Specifically, [42] is one of the early attempts to propagate messages between the bipartite node and edge sub-graphs for SGG. [62] composes a sequence of all objects in a scene and then leverages a bidirectional LSTM to propagate the messages along a flat chain structure. [63, 134] further predict whether relationships exist between object pairs to construct a tree or sparse graph structure for message propagation. Moreover, instead of holistically propagating messages between nodes, [66] performs message propagation over sub-graphs that are generated by clustering object pairs containing similar interactions. In [133], the statistical co-occurrence knowledge of objects are additionally exploited to aggregate the messages from different neighbors. Recently, [65] performs adaptive message propagation using its estimated relationship confidence in order to reduce the noise in context modeling. Several other works introduce multi-task learning [137], external knowledge base [139] or casual inference [78] to enhance scene graph generation.

Salient Relation Detection. The traditional scene graph aims to recognize all visual relations in a scene and treat them equally, which is somewhat contradictory to the human visual perception that commonly focuses on the most salient visual relations at first glance. Recently, the research community starts to pay more attention to the detection of salient visual relations, which can benefit several downstream tasks (e.g., cross-modal retrieval) that require a holistic understanding of key gists in an image. In particular, [88] differentiates visual relations by importance from human perception, and presents an attention-based mechanism for emphasizing the salient relations in existing visual relation detection benchmarks. Furthermore, [43, 89] clearly define the task of salient relation detection and construct datasets (i.e, VG-KR and ViROI) with salient relation annotations to support this new task.

Summary. In short, our approach is also a type of message passing method for scene graph generation, but focuses on the latter challenging task of SGG with relations in multiple saliency levels. HetH-RRM proposed in [43] is perhaps the most related work, which utilizes two separate branches to tackle this task: the structured context encoding branch that employs TreeLSTM over a hierarchical tree of objects for scene graph generation, and the relation ranking branch to prioritize key relations. Compared to HetH-RRM, our SMP differs in multiple ways: 1) we utilize a more detailed design for enriching relation representations with contextual information mined from three different perspectives (i.e., appearance, semantic, and spatial); 2) we integrate both scene graph generation and relation saliency estimation into a unified architecture, which naturally triggers the interaction in between and thus boosts SGG with key relations.

3.3 Approach

3.3.1 Notation and Overview

To comprehensively describe the objects and visual relations in an image, the typical scene graph $G = (O, R)$ is commonly constructed by treating all objects $O = \{o_1, ..., o_N\}$ as graph nodes and the pairwise object relations $R = \{r_1, ..., r_M\}$ as graph edges. Each object $o_i \in O$ is represented as $o_i = (l_i, b_i)$, where l_i denotes the corresponding class label and $b_i \in \mathbb{R}^4$ is the spatial location of bounding box. Each object relation $r_m \in R$ is a $\langle$*subject-predicate-object*$\rangle$ triplet, and we represent it as $r_m = (o_i, p_{ij}, o_j)$, where p_{ij} is the label of predicate (i.e., visual relation) that characters the visual interaction between o_i and o_j. Considering that the typical scene graph fails to model the inherent perceptual saliency of each visual relation, we upgrade each graph edge with an attribute of relation saliency score s_{ij} to better align with the human perception of a scene. Accordingly, each graph edge is represented as $r_m = (o_i, p_{ij}, s_{ij}, o_j)$. The underlying assumption is that the higher the relation saliency score, the more important the visual relation in the visual scene from the human perception perspective.

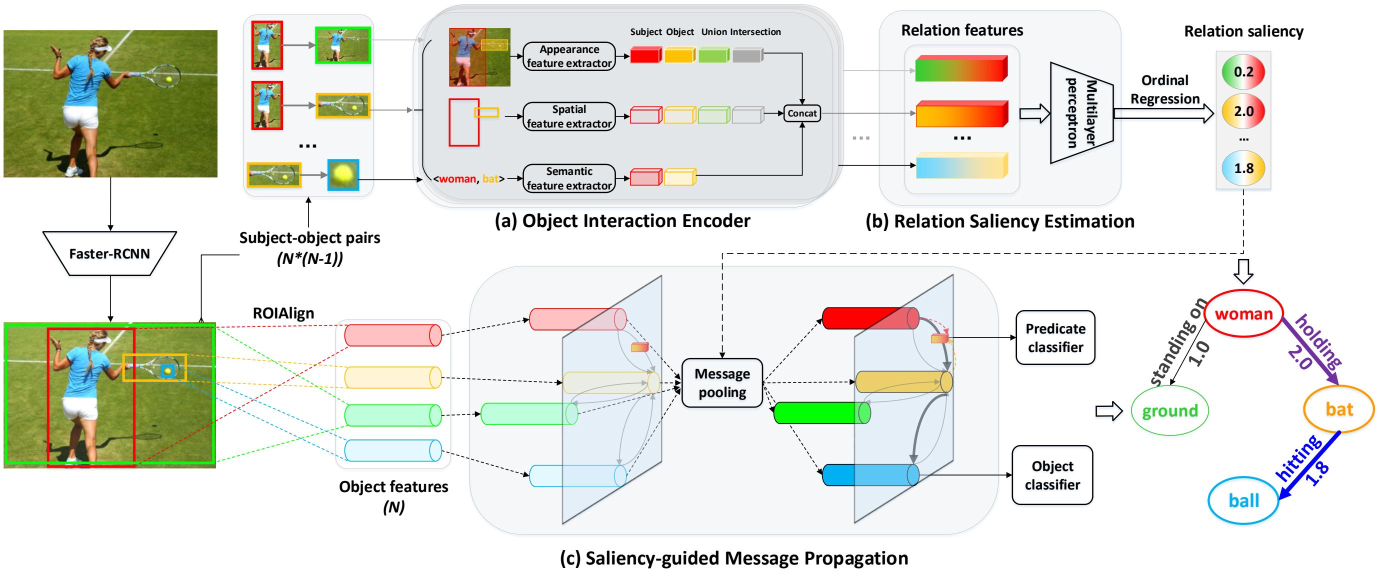


Figure 3.2: Overview of our Saliency-guided Message Passing (SMP) for scene graph generation with key relations. Faster R-CNN is first utilized to detect a set of image regions. Next, the Object Interaction Encoder (a) is employed to fully construct relationship representations from three different perspectives (i.e., appearance, spatial, and semantic). A Relation Saliency Estimation branch (b) is then leveraged to model the saliency level of each visual relation via ordinal regression. Conditioned on the object and relation features (coupled with the estimated relation saliency), Saliency Guided Message Propagation (c) is performed to iteratively improve the predictions of objects and the most salient visual relations through message passing. The final output scene graph is further equipped with the estimated relation saliency.

In this work, we present a Saliency-guided Message Passing architecture (SMP) to generate such kind of scene graph with key/salient relations. SMP first interprets the whole image as a set of image regions via an object detector (e.g. Faster R-CNN [21]). Next, in Section 3.3.2, we present an object interaction encoder to contextually learn the relation representation by exploiting the appearance, semantic, and spatial relations between objects. In Section 3.3.3, a relation saliency estimation branch is further devised to model the saliency level of each relation by formulating it as an ordinal regression problem. Finally, Section 3.3.4 details the process of saliency-guided message propagation in our SMP that boosts scene graph generation. An overview of our scene graph generation architecture is illustrated in Figure 3.2.

3.3.2 Relation Representation Learning

Given N regions detected in an input image, we construct $N \times (N-1)$ possible subject-object pairs. For each possible subject-object pair, we need to extract its relation representation via an object interaction encoder. Recent advances in visual relation detection [42, 62, 64, 78, 136] have demonstrated that modeling the contextual information between objects or exploiting the union/intersection box in between does enhance relation representation learning for scene graph generation. Our work takes a step forward and constructs a more detailed design of object interaction encoder, that fully mines the relation contexts from both union and intersection boxes in three different perspectives (i.e., appearance, semantic, and spatial), for relation representation learning. Concretely, in the object interaction encoder, three different feature extractors (i.e., appearance, spatial, and semantic feature extractors) are utilized to learn relation representation based from each perspective.

Appearance Feature Extractor. Given the inputs of subject region, object region, their union box, and intersection box, we capitalize on *ROIAlign* [163] to extract the appearance feature map (size: $7 \times 7 \times D$, $D = 512 / 256$ in different backbones) for each region. Next, two fully-connected (FC) layers with ReLU activations are leveraged to transform each appearance feature into a low-dimensional vector. Please note that for the special case of no intersection of two objects, a zero vector is used to represent the appearance feature from the intersection region. We concatenate all the four vectors to compose the output appearance relation feature.

Spatial Feature Extractor. Similarly, we exploit the spatial contextual information from the subject region, object region, their union box, and intersection box. For each region, we measure the spatial feature $\mathbf{b} \in \mathbb{R}^9$ based on the bounding box coordinates (x_1, y_1, x_2, y_2):

$$\mathbf{b} = (\frac{c_x}{W}, \frac{c_y}{H}, \frac{w}{W}, \frac{h}{H}, \frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \frac{wh}{WH}), \tag{3.1}$$

where $(c_x, c_y) = (\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2})$ is the center, $(w, h) = (x_2 - x_1, y_2 - y_1)$ denotes the width and height of region, and (W, H) is the width and height of the whole image. Please note that a 9-dimensional zero vector is used as the spatial feature for an empty intersection box. Then, we employ a FC layer to map the spatial feature into a high-dimensional vector. The final spatial relation feature is thus achieved by concatenating all the four spatial features.

Semantic Feature Extractor. Here we adopt an embedding layer to encode the label names of subject and object: $\mathbf{z} = \mathbf{W}_e\mathbf{l} \in \mathbb{R}^{300}$, where $\mathbf{W}_e$ is the word embedding initialized from Glove [164] and $\mathbf{l}$ is a one-hot vector over label vocabulary. The concatenation of two encoded name features is taken as the semantic relation feature.

Finally, we obtain the relation representation by concatenating the appearance, spatial and semantic relation features.

3.3.3 Relation Saliency Estimation

Recall that our target is to recognize the visually salient relationships in an image, which naturally aligns with human perception of a scene. Hence, we involve an additional relation saliency estimation branch to predict the relation saliency for each visual relation. Considering that the relation saliency is annotated on a multiple point ordinal scale (e.g., 0-no relation, 1-perceptually less important relation, 2-salient relation in VG-KR dataset), we naturally formulate the relation saliency estimation task as Ordinal Regression (OR) problem.

Inspired by [165, 166, 167], here we tackle the OR problem by decomposing it into multiple binary classification sub-problems for relation saliency estimation. Formally, given the input relation feature $\mathbf{u}_i$ of visual relation r_i, we aim to predict the corresponding relation saliency S_{r_i} over S different levels $\{0, 1, ..., S-1\}$ ($S = 3$ in VG-KR dataset). Such process can be transformed into $S-1$ simpler binary classification subtasks, and each subtask targets for discriminating whether $S_{r_i} > k$, where $k \in \{0, 1, ..., S-2\}$. The relation saliency estimation branch is implemented as a multi-layer perception, which takes $\mathbf{u}_i$ as input and produces $S-1$ output units via Sigmoid activations. The output value of the k-th unit is denoted as $\hat{y}_k = P(S_{r_i} > k)$, which reflects the probability of $S_{r_i} > k$. For each relation feature $\mathbf{u}_i$ coupled with its ground-truth saliency level s_{r_i}, the objective of relation saliency estimation is thus measured as:

$$L_{saliency}(\mathbf{u}_i, s_{r_i}) = \sum_{k=0}^{S-2} BCELoss(\hat{y}_k, \mathbf{1}\{s_{r_i} > k\}) + \sum_{k=0}^{S-3} max(\hat{y}_{k+1} - \hat{y}_k, 0), \tag{3.2}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, and $BCELoss(\cdot)$ denotes binary cross entropy loss. The first term aggregates the classification losses for all subtasks, and the second term is involved to suppress the inconsistency of saliency estimation as in [168] (e.g., the unreasonable case of $P(S_{r_i} > 1) > P(S_{r_i} > 0)$). At inference, we calculate the expected relation saliency for an unseen sample r' as the aggregation of $S-1$ output probabilities:

$$\mathbf{E}[S_{r'}] = \sum_{k=0}^{S-2} P(S_{r'} > k) = \sum_{k=0}^{S-2} \hat{y}_k. \tag{3.3}$$

3.3.4 Saliency-guided Message Propagation

One natural way for scene graph generation is to explore contextual information between objects and all the connected edges (i.e., relations) via message passing [42] to make better predictions on objects and predicates. Nevertheless, such way not only leads to a rise in computational cost, but also involves more unremarkable relations and the overall stability of message passing will be inevitably affected. To alleviate this issue, we devise the saliency-guided message propagation module that triggers the interactions between relation saliency estimation and scene graph generation by facilitating message passing under the guidance of visual relation saliency. The spirit behind follows the philosophy that this module filters out the unremarkable relations, and performs message propagation over the objects and the most salient relations, which iteratively strengthens the predictions of objects and relations, especially for the salient ones.

Technically, let $\mathbf{v}_i$ and $\mathbf{e}_{ij}$ denote the feature of graph node i and the feature of graph edge $i \rightarrow j$, where $i \rightarrow j$ represents the directed edge from node i to node j and the corresponding relation saliency value is s_{ij}. Similar to [42], we iteratively perform message propagation, and two Gated Recurrent Units (GRUs) are employed to memorize and refine the node and edge representations. Note that $\mathbf{v}_i$ and $\mathbf{e}_{ij}$ are initially set as the region feature via *ROIAlign* and the output relation feature through object interaction encoder.

Each message propagation iteration consists of two steps: message pooling and feature refining. Message pooling aims to aggregate multiple incoming messages from connected object neighbors. In our design, for each node, we only exploit the messages from top K ($K = 4$) connected neighbors with the most salient relations (ranked by the expected relation saliency). The message for node i (i.e., $\mathbf{m}_i$) is computed as the concatenation of messages from outbound and inbound nodes, followed by a mapping function ϕ_v:

$$\mathbf{m}_i = \phi_v([\sum_{j:s_{ij}\in top\ K} f_{out}(\mathbf{v}_i, \mathbf{v}_j)\mathbf{v}_j, \sum_{j:s_{ji}\in top\ K} f_{in}(\mathbf{v}_i, \mathbf{v}_j)\mathbf{v}_j]), \tag{3.4}$$

where f_{out} and f_{in} denote the functions for learning attention weights in message aggregation. We implement f_{out}/f_{in} as a FC layer with Softmax normalization, and ϕ_v as a FC layer with LeakyReLU non-linearity activation. The message for edge $i \to j$ (i.e., $\mathbf{m}_{ij}$) is obtained by concatenating the relation features of node i and j:

$$\mathbf{m}_{ij} = \phi_e([\mathbf{v}_i, \mathbf{v}_j]), \tag{3.5}$$

where ϕ_e is the mapping function implemented by a FC layer with LeakyReLU non-linearity activation.

The feature refining step further enriches each node or edge representation with the contextual information derived from the corresponding pooled messages. Thus, the node and edge representations are refined as:

$$\begin{aligned} \mathbf{v}_i &= GRU_v(\mathbf{v}_i, \mathbf{m}_i) \\ \mathbf{e}_{ij} &= GRU_e(\mathbf{e}_{ij}, \mathbf{m}_{ij}), \end{aligned} \tag{3.6}$$

where GRU_v and GRU_e denote two GRUs for recording the refined node and edge representations respectively.

Finally, after T ($T = 2$ as in [42]) iterations of message propagation, we obtain the strengthened context-aware node and edge representations, which are leveraged for predicting objects and predicates via two classifiers.

3.3.5 Training and Inference

At the training stage, based on the strengthened graph node representation $\mathbf{v}_i$, we predict its object label via the object classifier $f_{obj} : \mathbb{R}^{d_v} \to \mathbb{R}^{C_{obj}}$, where d_v is the dimension of $\mathbf{v}_i$ and C_{obj} is the object class number. In practice, the object classifier is implemented as a 2-layer MLP. By denoting the ground-truth object label as y_i^{obj}, the object classification loss is thus measured as:

$$L_{object} = \frac{1}{N_{obj}} \sum_{i}^{N_{obj}} CrossEntropy(f_{obj}(\mathbf{v}_i), y_i^{obj}), \tag{3.7}$$

where N_{obj} is the number of objects/graph nodes, and $CrossEntropy(\cdot)$ denotes the cross-entropy loss function.

Similarly, based on graph edge representation $\mathbf{e}_{ij}$, the predicate classification loss is calculated as:

$$L_{relation} = \frac{1}{N_{rel}} \sum_{i,j}^{N_{rel}} CrossEntropy(f_{rel}(\mathbf{e}_{ij}), y_{ij}^{rel}), \tag{3.8}$$

where N_{rel} is the number of relations/graph edges, f_{rel} is the predicate classifier (also implemented as a 2-layer MLP), and y_{ij}^{rel} is the ground-truth predicate label.

Together with relation saliency estimation objective in Eq.(3.2), the overall objective of our SMP is:

$$L_{SMP} = L_{saliency} + L_{object} + L_{relation}. \tag{3.9}$$

During the inference of relation ranking, we compute the ranking score for a predicted relation triplet $r' = (s, p, o)$ as:

$$C_{r'} = P(s) \cdot P(o) \cdot P(p) \cdot P(r' \, is \, salient), \tag{3.10}$$

where $P(s), P(o), P(p)$ are the predicted probabilities of subject, object and predicate, and $P(r' \, is \, salient) = P(S_{r'} > 1)$ in VG-KR for example.

3.4 Experiments

3.4.1 Datasets and Experimental Settings

Datasets. VG-KR [43] is a recently constructed dataset for the challenging task of SGG with key relations. It goes beyond the typical Visual Genome [53] dataset and contains the annotations of relation saliency. The dataset consists of 26,992 images belonging to 200 object categories and 80 predicate categories. There are 250,755 predicates in total, out of which 101,312 predicates are annotated as key/salient relations. We follow the same split in [136] to compose the training, validation, and testing sets.

VG150 [42] dataset is a widely adopted subset of Visual Genome for evaluating conventional SGG task. It contains 10,8073 images, and covers 150 object categories & 50 predicate categories. There are totally 622,705 annotated instances for visual relations without differentiating saliency. We follow the same split in [42, 43, 62, 65, 78] for evaluating our SMP.

Evaluation. Following [43], we evaluate our SMP for the task of SGG with key relations under two protocols: predicate classification (**PREDCLS**) and scene graph classification (**SGCLS**). For PREDCLS, given the object categories and their bounding boxes, the target is to predict the relationships in between. For SGCLS, given the bounding boxes, it aims to infer both object categories and relationships to compose a scene graph. Key Relation Recalls among top K (K=1,5) predicted relationship triplets (i.e., kR@K) are adopted as the metrics for evaluating the ability of recognizing salient relationships, which are computed under two different match rules: **Triplet Match**

	Triplet Match				Tuple Match			
	SGCLS		PREDCLS		SGCLS		PREDCLS	
Model	kR@1	kR@5	kR@1	kR@5	kR@1	kR@5	kR@1	kR@5
VCTree-SL† [134]	5.7	14.2	11.4	30.2	8.4	22.2	16.1	46.4
MOTIFS† [62]	5.9	14.5	11.3	30.0	8.5	21.8	16.0	46.2
MOTIFS* [62]	5.9	14.7	11.5	30.3	8.4	22.3	16.1	46.5
MOTIFS-TDE* [78]	1.2	3.1	4.8	13.5	4.5	12.5	8.6	24.8
HetH† [43]	6.1	15.1	11.6	30.4	8.6	22.7	16.4	47.1
MOTIFS-RRM† [43]	8.6	16.4	16.7	33.8	13.8	26.3	27.9	57.1
HetH-RRM† [43]	9.2	17.1	17.5	35.0	14.6	27.3	28.9	59.1
SMP	10.2	18.2	19.8	**38.4**	16.4	29.7	33.3	**66.6**
SMP (ResNeXt-101-FPN)	**11.5**	**20.3**	**20.4**	38.1	**18.3**	**33.3**	**34.5**	**66.6**

Table 3.1: Performance comparison for scene graph generation on VG-KR dataset (Default backbone: VGG-16). † indicates the results are referred from [43], and * denotes our re-implemented results.

(all of subject, predicate, and object should be same as the ground truth relationship) and **Tuple Match** (only subject and object are consistent with the ground truth). For conventional SGG task, we additionally include the evaluation under the scene graph detection (**SGDET**) protocol, which assumes both object boxes and categories are unknown to generate a scene graph. All protocols adopt Relation Recall among top K (K=50,100) predictions under the triplet match rule as evaluation metrics (i.e., R@K).

Implementation Details. For fair comparison with [43] in terms of SGG with key relations, we adopt the same pre-trained Faster R-CNN (backbone: VGG-16), and we also report the performances of our SMP based on Faster R-CNN with a stronger backbone of ResNeXt-101-FPN [169]. For conventional SGG, we use the same pre-trained Faster R-CNN (backbone: ResNeXt-101-FPN) on VG150 as in [65, 78]. The whole architecture of our SMP model is trained by SGD optimizer over a single Nvidia 2080ti GPU. The batch size and the initial learning rate is set as 6 and 0.006, respectively. The learning rate will be decayed by 10 for two times after validation plateaus. The max training iteration number is set as 50K. Note that we solely optimize the relation saliency estimation branch at the first 5K iterations for a better initialization. After that, all the three modules of our SMP are jointly trained in an end-to-end manner.

3.4.2 Experiments on Scene Graph Generation

Performance Comparison On VG-KR. Table 3.1 shows the performances of different techniques for scene graph generation with key relations task on the VG-KR dataset. Overall, the results with regard to all evaluation metrics consistently indicate that our SMP achieves superior performances against other state-of-the-art approaches. The baselines include both traditional SGG models (MOTIFS [62], VCTree-SL [134]) and the relation-saliency-based methods (HetH-RRM and MOTIFS-RRM [43]) particularly designed for SGG with key relations. Specifically, by integrating the SGG architecture with another branch for relation ranking, HetH-RRM and MOTIFS-RRM

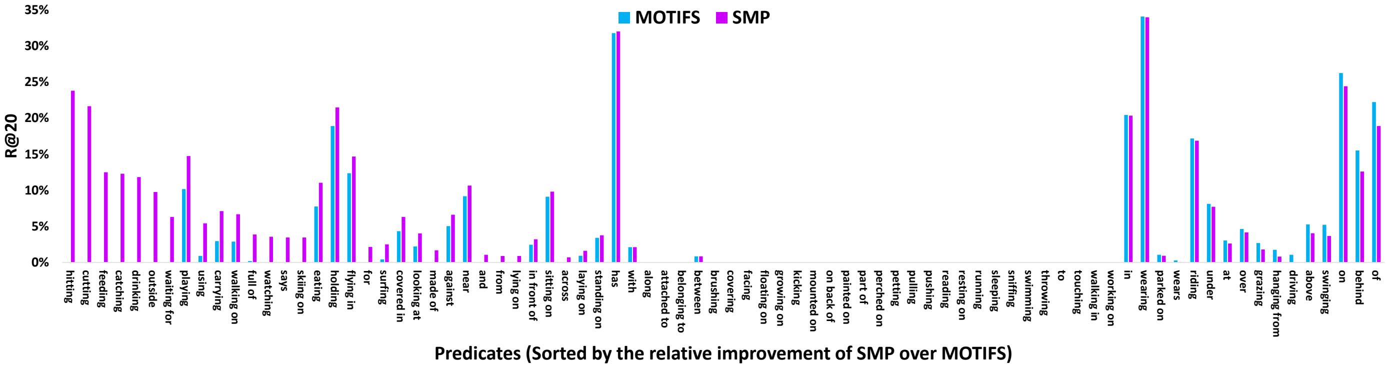


Figure 3.3: Performance comparison for each predicate category under SGCLS on VG-KR.

outperform the traditional SGG models (MOTIFS and VCTree-SL), which basically verifies the effectiveness of exploiting relation saliency in the specific task of SGG with key relations. However, the performances of HetH-RRM are lower than our SMP that triggers the interaction between relation saliency estimation and scene graph generation in a unified architecture for enhancing the contextual-aware relation features and eventually boosting SGG with key relations. In addition, by leveraging Faster R-CNN with a stronger backbone, the performances of SMP are further boosted up.

In addition, we compare our SMP with the baseline (MOTIFS) with respect to recall rates for each predicate category under SGCLS. Figure 3.3 shows the recall rates among top 20 (R@20) for each predicate, where the predicates are sorted by the relative performance improvements of SMP against MOTIFS. As shown in this figure, our SMP manages to recognize the visual relations with “verb” predicates (e.g., “hitting”, “feeding”, “cutting”), which tend to be the salient relationships in VG-KR. Instead, the MOTIFS model fails to recognize these visual relations ($R@20 = 0\%$), but prefers to detect some high-frequent predicates (e.g., “of”, “behind”, “on”).

Performance Comparison on VG150. Our SMP can be easily generalized for the traditional SGG task by directly classifying visual relations into two saliency levels (i.e., no relation and salient relations). In this way, the relation saliency estimation module is utilized to predict whether there exists a relationship between objects. Table 3.2 shows the performance comparison against state-of-the-art methods on VG150. Note that for fair comparison, all results are obtained by using the same pre-trained Faster R-CNN detector and codebase [170]. As shown in this table, our SMP manages to achieve competitive performances against state-of-the-arts under all evaluation protocols. The results basically demonstrate the advantage of our saliency-guided message propagation mechanism for alleviating noise effect in context modeling.

3.4.3 Experimental Analysis

Ablation Study. Here we conduct ablation study to examine how each design in our SMP influences the overall performance. Table 3.3 shows the results by considering one more design for SGG with key relations in SMP. We start from a base model (**Base**)

	PREDCLS		SGCLS		SGDET	
Model	R@50	R@100	R@50	R@100	R@50	R@100
MOTIFS-TDE [78]	47.2	51.6	25.4	27.9	19.4	23.2
BGNN† [65]	59.2	61.3	37.4	38.5	31.0	35.8
IMP* [42]	61.1	63.1	37.5	38.5	25.9	31.2
MSDN† [137]	64.6	66.6	38.4	39.8	31.9	36.6
RelDN† [61]	64.8	66.7	38.1	39.3	31.4	35.9
GPS-Net† [68]	65.2	67.1	39.2	37.8	31.1	35.9
G-RCNN† [63]	65.4	67.2	38.5	37.0	29.7	32.8
VCTree† [134]	65.5	67.4	38.9	39.8	31.8	36.1
MOTIFS† [62]	66.0	67.9	39.1	39.9	32.1	**36.9**
SMP	**66.3**	**68.0**	**39.9**	**40.7**	**32.6**	**36.9**

Table 3.2: Performance comparison for scene graph generation on VG150 dataset. † indicates that the results are referred from [65], and * denotes our re-implemented results. All results are based on the same pre-trained Faster-RCNN (backbone: ResNeXt-101-FPN) and codebase [170] for fair comparison.

which is a degraded version of our SMP with two separate branches: one for relation saliency estimation via ordinal regression and the other for SGG through message passing. Note that this ablated Base model solely exploits the contextual information from appearance perspective to represent each object relation. After that, by additionally modeling the Contextual information from Semantic and Spatial perspectives (**CSS**), Base+CSS strengthens the relation representation for relation saliency estimation, and thus leads to a performance boost. Finally, by further guiding the process of message passing with the estimated relation saliency, our SMP reaches the highest performance. This verifies the advantage of unifying both relation saliency estimation and visual relation detection for SGG with key relations.

Effect of K for Message Propagation. We vary the number of selected connected neighbors (K) from 0 to 7 for performing message propagation (Eq. (3.4)), aiming to explore the relationship between the performance and the neighbor number K. We show such performance comparisons under SGCLS protocol (kR@5) in Figure 3.4. In the extreme case of $K = 0$, no connected neighbor is selected to perform message propagation, and the model degenerates to basic object and predicate classifiers that produce predictions based on isolated object/relation features. Next, enlarging the neighbor number K generally increases the performances, which basically demonstrates the effectiveness of performing message propagation among the objects and the top-K salient relations. Specifically, the best performance is achieved when K is 4. Furthermore, when K increases more than 4, the performances begin to drop. We speculate that this may be the result of noise effect in context modeling when selecting more neighbors.

3.4.4 Evaluation on Downstream Applications

As an extractor of scene graph with key relations, here we follow recent vision-language pre-training techniques [35, 171, 172, 173, 174] and test the generalization ability of

	Triplet Match		Tuple Match	
Model	kR@1	kR@5	kR@1	kR@5
Base	9.7	17.5	15.8	29.2
Base+CSS	9.9	17.8	16.2	29.3
SMP	**10.2**	**18.2**	**16.4**	**29.7**

Table 3.3: Ablation study on VG-KR under SGCLS protocol. **Base**: a degraded version of SMP with two separate branches for relation saliency estimation and SGG, which solely exploits the contextual information from the appearance perspective. **CSS**: the exploration of contextual information from semantic and spatial perspectives.

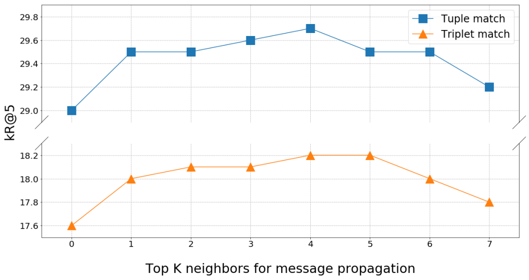
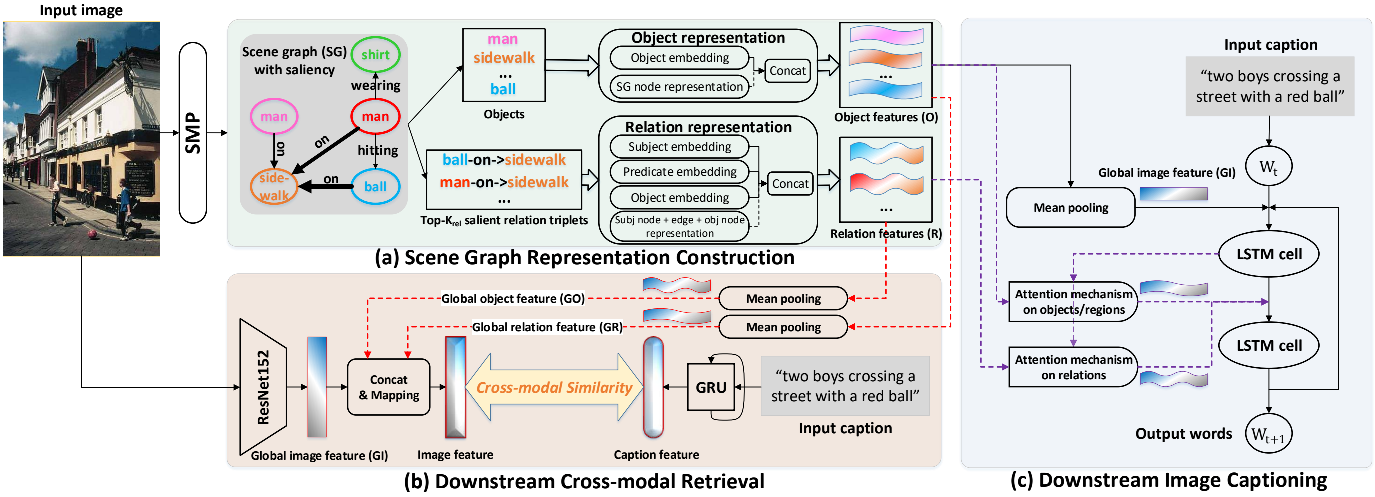


Figure 3.4: Performance comparisons under SGCLS protocol (kR@5) by using different numbers of selected top connected neighbors (K) for message propagation.



(a) Scene Graph Representation Construction

(b) Downstream Cross-modal Retrieval

(c) Downstream Image Captioning

Figure 3.5: Overview of employing learned scene graphs via our SMP for downstream tasks. (a) Scene graph representation is constructed by encoding objects and top-K_{rel} salient relation triplets. Then, object and relationship representations of the scene graph are integrated into typical architectures for cross-modal retrieval (b) and image captioning (c). (Dotted lines indicate optional inputs.)

learned scene graphs via our SMP on two downstream applications/tasks, i.e., cross-modal retrieval [175, 176, 177] and image captioning [32, 34, 178, 179, 180, 181], as shown in Figure 3.5. Experiments of both tasks are conducted on Flickr30K dataset [182] with the commonly adopted split [183] for evaluation, which assigns 29k/1k/1k images for train/val/test. Each image in Flickr30K is annotated with 5 captions.

Experiments on Cross-modal Retrieval. For each image, we select the top-K_{rel} predicted salient relations (top-2 or top-5), and perform mean pooling over all individual representations of the selected relations to achieve the global relation feature. We describe this procedure in detail as follows:

- By denoting a relation triplet as $r = (sbj, pred, obj)$, where sbj, $pred$ and obj represent subject, predicate and object labels respectively, we encode each label of subject/predicate/object as a one-hot vector. For example, for subject label sbj, its one-hot vector is constructed as $\mathbf{s}_{oh} = [0, ..., 1, ..., 0] \in \{0, 1\}^{C_{sbj}}$, where only the sbj^{th} element that corresponds to the subject label sbj is set as 1 and all other elements are set as 0. C_{sbj} denotes the size of subject vocabulary, i.e., the number of all subject categories in the dataset. Similarly, the one-hot vectors for object $\mathbf{o}_{oh}$ and predicate $\mathbf{p}_{oh}$ are constructed based on the object and predicate vocabulary.

- Next, we utilize embedding layers to separately encode the one-hot vectors of subject, predicate and object labels in a relation triplet. Taking the input one-hot vector $\mathbf{s}_{oh}$ of subject as an example, we achieve the subject label embedding vector $\mathbf{s}_e = \mathbf{E}_{sbj}\mathbf{s}_{oh} \in \mathbb{R}^{d_{sbj}}$, where $\mathbf{E}_{sbj} \in \mathbb{R}^{d_{sbj} \times C_{sbj}}$ denotes the embedding layer. Similarly, we obtain the object label embedding $\mathbf{o}_e$ and predicate label embedding $\mathbf{p}_e$. After that, we take the concatenation of all the three label embeddings, i.e., $\mathbf{r}_e = cat[\mathbf{s}_e, \mathbf{p}_e, \mathbf{o}_e]$, as the representation for relation triplet r.

- Finally, we perform mean pooling operation over the relation triplet representations of all selected top-K_{rel} relations $\{\mathbf{r}_{e,1}, ..., \mathbf{r}_{e,K_{rel}}\}$ in an image, leading to the global relation feature, i.e., $\mathbf{r}_{global} = \frac{1}{K_{rel}} \sum_k \mathbf{r}_{e,k}$.

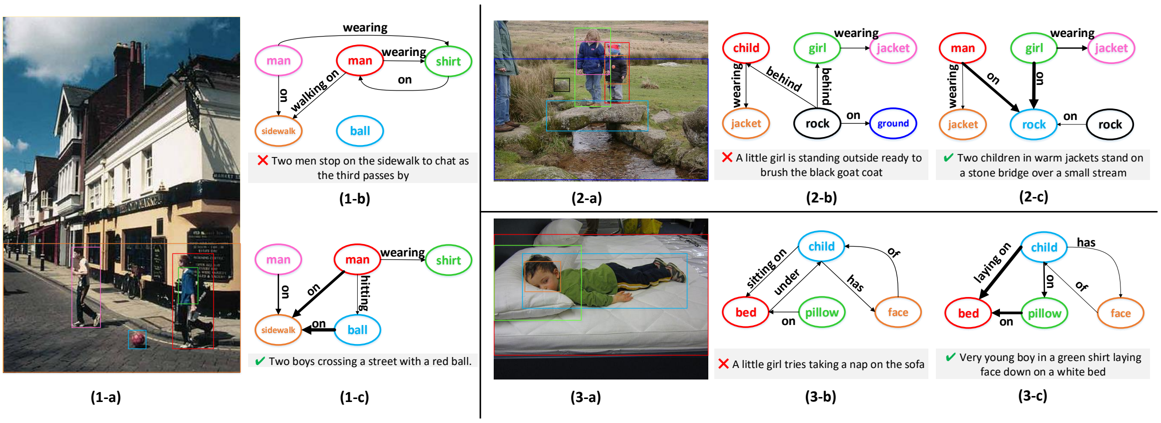
Similarly, the mean-pooled representation of all the label embeddings of object entities in this image is taken as the global object feature. The holistic image representation is thus obtained by concatenating the global image feature (backbone: ResNet152 [39]), the global object feature, and the global relation feature. To extract the representation of each sentence, we utilize the same GRU-based text encoder as in [184]. Then, we exploit VSE++ [184] to learn the visual-semantic embedding for cross-modal retrieval. During the evaluation stage, we conduct image-to-caption retrieval and caption-to-image retrieval, and report Recall@1-10 for each kind of retrieval. An overview of integrating scene graph for cross-modal retrieval is depicted in Figure 3.5-(b).

Table 3.4 details the performances of different image representations for cross-modal retrieval. Note that we additionally involve several runs using different global relation features derived from existing SGG techniques (e.g., MOTIFS [62] and MOTIFS-TDE [78]) for performance comparison. As expected, only utilizing the global image feature is inferior to the concatenation of global image and object features. A performance boost is further attained when additionally exploiting the global relation feature. Most specifically, by delving into visual relation saliency for SGG, the global relation features from our SMP outperform the ones from typical SGG techniques. The results basically highlight the merit of learned scene graphs with salient relations on the cross-modal retrieval downstream task.

Figure 3.6 further showcases examples with the constructed scene graphs and image2caption retrieval results generated by MOTIFS [62] and our SMP. From these exemplar results, it is easy to see that the two constructed scene graphs can reflect somewhat relevant objects and visual relations in between, while our SMP can capture more salient visual relations to boost image2caption retrieval. Specifically, as shown in

Image Feature	SGG Model	Image2caption R@1/5/10	Caption2image R@1/5/10
GI		46.0 / 73.9 / 83.5	32.8 / 62.3 / 73.6
GI + GO		50.0 / 78.7 / 86.7	35.4 / 65.2 / 75.3
GI + GO + GR (top-2)	MOTIFS [62]	50.1 / 80.1 / 87.8	35.7 / 66.0 / 76.7
	MOTIFS-TDE [78]	49.8 / **80.3** / 87.6	35.3 / 66.1 / 76.8
	SMP (ours)	**51.6** / **80.3** / **88.3**	**36.6** / **67.2** / **76.9**
GI + GO + GR (top-5)	MOTIFS [62]	51.6 / 80.8 / 88.5	35.4 / 67.0 / 76.6
	MOTIFS-TDE [78]	51.9 / 80.9 / 88.2	36.6 / 66.3 / 76.7
	SMP (ours)	**54.7** / **82.4** / **89.1**	**38.1** / **67.3** / **77.3**

Table 3.4: Performance comparison with different image representations for cross-modal retrieval downstream task. **GI**: global image feature. **GO**: global object feature. **GR**: global relation feature.



✗ Two men stop on the sidewalk to chat as the third passes by

(1-b)

✓ Two boys crossing a street with a red ball.

(1-a) (1-c)

✗ A little girl is standing outside ready to brush the black goat coat

✓ Two children in warm jackets stand on a stone bridge over a small stream

(2-a) (2-b) (2-c)

✗ A little girl tries taking a nap on the sofa

✓ Very young boy in a green shirt laying face down on a white bed

(3-a) (3-b) (3-c)

Figure 3.6: Scene graph generation and image2caption retrieval results on Flickr30K. Given each input image (#-a), the scene graphs (containing top-5 relations) are produced by MOTIFS (#-b) and our SMP (#-c), coupled with the corresponding top-1 retrieved captions.

the first example of Figure 3.6 (i.e., (1-a), (1-b), (1-c)), compared to the constructed scene graph of MOTIFS that misses the key relations between "*sidewalk*"/"*man*" and "*ball*", our learned scene graph via SMP accurately recognizes the most salient relations (i.e., ⟨*man-hitting-ball*⟩ and ⟨*ball-on-sidewalk*⟩), which are correlated to the key object ("*ball*") and thus encourage the coverage of "*ball*" in the retrieved caption.

Experiments on Image Captioning. Here we adopt a basic model (Up-Down [33]) to generate sentences that describe image contents. The original Up-Down model first extracts region features via a pre-trained detector (Faster R-CNN), and further feeds them into a sentence decoder (implemented as two-layer LSTM). The first layer in this decoder takes the mean-pooled region features (i.e., global image feature) and the embedding of the previous word as inputs, and the output of the first layer plus the aggregated region features with attention mechanism are fed into the second layer. In our context, we only keep the sentence decoder and replace the input region features of the pre-trained detector with our extracted scene graph representation. Specifically,

Image Feature	SGG Model	Bleu@4	METEOR	ROUGH-L	CIDEr	SPICE
GI		21.1	17.6	43.6	37.4	12.5
GI + O		23.1	18.8	45.9	43.6	13.9
GI + R (top-2)	MOTIFS [62]	22.4	18.7	45.4	42.0	13.4
	MOTIFS-TDE [78]	22.4	18.9	45.4	42.0	13.6
	SMP (ours)	**23.0**	**19.2**	**45.9**	**44.2**	**14.0**
GI + R (top-5)	MOTIFS [62]	**23.1**	18.8	45.4	43.0	13.7
	MOTIFS-TDE [78]	**23.1**	19.0	45.5	43.2	13.8
	SMP (ours)	**23.1**	**19.2**	**46.3**	**44.8**	**14.2**
GI + O + R (top-2)	MOTIFS [62]	23.1	19.1	45.9	43.9	13.9
	MOTIFS-TDE [78]	23.2	19.2	45.9	44.1	14.0
	SMP (ours)	**23.3**	**19.4**	**46.3**	**45.3**	**14.3**
GI + O + R (top-5)	MOTIFS [62]	23.2	19.3	46.1	44.2	14.2
	MOTIFS-TDE [78]	23.2	19.2	46.0	44.3	14.3
	SMP (ours)	**23.4**	**19.5**	**46.5**	**45.6**	**14.5**

Table 3.5: Performance comparison with different image representations for image captioning downstream task. **GI**: global image feature. **O**: object features. **R**: relation features.

		Dataset for pre-training			Metrics	
Methods	Backbone	Name	Annotation	#images	Bleu@4	METEOR
Cornia et.al [34]	ResNet-50	ImageNet	Image-level labels	∼4M	21.3	20.0
Unified VLP [35]	Transformer	Conceptual Captions	Image-caption pairs	∼3M	30.1	23.0
Ours (GI + O + R (top-5))	VGG-16	VG-KR	Scene graphs	∼19K	23.4	19.5

Table 3.6: Evaluation for image captioning downstream task on Flick30K benchmark leaderboard. **GI**: global image feature. **O**: object features. **R**: relation features.

	Metrics (on "Karpathy" test split, without CIDEr optimization)				
Methods	Bleu@4	METEOR	ROUGH-L	CIDEr	SPICE
RDN [185]	36.8	27.2	56.8	115.3	20.5
Ours (GI + O)	35.8	27.7	56.5	112.8	20.8
Ours (GI + O + R (top-5))	36.5	27.7	56.7	114.6	21.0

Table 3.7: Evaluation for image captioning downstream task on MSCOCO. **GI**: global image feature. **O**: object features. **R**: relation features.

we include two baselines by feeding 1) only global image feature (i.e., GI) or 2) both global image feature plus region features (i.e., GI + O) into the sentence decoder. Next, we exploit different ways of incorporating scene graph representations into sentence decoder: 1) directly replacing the region features with top-2 or top-5 salient relation features (i.e., GI + R (top-2/5)); 2) additionally feeding top-2 or top-5 salient relation features into sentence decoder (i.e., GI + O + R (top-2/5)). Note that we take the concatenation of subject, predicate, object after context encoding, and their label embeddings as the representation for each relation. Standard metrics in image captioning (e.g., Bleu@4, METEOR, ROUGH-L, CIDEr and SPICE) are utilized for evaluation. An overview of integrating scene graphs for image captioning is illustrated in Figure 3.5-(c).

Table 3.5 shows the performance comparison with different image representations for the image captioning task. Similarly, we involve several additional runs that utilize different relation features extracted by existing SGG techniques (e.g., MOTIFS [62] and MOTIFS-TDE [78]). In general, when additionally incorporating relation features into the baseline (GI + O), the runs of GI + O + R (top-2/5) lead to consistent performance improvements across most metrics for each SGG technique. The results basically validate the merit of exploiting the relation features for image captioning. In between, the runs of GI + O + R (top-2/5) for our SMP exhibit better performances than the ones of conventional SGG techniques, which demonstrate the efficacy of our learned scene graph with salient relations. Moreover, taking a close look at the comparison between GI + O and GI + R (top-2/5) for each SGG technique, we find that the relation features learned by conventional SGG techniques result in inferior performances than the original region features. Instead, our relation features learned by SMP (GI + R (top-2/5)) manage to outperform the region features, which again confirms the promising generalization ability of learned scene graphs with salient relations. Also, we evaluate the image captioning downstream task on Flick30K benchmark leaderboard[2], as presented in Table 3.6. In this table, our run (GI + O + R (top-5)) achieves comparable results with state-of-the-art methods (e.g., Cornia et.al [34]). Note that although Unified VLP [35] achieves superior performances on Flick30K, it is somewhat unfair to compare our run with Unified VLP [35], since Unified VLP [35] adopts a large-scale pre-training dataset (∼3M image-caption pairs) and a stronger Transformer-based encoder-decoder structure.

Moreover, we additionally evaluate our run on MSCOCO dataset for image captioning downstream task. Table 3.7 presents our results on the MSCOCO leaderboard[3]. Please note that here we adopt the same pre-trained image region features as in RDN [185, 186], and exploit visual relation features learned by our SMP scene graph model. As shown in this table, although RDN additionally exploits inherent properties of language by upgrading the sentence decoder to enhance both long sequence dependency and position perception of words, our runs equipped with a basic sentence decoder still manage to achieve comparable performances. In addition, when comparing our run that exploits relation features (GI + O + R (top-5)) against our baseline run (GI + O), we observe that using additional relation features learned by our proposed SMP can boost the image captioning performance.

3.5 Conclusion

In this chapter, we present Saliency-guided Message Passing (SMP), which steers scene graph generation with guidance from the visual relation saliency. Particularly, we study the problem from the viewpoint of unifying both relation saliency estimation and visual relation detection in a single framework. To materialize our idea, we construct an object interaction encoder to contextually enhance relation representations by jointly

[2] https://paperswithcode.com/sota/image-captioning-on-flickr30k-captions-test

[3] https://paperswithcode.com/sota/image-captioning-on-coco-captions

exploiting the appearance, semantic, and spatial relations between objects. A relation saliency estimation branch is further utilized to predict the saliency of each object relation through ordinal regression. After that, we perform message passing over the objects and the most salient relations to facilitate scene graph generation. Extensive experiments conducted on VG-KR dataset demonstrate the efficacy of our SMP for the specific task of scene graph generation with key relations. The evaluations on the learned scene graph with salient relations via SMP also validate its potential of generalizing to downstream tasks.

Chapter 4

Scene Graph Generation towards Real-world Scenarios

Our work in the previous chapter addresses the saliency issue in SGG. Actually, there are two other obstacles (i.e., expensive manual annotations, and closed-set prediction) that limit existing SGG methods in practical scenarios. In this chapter[1], we push the SGG task towards real-world scenarios by taking full advantage of pre-trained visual-semantic space to address these obstacles.

We know that SGG aims to abstract an image into a graph structure, by representing objects as graph nodes and their relations as labeled edges. However, two knotty obstacles limit the practicability of current SGG methods in real-world scenarios: 1) training SGG models requires time-consuming ground-truth annotations, and 2) the closed-set object categories make the SGG models limited in their ability to recognize novel objects outside of training corpora. To address these issues, we novelly exploit a powerful pre-trained visual-semantic space (VSS) to trigger language-supervised and open-vocabulary SGG in a simple yet effective manner. Specifically, cheap scene graph supervision data can be easily obtained by parsing image language descriptions into semantic graphs. Next, the noun phrases on such semantic graphs are directly grounded over image regions through region-word alignment in the pre-trained VSS. In this way, we enable open-vocabulary object detection by performing object category name grounding with a text prompt in this VSS. On the basis of visually-grounded objects, the relation representations are naturally built for relation recognition, pursuing open-vocabulary SGG. We validate our proposed approach with extensive experiments on the Visual Genome benchmark across various SGG scenarios (i.e., supervised / language-supervised, closed-set / open-vocabulary). Consistent superior performances are achieved compared with existing methods, demonstrating the potential of exploiting pre-trained VSS for SGG in more practical scenarios.

[1] ***Yong Zhang***, *Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. "Learning to Generate Language-supervised and Open-vocabulary Scene Graph using Pre-trained Visual-Semantic Space." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.*

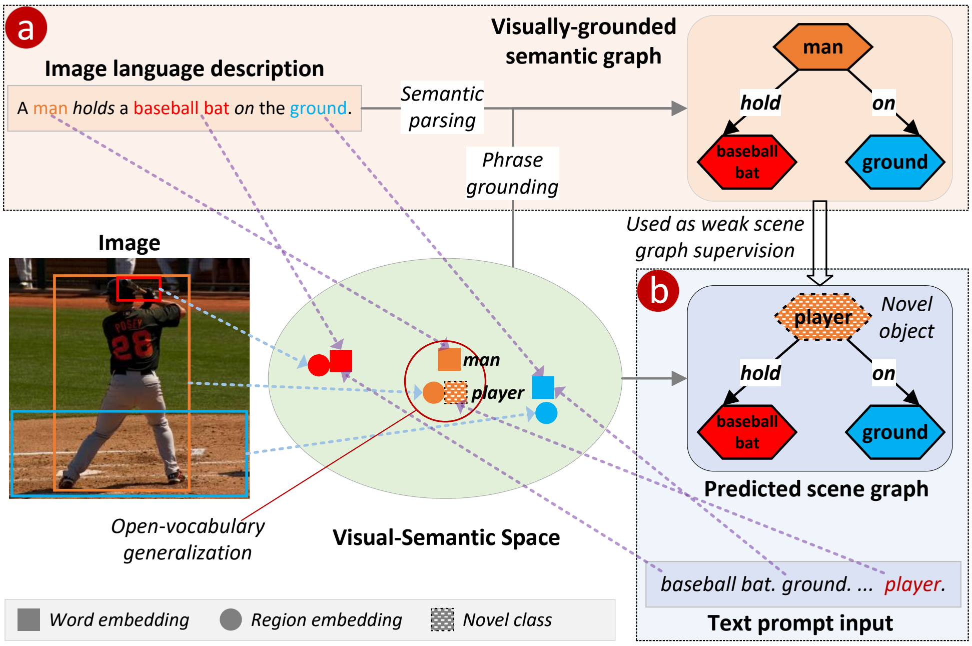


Figure 4.1: An illustration of exploiting a pre-trained visual-semantic space (VSS) to trigger language-supervised and open-vocabulary scene graph generation (SGG). (a) We acquire weak scene graph supervision by semantically parsing the image language description and grounding noun phrases on image regions via VSS. (b) At SGG inference time, thanks to the open-vocabulary generalization naturally rooted in VSS, the novel object name (e.g., player) in the text prompt input can be well aligned to one image region, which is regarded as its detection.

4.1 Introduction

Scene graph [4] is a structured representation for describing image semantics. It abstracts visual objects as graph nodes and represents their relations as labeled graph edges. The task of scene graph generation (**SGG**) [42, 53, 62, 65, 75, 76, 78, 134, 138, 139, 141, 160] plays an important role for fine-grained visual understanding, which has shown promising results in facilitating various downstream applications, such as image-text retrieval [6, 7, 8], image captioning [9, 12, 13, 31, 145], cross-media knowledge graph construction [18, 150, 151] and robot planning [112].

Though great effort has been made, SGG of the current stage still faces two knotty obstacles that limit its practicability in real-world scenarios. 1) Training SGG models requires massive ground-truth scene graphs that are expensive for manual annotation. Annotators have to draw bounding boxes for all objects in an image and connect possible interacted object pairs, and assign object/relation labels. Since assigned labels might be ambiguous, further verification and canonicalization processing are usually required [53]. Finally, a scene graph in the form of a set of $\langle subject, predicate, object\rangle$ triplets with subject and object bounding boxes is constructed. Such annotating process is time-consuming and tedious, costing much human labor and patience. 2) Almost all existing SGG methods [42, 62, 65, 75, 76, 78, 134, 159, 160] involve a pre-defined closed set of object categories, making them limited in recognizing novel objects outside of training corpora. However, real-world scenes contain a boarder set of visual concepts than any pre-defined category pool. It is very likely to encounter unseen/novel categories. When this happens, current SGG models either classify novel objects to a known category or fail to detect them like background regions. Accordingly, the prediction of their interactions/relations with other objects is negatively affected or just neglected. This may lead to problems. For example, a real-world robot may take inappropriate actions using such closed-set SGG models [94, 112].

Recently, there is a trend of leveraging free-form language supervision for benefiting visual recognition tasks via large-scale language-image pre-training [187, 188, 189, 190, 191, 192, 193]. These methods (e.g., CLIP [187]) perform pre-training on massive easily-obtained image-text pairs to learn a visual-semantic space (VSS), and have demonstrated great zero-shot transferability. Especially, the recent grounded language-image pre-training (GLIP) [189] has learned an object-level and semantic-rich VSS. Based on the learned VSS, it has established new state-of-the-art performances in phrase grounding and zero-shot object detection. This indicates such pre-trained VSS has powerful multi-modal alignment ability (i.e., image regions and text phrases that have similar semantics get close embeddings) and open-vocabulary generalization ability (i.e., covering virtually any concepts in the pre-training image-text corpus). This inspires our thought of addressing the aforementioned obstacles in SGG using the pre-trained VSS. On the one hand, taking advantage of its multi-modal alignment ability, we can cheaply acquire scene graph supervision from an image description (e.g., retrieving image regions aligned with noun phrases and re-arranging the description into a scene-graph-like form). On the other hand, by leveraging its open-vocabulary generalization ability, it is promising to enable novel category prediction in SGG.

In this work, we investigate the opportunity of fully exploiting the VSS learned by language-image pre-training to trigger language-supervised and open-vocabulary SGG. Specifically, we obtain weak scene graph supervision by semantically parsing an image language description into a semantic graph, then grounding its noun phrases over image regions through region-word alignment in the pre-trained VSS (Figure 4.1 (a)). Moreover, we propose a novel SGG model, namely Visual-Semantic Space for Scene graph generation (**VS**^{3}). It takes a raw image and a text prompt containing object category names as inputs, and projects them into the shared VSS as embeddings. Next, VS^3 performs object detection by aligning the embeddings of category names and image regions. Based on high-confidence detected objects, VS^3 builds relation representations for object pairs with a devised relation embedding module that fully mines relation patterns from visual and spatial perspectives. Finally, a relation prediction module takes relation representations to infer relation labels. The predicted scene graph is composed by combining object detections and inferred relation labels. During training, visually-grounded semantic graphs parsed from image descriptions could be used as weak scene graph supervision, achieving language-supervised SGG. At SGG inference time, when using a text prompt input containing novel categories, VS^3 manages to detect novel objects thanks to the open-vocabulary generalization ability naturally rooted in VSS, hence allowing for open-vocabulary SGG (Figure 4.1 (b)).

In summary, we have made the following contributions: (1) the exploitation of a pre-trained VSS provides an elegant solution for addressing obstacles to triggering both language-supervised and open-vocabulary SGG, making a solid step toward real-world usage of SGG. (2) The proposed VS^3 model is a new and versatile framework, which effectively transfers language-image pre-training knowledge for benefiting SGG. (3) We fully validate the effectiveness of our approach through extensive experiments on the Visual Genome benchmark, and have set new state-of-the-art performances spanning across all settings (i.e., supervised / language-supervised, closed-set / open-vocabulary).

4.2 Related Work

Fully supervised SGG. The concept of scene graph as a structured image representation is first introduced in [4]. Next, the Visual Genome benchmark [53] is manually annotated with large-scale scene graphs on images. Such annotated dataset triggers a series of innovations [42, 62, 65, 75, 76, 78, 134, 138, 139, 141, 160] for the fully supervised SGG task. Typically, an object detector (e.g., Faster-RCNN [21]) is trained to retrieve image regions as scene graph nodes. Then, relation representations of object pairs are constructed from visual, spatial and language perspectives, and are used for relation classification to label scene graph edges. To achieve desirable SGG, researchers have devised message-passing mechanisms [42, 62, 65, 75, 76, 134] to exploit contextual information, derived contrastive loss functions [61] or incorporated external knowledge [138, 139, 141]. However, all these methods rely on training with expensive scene graph annotations. Our proposed VS^3 model is compatible with fully supervised SGG, but we seek to make SGG training cheaper, which is more practical in real-world applications.

Language-supervised SGG. This task aims to train SGG models using language descriptions. It has recently attracted increasing attention [69, 85, 86, 87, 159], which is also referred as weakly supervised SGG in [69, 86, 159]. Particularly, based on a graph alignment algorithm, VSPNet [69] first proposes to supervise SGG training with scene graphs that have no object locations. Subsequent works [85, 86, 159] extract entities and relations from image captions to compose such unlocalized scene graph, which is achieved via an off-the-shelf language parser [142, 194]. They next follow a common paradigm: first grounding text entities on image regions, and then leveraging grounded scene graphs as pseudo labels to train standard SGG models. To acquire entity groundings, Shi *et al* [86] devise an efficient graph matching module optimized via contrastive learning; Zhong *et al* [85] simply match text entity names with predicted object labels from a pre-trained object detector using semantic rules such as WordNet [195] synsets matching. More recently, Li *et al* [159] integrate interaction-aware knowledge distilled from pre-trained language-image models [196] for enhancing grounding reliability. Instead, we propose to obtain groundings through region-word alignment in a pre-trained VSS, which is much simple yet more effective to collect scene graph supervision from language.

Language-image pre-training. This has been shown effective for boosting various vision-language downstream tasks [173, 196, 197, 198, 199], e.g., image-text retrieval, image captioning. Also, recent studies present remarkable results on transferring pre-trained language-image knowledge to solve vision recognition problems, such as zero-shot image classification [187, 200], open-vocabulary object detection

[188, 189, 190, 191, 192] and zero-shot semantic segmentation [193]. For example, CLIP [187] and ALIGN [200] learn separate encoders to embed image and text into a shared space by pre-training on massive image-text pairs using a contrastive loss. They have demonstrated remarkable generalization ability on zero-shot image classification after pre-training. Distinct from CLIP and ALIGN that learn image-level representations, GLIP [189] focuses on learning object-level visual representations through region-word alignment. It has attained strong zero-shot and few-shot transferability to various object-level recognition tasks such as object detection and phrase grounding. Most recently, He *et al* [201] investigate a visual-relation pre-training and prompt-based fine-tuning method for open-vocabulary SGG. However, its image encoder relies on a pre-trained region proposal extractor, which is a bottleneck for achieving open-vocabulary SGG under the SGDET protocol. Unlike [201], our proposed VS^3 directly encodes an image into region tokens, avoiding the bottleneck of region proposals. More importantly, our approach addresses obstacles to achieving both language-supervised and open-vocabulary SGG using a unified framework, while He *et al* [201] only focus on the latter.

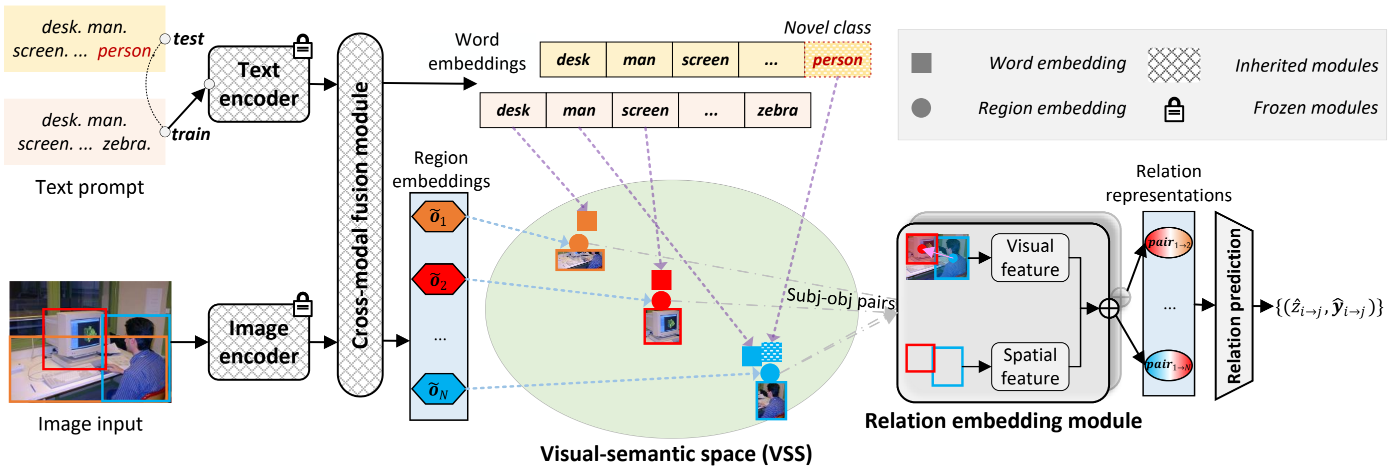


Figure 4.2: An overview of the proposed Visual-Semantic Space for Scene graph generation (VS^3) model. It inherits the image encoder, the text encoder and the cross-modal fusion module from GLIP [189], so as to project image regions and text prompt words in a pre-trained visual-semantic space (VSS). Object regions are detected by aligning the embeddings of category names and image regions in VSS. Next, high-confidence detected results are retained to compose subject-object pairs. After that, the relation embedding module constructs their relation representations by extracting visual and spatial features, on which relation prediction is performed. At test time, thanks to the open-vocabulary generalization ability of VSS, VS^3 manages to detect novel objects by switching to a text prompt containing novel classes.

4.3 Approach

4.3.1 Notation & Overview

The task of scene graph generation (SGG) aims to map an image into an abstract graph $SG = \{O, R\}$, where graph nodes $O = \{o_1, ..., o_N\}$ correspond to image objects, and graph edges $R = \{r_1, ..., r_M\}$ represents their relations. Each object $o_i = \{\mathbf{b}_i, l_i\} \in O$ contains the bounding box coordinates $\mathbf{b}_i \in \mathbb{R}^4$ and its class label information $l_i \in \mathcal{C}_o$, where $\mathcal{C}_o$ denotes the set of object categories. Each relation $r_m \in R$ is a $\langle subject, predicate, object \rangle$ triplet, and we represent it as $r_m = r_{i \to j} = \{o_i, p_{ij}, o_j\}$, in which p_{ij} is the predicate/relation label belonging to category set $\mathcal{C}_r$.

Most existing SGG methods require expensive manually annotated scene graphs SG as supervision. And they involve a closed set of object categories $\mathcal{C}_o$ in both training and inference. These issues limit SGG for practical usage. In this work, we propose to fully exploit a pre-trained VSS to push SGG towards language-supervised and open-vocabulary scenarios. We illustrate the definitions of different SGG settings in Table 4.1. Concretely, from the perspective of scene graph supervision SG, SGG is categorized into fully supervised and language-supervised, using SG from manual annotation and language respectively. From the other perspective of object categories $\mathcal{C}_o^{target}$ at inference, it is referred to as open-vocabulary or closed-set according to whether or not $\mathcal{C}_o^{target}$ contains novel objects.

$\mathcal{C}_o^{target}$ SG	**Not containing novel object classes**	**Containing novel object classes**
Manually annotated	fully supervised & closed-set	fully supervised & open-vocabulary
Automatically parsed from image descriptions	language-supervised & closed-set	language-supervised & open-vocabulary

Table 4.1: Definitions of different SGG settings, according to scene graph supervision SG and the object set at inference $\mathcal{C}_o^{target}$.

Next, in Section 4.3.2, we present a new SGG model named VS^3, which is versatile to handle with all SGG settings in Table 4.1. In Section 4.3.3, we devise a scheme to obtain scene graph supervision from language descriptions, allowing for language-supervised SGG. Finally, Section 4.3.4 details the strategy of transferring the proposed VS^3 to pursue open-vocabulary SGG.

4.3.2 The Proposed VS^3 Model

We propose the VS^3 model for tackling the SGG task by extending the GLIP [189] framework with relation recognition modules, as shown in Figure 4.2.

Preliminary. GLIP unifies object detection and phrase grounding into one framework. It has an image encoder Enc_I (e.g., Swin Transformer backbone [202]) and a text encoder Enc_L (e.g., BERT [203]). Enc_I extracts region/box features $\bar{O} \in \mathbb{R}^{\bar{N} \times d}$ from an input image, where $\bar{N}$ is the number of regions and d is the feature dimension. Enc_L encodes a text input into contextualized word/token embeddings $\bar{P} \in \mathbb{R}^{\bar{T} \times d}$, where $\bar{T}$ is the text length. Further, GLIP uses a cross-modal fusion module to achieve feature communication between $\bar{O}$ and $\bar{P}$, resulting in enriched region embeddings $\tilde{O} \in \mathbb{R}^{\bar{N}\times d}$ and word embeddings $\tilde{P} \in \mathbb{R}^{\bar{T}\times d}$. Finally, the region-word alignment scores $\hat{S}_{ground} = \tilde{O}\tilde{P}^{\top} \in \mathbb{R}^{\bar{N}\times\bar{T}}$ and the predicted locations $\hat{B} = box\_predictor(\tilde{O}) \in \mathbb{R}^{\bar{N}\times 4}$ are supervised by ground-truth text grounding data. After large-scale pre-training, Enc_I and Enc_L embed the input image and text into a joint VSS, which aligns multimodal embeddings and covers open-vocabulary concepts. VS^3 inherits Enc_I, Enc_L and the cross-module fusion module from GLIP.

Text prompt. Considering that object detection has been reformulated as phrase grounding, VS^3 also requires a text prompt input except for the image input. Following GLIP's design, we set the text prompt for object detection in the form of "$name(c_1)$. $name(c_2)$. ... $name(c_{|\mathcal{C}_o|})$.", where $c_i \in \mathcal{C}_o$ and $name(c_i)$ gets the category name of c_i (e.g., *person*). Hence, an object is detected according to the alignment score between a region embedding $\tilde{\mathbf{o}}_i \in \tilde{O}$ (the ith row) and the category name embeddings $\tilde{P}$.

Relation embedding module. To further enable VS^3 with relation recognition ability, we devise the relation embedding module to build relation representations. Based on the region/box features $\tilde{O}$ after cross-modal fusion, we first sample a subset of regions $\tilde{O}' \in \mathbb{R}^{N'\times d}$ that are most likely to be valid objects. This is achieved by matching predicted bounding boxes $\hat{B}$ with ground-truth objects during training, and by retaining top-N' regions with the highest confidence scores after non-maximum suppression (NMS) at inference. Next, we construct relation representations for all possible subject-object pairs. Given an object pair $(\tilde{\mathbf{o}}_i, \tilde{\mathbf{o}}_j)$ and their normalized bounding boxes $(\mathbf{b}_i, \mathbf{b}_j)$, the pairwise relation representation is represented as $\mathbf{pair}_{i\to j} = cat[\mathbf{pair}^{visual}_{i\to j}, \mathbf{pair}^{spatial}_{i\to j}]$. This is the concatenation of features mined from the visual and spatial perspectives. The visual feature is computed by

$$\mathbf{pair}^{visual}_{i\to j} = \boldsymbol{f}_{diff}(\tilde{\mathbf{o}}_i - \tilde{\mathbf{o}}_j) + \boldsymbol{f}_{sum}(\tilde{\mathbf{o}}_i + \tilde{\mathbf{o}}_j), \tag{4.1}$$

where $\boldsymbol{f}_{diff}$ and $\boldsymbol{f}_{sum}$ are two mapping functions implemented as 2-layer MLPs (multilayer perceptron). By defining the normalized center coordinates of two involved objects as (ct^x_i, ct^y_i) and (ct^x_j, ct^y_j), the spatial feature is measured as

$$\mathbf{pair}^{spatial}_{i\to j} = cat[\mathbf{b}_i, \mathbf{b}_j, dx, dy, dis, \theta, A_i, A_j, I, U], \tag{4.2}$$

where $dx = ct^x_i - ct^x_j, dy = ct^y_i - ct^y_j, dis = \sqrt{dx^2 + dy^2}, \theta = arctan(\frac{dy}{dx})$. A_i, A_j, I, U denote the areas of the subject, the object, their intersection, and union boxes, respectively.

Relation prediction. Conditioned on the relation representation $\mathbf{pair}_{i\to j}$ of each object pair, we predict a relateness score $\hat{z}_{i\to j} = f_{relateness}(\mathbf{pair}_{i\to j}) \in [0, 1]$ and a semantic label probability $\hat{\boldsymbol{y}}_{i\to j} = \boldsymbol{f}_{semantic}(\mathbf{pair}_{i\to j}) \in [0, 1]^{|\mathcal{C}_r|}$. The relateness $\hat{z}_{i\to j}$ represents the probability that relations exist between the object pair. $f_{relateness}$ is implemented with an MLP coupled with Sigmoid activation. $\boldsymbol{f}_{semantic}$ is implemented with another MLP using Softmax activation. The total loss for relation recognition

L_{rel_rcg} is measured as

$$L_{relateness} = FL(\hat{z}_{i \to j}, z_{i \to j}), \tag{4.3}$$

$$L_{semantic} = CE(\hat{\boldsymbol{y}}_{i \to j}, \boldsymbol{y}_{i \to j}), \tag{4.4}$$

$$L_{rel\_rcg} = L_{relateness} + L_{semantic}, \tag{4.5}$$

where FL and CE denote focal loss [120] and cross-entropy loss functions. $z_{i \to j}$ and $\boldsymbol{y}_{i \to j}$ represent the ground-truth relateness label and predicate category label respectively.

Training & inference. During training, we initialize parameters from pre-trained GLIP models for inherited modules in VS^3. To ease training difficulty, we freeze the image encoder and text encoder, and only fine-tune the cross-modal fusion module and devised modules for relation recognition. This also avoids the degeneration of the pre-trained VSS. At inference, by retaining high-confidence detected objects and further predicting their relations, we generate an image scene graph representation.

4.3.3 Obtaining Language Scene Graph Supervision

Ground-truth scene graphs are time-consuming to annotate. Alternatively, we can parse semantic graphs from image language descriptions, and obtain noun phrase groundings through region-word alignment in the pre-trained VSS (implemented with an off-the-shelf GLIP). This is a much cheaper way to obtain weak scene graph supervision.

Semantic graph parsing. Concretely, for each image language description, we parse it into a semantic graph $SG^{text} = \{O^{text}, R^{text}\}$ using the Standard Scene Graph Parser based on [142]. The parser not only extracts noun phrases as entities/objects (O^{text}), but also extracts the words describing their relations (R^{text}). For example, the sentence "*a woman is playing the piano in the room.*" is parsed to the SG^{text}, of which $O^{text} = \{woman, piano, room\}$ and $R^{text} = \{\langle 0, playing, 1\rangle, \langle 0, in, 2\rangle\}$ (numbers denote object indices). Considering that parsed object/relation words are free-form, we map them to our concerning categories (e.g., VG150 object/relation categories in experiments) by rules such as direct string matching and WordNet [195] synsets matching following [85].

Semantic graph grounding. Note that each element of O^{text} only contains a text label name so far, and its bounding box information is still missing. To obtain grounding boxes, we construct a text prompt using triplets in SG^{text}, e.g., "*woman playing piano. woman in room.*". Then, we feed such text prompt together with the raw image into a pre-trained GLIP, in order to acquire grounding boxes of O^{text}. Specifically, for each element in O^{text}, we select the image region that has the highest alignment score with its category name as its grounding box. Since there might be multiple objects in O^{text} that actually refer to the same object, we perform a post-processing NMS to merge boxes with the same label and high IoU (intersection over union) scores (≥ 0.9). Finally, with box information, the visually-grounded SG^{text} is ready to be used as weak supervision for training SGG models, e.g., the proposed VS^3.

4.3.4 Transferring to Open-vocabulary SGG

Open-vocabulary SGG [201] aims to train SGG models that can recognize objects of novel categories and their involved relations. Formally, we train the SGG model with scene graphs containing objects in the base category set $\mathcal{C}_o^{base}$. At inference, the object category set is $\mathcal{C}_o^{target}$, which contains novel categories in $\mathcal{C}_o^{novel} = \mathcal{C}_o^{target} \backslash \mathcal{C}_o^{base} \neq \emptyset$.

Back to our proposed VS^3, an open-vocabulary VSS is maintained by freezing the image and text encoders. Taking this advantage, we devise a scheme to adapt VS^3 for open-vocabulary SGG. Concretely, during training, we set the text prompt as "$name(c_1)$. $name(c_2)$. ... $name(c_{|\mathcal{C}_o^{base}|})$.", where $c_i \in \mathcal{C}_o^{base}$. And only relation triplets involving base object categories are kept for training. At inference, the text prompt is switched to be "$name(c_1)$. $name(c_2)$. ... $name(c_{|\mathcal{C}_o^{target}|})$.", where $c_i \in \mathcal{C}_o^{target}$. In this way, a novel object class (e.g., *lady*) may have an embedding close to a base category (e.g., *woman*) embedding. This makes the novel class also able to find well-aligned image regions. Note that relation representations are constructed from visual and spatial cues, which are usually class-agnostic. Hence, the following relation recognition in VS^3 will not be affected when encountering novel objects.

4.4 Experiments

4.4.1 Datasets and Experimental Settings

Datasets. To evaluate the SGG task, we adopt the widely-used ***VG150*** version [42] of the Visual Genome (VG) dataset [53]. VG150 retains the most frequent 150 categories and 50 relation/predicate categories in VG. It contains ∼108K images, of which 70% images are used for training (including 5K for validation), and the remaining 30% images are used for testing. The annotated scene graph of each image has 11.5 objects and 6.2 relation triplets on average. In addition, images of VG are densely annotated with region descriptions, about 50 descriptions for each image. We refer to these descriptions as ***VG caption***, which provides a text source for evaluating the language-supervised SGG setting. Moreover, we consider the challenging setting of using image-text pairs in ***COCO caption*** [204] for training SGG models. This dataset contains 123K images in total. Each image has 5 human-annotated captions. We keep ∼106k images by filtering out those images that also exist in the VG150 test split.

Evaluation protocols and metrics. We mainly adopt the SGDET [42, 78] protocol, which generates a scene graph from the input image without any given box information. We report the performance on Recall@K (K=20/50/100) following previous works [42, 78, 85, 86, 87, 159], which measures the fraction of correctly predicted relation triplets in top K predictions. A triplet prediction is considered as correct when its subject, object, predicate labels and both the subject and object regions match with (same label or IoU>0.5) a ground-truth triplet. Note that we obtain triplet predictions using graph constraint, which limits each subject-object pair to have only the most confident predicate. All recall metrics across different SGG settings in exper-

Methods	Original test split (∼26k)	New test split (∼15k)	Difference	Relative
IMP [42]	18.59 / 26.36 / 31.62	18.45 / 26.35 / 31.57	-0.14 / -0.01 / -0.05	0.75%
VTransE [205]	23.06 / 29.99 / 34.69	23.06 / 29.91 / 34.59	-0.00 / -0.08 / -0.10	0.29%
VCTREE [134]	24.51 / 31.29 / 35.98	24.45 / 31.19 / 35.87	-0.06 / -0.10 / -0.11	0.32%
MOTIFS [62]	25.29 / 32.30 / 37.08	25.16 / 32.21 / 36.94	-0.13 / -0.09 / -0.14	0.51%
VS^3 *(Swin-T)*	-	26.10 / 34.53 / 39.18	-	-
VS^3 *(Swin-L)*	-	**27.81** / **36.63** / **41.50**	-	-

Table 4.2: Performance comparisons in the evaluation metrics (R@20/50/100) between the original VG150 test split and the new test split (removing invalid images that have already been seen during GLIP [189] pre-training). We observe the evaluation differences between these two splits ($<$ 0.15% variations) are somewhat trivial in comparison with the performance differences between different SGG methods (e.g., VTransE improves over IMP by $>$ 3%). Difference = New test split metrics - Original test split metrics; Relative = max(|Difference| / Original test split metrics).

iments are computed over VG150 test images. Considering that the adopted GLIP pre-trained VSS has seen part of images in the original VG150 test split (∼26k) during pre-training, we exclude these images and get a new split of ∼15k test images. We have validated that such VG150 test split is sufficiently large for computing stable metrics as the original, by comparing computed metrics of several SGG models (in codebase [170]) on these two splits ($<$ 0.15 points variation, see note here[2]).

Implementation details. We initialize VS^3 from pre-trained GLIP [189] models, i.e., the GLIP-T and the larger GLIP-L trained with more data. Both construct a VSS of dimension $d = 256$. We retain the top 36 object detections per image for pairwise relation recognition. The whole framework is fine-tuned on 8 Nvidia 2080Ti GPUs with AdamW optimizer. During fine-tuning, we freeze the parameters of the image and text encoder; and set the learning rate for the cross-modal fusion module as 1e-5 and 10x larger learning rates for the relation embedding and prediction modules. The maximum fine-tuning epoch number is 10, with learning rates dropping by 10x after 6 epochs.

[2] As mentioned, we compute all recall metrics over test images of the VG150 dataset. However, considering that the adopted GLIP pre-trained visual-semantic space (VSS) has seen part of images in the original VG150 test split (∼26k) during pre-training [189], we exclude these overlapped images and achieve a new split of ∼15k test images. Our approach adopts the same VG150 train split and computes evaluation metrics over the new test split.

Since the new test split is a subset of the original test split, one concern is whether or not the performances over these two test splits would exhibit significant variations. In an effort to delve into this concern, we compare the performances of several classic SGG methods on the original and the new test split. These methods are IMP [42], VTransE [205], VCTREE [134] and MOTIFS [62], with stable implementations in codebase [170]. Table 4.2 summarizes the results. We observe that the computed recalls show only $<$ 0.15% variations (relatively $<$ 0.8%) between the two different test splits. Such results basically validate that the new VG150 test split can lead to stable recall metrics with mostly the same performance trends as in the original split. The performance variations between these two test splits are trivial in comparison with the performance differences between different SGG methods. For example, in Table 4.2, our VS^3 *(Swin-L)* achieves $>$ 2.65% performance boosts than the mentioned baselines. Accordingly, we directly compare the performances obtained on the new VG150 test split by our method against the performances reported in previous works on the original VG150 test split.

4.4.2 Fully Supervised SGG

Setup. We first evaluate our proposed VS^3 under the conventional fully supervised SGG setting. This setting trains SGG models using manually annotated scene graphs, consisting of object labels coupled with bounding boxes, and relation labels. We adopt VG150 for training and evaluation following previous methods [42, 62, 73, 74, 75, 76, 85, 134, 205, 206]. All these methods involve a closed set of object categories. Specifically, the text prompt input of VS^3 is constructed from VG150 object category names, i.e., "*airplane. animal. ... zebra.*". We train VS^3 by fine-tuning over two GLIP variants: GLIP-T with the Swin-T [202] backbone, GLIP-L with the Swin-L [202] backbone.

Comparison with state-of-the-arts. The results are summarized in Table 4.3. Our proposed VS^3 model using the Swin-T backbone already achieves competitive recall metrics. When upgrading to the larger Swin-L variant, the performance improvements become significant (1.8 to 3.4 points improvement than the previous best results). Note that previous methods [42, 62, 75, 76] build their models upon an off-the-shelf object detector, and they usually design heavy message-passing modules to incorporate context information. Instead, VS^3 devises a light-weighted relation recognition head (including the relation embedding and prediction modules) over a pre-trained VSS. The superior performances clearly suggest the merits of transferring language-image pre-trained models for boosting SGG.

Ablation on relation representation. Next, we carry out ablation studies on relation representation construction in the relation embedding module. As shown in Table 4.3, by removing visual and spatial feature components, the relation triplet recalls drop accordingly. Also notice that the removal of visual features which is built from subject and object region embeddings leads to relatively larger performance drops than spatial. This suggests the region embeddings in the pre-trained VSS provide strong cues for relation recognition. Overall, these observations validate the effectiveness of our design to mine relation patterns.

Additional results on mean recalls. We report additional results on the unbiased metric under the same fully supervised SGG in Table 4.4. Concretely, the adopted unbiased metric is mean Recall@K (mR@K), which averages Recall@K across all predicate categories. The results show that the pre-trained VSS does not mitigate the bias issue very significantly, since the bias issue is naturally rooted in the pre-training image-text corpus. Note that we can easily apply debias techniques (e.g., reweight, TDE [78]) in our VS^3 framework to further mitigate the bias issue.

4.4.3 Language-supervised SGG

Setup. Language-supervised SGG [85, 86, 87, 159] explores to train SGG models with language descriptions of images. Concretely, we parse each image description into a semantic graph, in the form of a set of $\langle subject, predicate, object \rangle$ triplets. Note that parsed object/relation phrases from language descriptions are free-form, we map them to VG150 categories by semantic rules following [85], such as WordNet [195] synsets matching. This makes the learned SGG model compatible for evaluating on VG150. Next, the parsed semantic graph is grounded to image regions using grounding

SGG model	Detector	Backbone	R@20	R@50	R@100
FCSGG [73]	-	HRNetW48	16.1	21.3	25.1
SGTR [74]	DETR	R-101	-	24.6	28.4
IMP [42]	Faster-RCNN	VGG-16	14.6	20.7	24.5
KERN [206]	Faster-RCNN	VGG-16	-	27.1	29.8
MOTIFS [62]	Faster-RCNN	VGG-16	21.4	27.2	30.3
GPS-Net [68]	Faster-RCNN	VGG-16	22.3	28.9	33.2
RelDN [61]	Faster-RCNN	VGG-16	21.1	28.3	32.7
VTransE [205]	Faster-RCNN	RX-101	23.0	29.7	34.3
MOTIFS [62]	Faster-RCNN	RX-101	25.1	32.1	36.9
VCTREE [134]	Faster-RCNN	RX-101	24.7	31.5	36.2
SGNLS [85]	Faster-RCNN	RX-101	24.6	31.8	36.3
RU-Net [76]	Faster-RCNN	RX-101	25.7	32.9	37.5
HL-Net [75]	Faster-RCNN	RX-101	26.0	33.7	38.1
VS^3	-	Swin-T	26.1	34.5	39.2
w/o visual	-	Swin-T	23.1	31.6	36.7
w/o spatial	-	Swin-T	24.3	32.8	37.8
VS^3	-	Swin-L	**27.8**	**36.6**	**41.5**

Table 4.3: Experimental results of fully supervised SGG. *w/o visual* and *w/o spatial* indicate removing spatial and visual features in the relation embedding module for relation representation. All metrics are computed under the SGDET protocol on VG150 test images.

SGG model	Detector	Backbone	mR@20	mR@50	mR@100
IMP [42]	Faster-RCNN	RX-101	2.8	4.2	5.3
VTransE [205]	Faster-RCNN	RX-101	3.7	5.0	6.0
VCTREE [134]	Faster-RCNN	RX-101	4.2	5.7	6.9
MOTIFS [62]	Faster-RCNN	RX-101	4.1	5.5	6.8
VS^3	-	Swin-T	**4.3**	**6.6**	**8.1**

Table 4.4: Experimental results of fully supervised SGG. All metrics are computed under the SGDET protocol on VG150 test images. Results of previous models come from Tang et al. [78].

methods, i.e., the pre-trained GLIP-L [189] in our approach. Finally, the visually-grounded semantic graphs are used as weak supervision to train our proposed VS^3 like the fully supervised setting.

Particularly, we have trained VS^3 with text triplets parsed from three different sources of text following [85, 159]. 1) The *unlocalized graph* setting uses ground-truth triplet annotations in VG. 2) The *VG caption* setting uses triplets that are automatically parsed from natural image descriptions in VG. 3) The *COCO caption* setting leverages triplets parsed from captions in COCO. This setting is the most challenging since COCO captions are image-level descriptions. Such captions are different from the region-level descriptions in VG, which focus on describing object interactions. Also, note that the number of annotated captions for each COCO image (average 5) is much less than the number for each VG image (average ∼50).

Comparison with state-of-the-arts. The experimental results compared with previous methods are presented in Table 4.5. All evaluation metrics are computed on the VG150 test set under the SGDET protocol. Specifically, under the unlocalized scene graphs setting, VS^3 with the Swin-T backbone (VS$^3{}_{(\textit{Swin-T})}$) obtains

	SGG model	Grounding	R@20	R@50	R@100
Unlocalized graph	VSPNet [69]	-	-	4.70	5.40
	LSWS [87]	-	-	7.30	8.73
	MOTIFS [62]	WSGM [86]	4.12	5.59	6.45
	MOTIFS [62]	SGNLS [85]	7.23	9.28	10.71
	MOTIFS [62]	Li et.al [159]	9.09	11.39	12.89
	Uniter† [207]	SGNLS [85]	7.81	10.03	11.50
	Uniter† [207]	Li et.al [159]	9.57	11.80	13.15
	VS$^3{}_{\textit{(Swin-T)}}$	GLIP-L [189]	18.02	23.89	28.19
	VS$^3{}_{\textit{(Swin-T+FreqBias)}}$	GLIP-L [189]	20.06	26.72	31.75
	VS$^3{}_{\textit{(Swin-L+FreqBias)}}$	GLIP-L [189]	**22.18**	**29.81**	**34.96**
VG caption	LSWS [87]	-	-	3.85	4.04
	MOTIFS [62]	SGNLS [85]	6.31	8.05	9.21
	MOTIFS [62]	Li et.al [159]	8.25	10.50	11.98
	Uniter† [207]	SGNLS [85]	-	9.20	10.30
	Uniter† [207]	Li et.al [159]	8.90	10.93	12.14
	VS$^3{}_{\textit{(Swin-T)}}$	GLIP-L [189]	11.78	16.25	19.7
	VS$^3{}_{\textit{(Swin-L)}}$	GLIP-L [189]	**13.01**	**17.38**	**20.54**
COCO caption	LSWS [87]	-	-	3.28	3.69
	MOTIFS [62]	Li et.al [159]	5.02	6.40	7.33
	Uniter† [207]	SGNLS [85]	-	5.80	6.70
	Uniter† [207]	Li et.al [159]	5.42	6.74	7.62
	VS$^3{}_{\textit{(Swin-T)}}$	GLIP-L [189]	5.59	7.30	8.62
	VS$^3{}_{\textit{(Swin-L)}}$	GLIP-L [189]	**6.04**	**8.15**	**9.90**

Table 4.5: Comparison with state-of-the-art language-supervised SGG methods, using weak scene graph supervision from three different text sources: unlocalized scene graphs, VG caption and COCO caption. All metrics are computed under the SGDET protocol on VG150 images. († indicates adapted for SGG.)

substantial improvements on recall metrics over existing best results (R@20/50/100 from 9.57/11.80/13.15 to 18.02/23.89/28.19). Since relation frequency statistics are available in this setting, we use them as frequency biases [62] in predicate classification, leading to further performance gains (VS$^3{}_{\textit{(Swin-T+FreqBias)}}$). When using the stronger Swin-L backbone (VS$^3{}_{\textit{(Swin-L+FreqBias)}}$), we attain the highest performances (R@20/50/100 = 22.18/29.81/34.96), which even outperform many fully supervised methods (see Table 4.3). The VG caption setting provides weaker scene graph supervision via language parsing. We observe that our approach also outperforms previous state-of-the-art methods significantly. As for the most challenging COCO caption setting, it suffers from the additional domain shift problem since it trains on COCO but evaluates on VG150. As expected, the performances are lower than in the two aforementioned settings. But when comparing with previous works using the same text source, our approach still manages to achieve better performances. Overall, our approach consistently surpasses previous methods for language-supervised SGG. This demonstrates the benefits brought by pre-trained language-image models in terms of both grounding box acquirement and task transferring to tackle SGG.

Ablation on scene graph parsing strategy. We also conduct ablation studies on different scene graph parsing strategies for obtaining language SGG supervision. The results are shown in Table 4.6. Note that each image in COCO is annotated with 5 captions, and these captions are usually complimentary in describing image content. At first, we compare the performances between training with triplets from a

SG from	*SG* parser	R@20/50/100
Single caption	Simple	5.07 / 6.25 / 7.36
All captions	Simple	5.42 / 6.82 / 7.93
All captions	Advanced	**5.59** / **7.30** / **8.62**

Table 4.6: Ablation on scene graph parsing strategies for language-supervised SGG. Results are obtained with VS$^3_{(Swin\text{-}T)}$ trained on scene graph supervision parsed from COCO captions.

single caption and all captions. We see the recalls achieve relative 10% performance boosts by replacing triplets from a single caption with all captions. This suggests the completeness of extracted scene graphs from image descriptions is a non-negligible factor for training a high-quality SGG model.

Moreover, we compare two language parsers for extracting $\langle subject, predicate, object\rangle$ triplets: the simple SG parser [194], and the advanced SG parser [142]. Both parsers apply pre-defined rules to extract object and relation concepts from the semantic graphs of image language descriptions. Compared with the simple SG parser, the advanced SG parser covers additional features for dealing with complex quantificational modifiers (e.g., *a lot of*), resolving pronouns (e.g., *it*) and handling plural nouns (e.g., *three men*). The performance boosts of the advanced SG parser over the simple one (recalls from 5.42/6.82/7.93 to 5.59/7.30/8.62), indicates that the quality of semantic parsing is also important for language-supervised SGG.

4.4.4 Open-vocabulary SGG

Setup. Following [201], we train the proposed VS^3 with the same 70% object categories of VG150 as base categories. With the aid of the pre-trained VSS, we hope VS^3 can generalize to recognize the remaining 30% novel objects and their involved relations at inference. Concretely, we compute evaluation metrics over two object category sets: 70% base + 30% novel objects (dubbed as open-vocabulary SGG (Ov-SGG) evaluation), and 30% novel objects (dubbed as zero-shot SGG (ZsO-SGG) evaluation).

In addition, we adopt the PREDCLS and SGDET evaluation protocols [42]. PREDCLS assumes object information given, yet SGDET generates scene graphs from the raw image using predicted objects. Since VS^3 detects objects in a one-stage manner, we implement PREDCLS by selecting image regions that best match the ground-truth objects in post-processing, then performing relation recognition. We neglect the SGCLS protocol that assumes bounding box information given. This is because given bounding boxes can be directly used as region proposals in two-stage detectors, while the adopted one-stage manner in VS^3 has no region proposal counterpart.

Fully supervised results. We first conduct experiments using manually annotated scene graphs. The results are presented in Table 4.7. For both Ov-SGG and ZsO-SGG, VS^3 achieves substantial performance improvements under PREDCLS. When upgrading to the stronger backbone Swin-L, more significant improvements are obtained. More importantly, we report performances for the challenging and more practical SGDET, which are neglected by all previous methods since their used object detector cannot handle open-vocabulary detection [201]. The SGDET performances

Method	Ov-SGG (70%+30%) PREDCLS	Ov-SGG (70%+30%) SGDET	ZsO-SGG (30%) PREDCLS	ZsO-SGG (30%) SGDET
IMP [42]	40.02 / 43.40	-	37.01 / 39.46	-
MOTIFS [62]	41.14 / 44.70	-	39.53 / 41.14	-
VCTREE [134]	42.56 / 45.84	-	41.27 / 42.52	-
TDE [78]	38.29 / 40.38	-	34.15 / 36.37	-
GCA [208]	43.48 / 46.26	-	42.56 / 43.18	-
EBM [209]	44.09 / 46.95	-	43.27 / 44.03	-
SVRP [201]	47.62 / 49.94	-	45.75 / 48.39	-
$\text{VS}^3_{\textit{(Swin-T)}}$	50.10 / 52.05	15.07 / 18.73	46.91 / 49.13	10.08 / 13.65
$\text{VS}^3_{\textit{(Swin-L)}}$	**55.88** / **58.18**	**23.13** / **28.49**	**54.44** / **57.35**	**21.51** / **27.62**

Table 4.7: Evaluation results (R@50/100) of fully supervised open-vocabulary SGG. Ov-SGG evaluates on 70% base categories + 30% novel categories in VG150, while ZsO-SGG only evaluates on 30% novel categories.

Method	*SG* supervision	Ov-SGG (70%+30%)	ZsO-SGG (30%)
$\text{VS}^3_{\textit{(Swin-T)}}$	Manual annotation	**15.07** / **18.73**	**10.08** / **13.65**
	VG caption	7.61 / 9.60	4.06 / 5.58
	COCO caption	4.39 / 5.63	3.65 / 4.73
$\text{VS}^3_{\textit{(Swin-L)}}$	Manual annotation	**23.13** / **28.49**	**21.51** / **27.62**
	VG caption	12.98 / 16.29	10.71 / 13.70
	COCO caption	6.76 / 8.45	6.26 / 7.89

Table 4.8: Evaluation results (R@50/100) of open-vocabulary SGG using three different scene graph supervisions: manual annotation, VG caption and COCO caption (language-supervised). Ov-SGG evaluates on 70% base categories + 30% novel categories in VG150, while ZsO-SGG only evaluates on 30% novel categories.

($R@50/100 = 10.08/13.65$) of ZsO-SGG using VS^3 are even higher than SGCLS metrics of SVRP ($R@50/100 = 9.30/11.32$ in [201]). This reveals the superiority of our approach to recognizing novel objects thanks to the open-vocabulary generalization ability of the pre-trained VSS.

Language-supervised results. Next, we evaluate the most challenging setting, i.e., open-vocabulary SGG using language supervision. To our knowledge, we are the first to propose such a new and practical SGG setting, and present the benchmark performances in Table 4.7. Not surprisingly, the recalls obtained via language-supervised training (i.e., *SG* from VG caption or COCO caption) are lower than supervised results (i.e., *SG* from annotated). When comparing $\text{VS}^3_{\textit{(Swin-T)}}$ and $\text{VS}^3_{\textit{(Swin-L)}}$ that is transferred from a stronger pre-trained model, the latter gets substantially higher Ov-SGG and ZsO-SGG performances. More importantly, we observe the performance gap between Ov-SGG and ZsO-SGG get closer in $\text{VS}^3_{\textit{(Swin-L)}}$, e.g., the $R@50$ gap under the VG caption setting becomes *12.98-10.71=2.27* from *7.61-4.06=3.55*. This is due to the better generalization ability for recognizing novel classes. Moreover, the superior performances obtained by VG caption over COCO caption, indicate that using dense region-level descriptions and avoiding domain shift will help improve language-supervised open-vocabulary SGG in practice.

Qualitative analysis. We further showcase qualitative results of open-vocabulary SGG in Figure 4.3. The results demonstrate that our approach manages to detect novel objects and their relations with other objects. We also find that, compared with

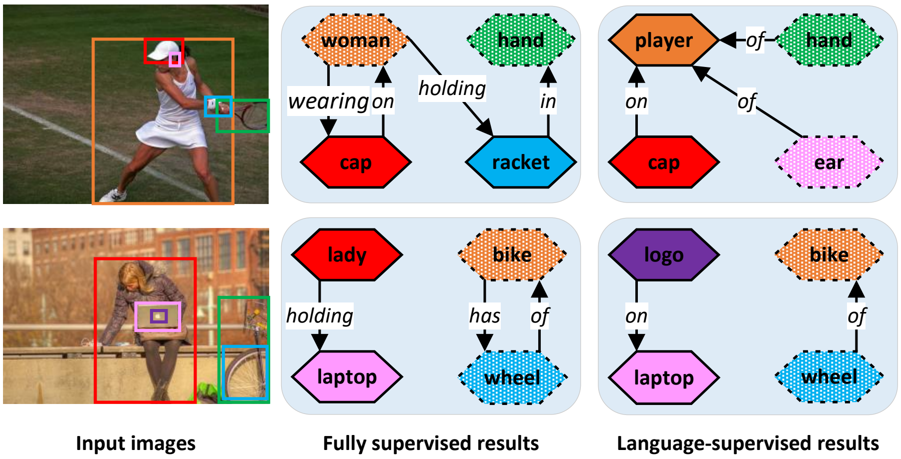


Figure 4.3: Qualitative results of open-vocabulary SGG, particularly from fully supervised and language-supervised (VG caption) settings. Note that dotted nodes denote novel objects. For clarity, we only show triplets among the top 20 predictions that depict relations of highlighted image regions (i.e., boxes on input images).

the fully supervised setting, the language-supervised results bias to predict simple relations such as ‘on’, ‘of’. Presumably, it’s because scene graph supervision parsed from language is more likely to extract such simple words as relation predicates.

4.5 Conclusion

In this chapter, we have proposed a novel approach to exploit a powerful pre-trained VSS for triggering language-supervised and open-vocabulary SGG. Particularly, we obtain cheap scene graph supervision by semantically parsing image language descriptions into semantic graphs and grounding the noun phrases through region-word alignment in the VSS. In addition, we devise the VS^3 model, which performs object detection as category name grounding in the VSS and naturally builds relation representations for relation recognition. Thanks to the open-vocabulary generalization ability of the VSS, VS^3 manages to detect novel objects and their relations with other objects, achieving open-vocabulary SGG. We validate our approach on the Visual Genome benchmark across supervised, language-supervised and open-vocabulary SGG settings, and have set new state-of-the-art performances. This demonstrates the merits of transferring pre-training knowledge to push SGG toward more practical scenarios.

Chapter 5

Human-centric Scene Graph Generation using Transformer

Our previous works generate scene graphs that involve relations of all kinds of objects. Note that human-object interactions (i.e., human-centric relations) are the most noteworthy relations in scene graph representations. In this chapter[1], we exploit the high-performing Transformer architecture for human-centric SGG - or the Human-Object Interaction (HOI) detection task. Its input is an image, and the output is a set of HOI triplets, or equivalently a human-centric scene graph representation.

We notice that recent high-performing HOI detection techniques have been highly influenced by Transformer-based object detectors (i.e., DETR). Nevertheless, most of them directly map parametric interaction queries into a set of HOI predictions through a vanilla Transformer in a one-stage manner. This leaves rich inter- or intra-interaction structures under-exploited. In this work, we design a novel Transformer-style HOI detector, i.e., Structure-aware Transformer over Interaction Proposals (STIP), for HOI detection. Such design decomposes the process of HOI set prediction into two subsequent phases, i.e., an interaction proposal generation is first performed, and then followed by transforming the non-parametric interaction proposals into HOI predictions via a structure-aware Transformer. The structure-aware Transformer upgrades the vanilla Transformer by encoding additionally the holistically semantic structure among interaction proposals as well as the local spatial structure of human/object within each interaction proposal, so as to strengthen HOI predictions. Extensive experiments conducted on V-COCO and HICO-DET benchmarks have demonstrated the effectiveness of STIP, and superior results are reported when comparing with the state-of-the-art HOI detectors.

[1] ***Yong Zhang***, *Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. "Exploring Structure-aware Transformer over Interaction Proposals for Human-Object Interaction Detection." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.*

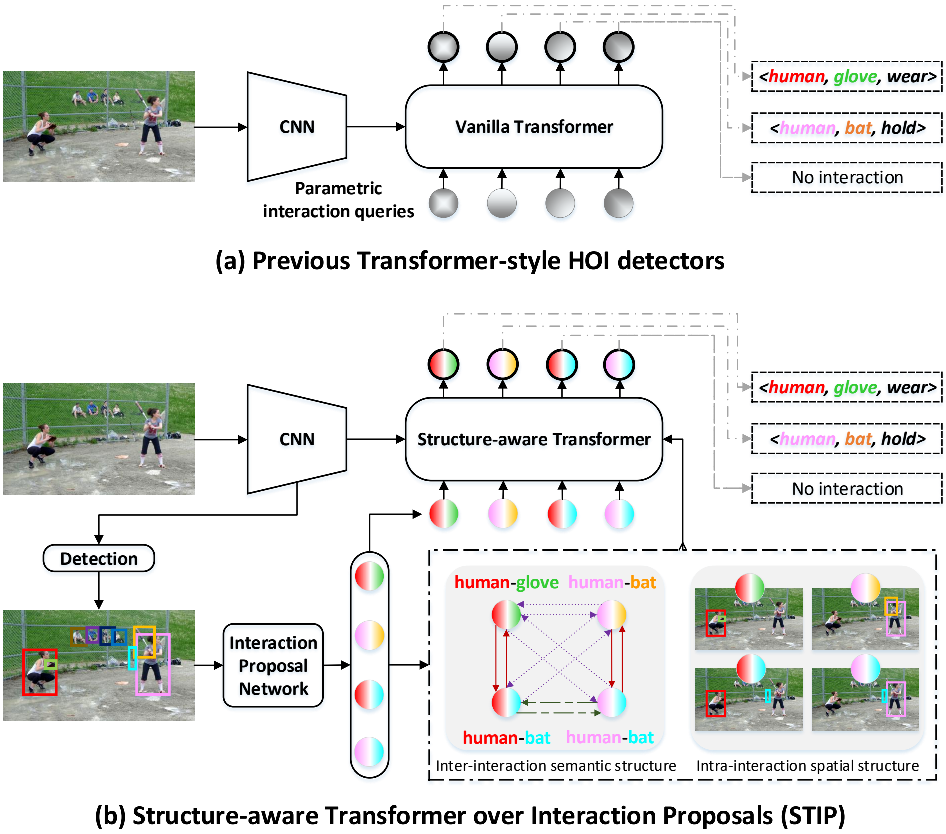

CNN
Vanilla Transformer
Parametric interaction queries
<human, glove, wear>
<human, bat, hold>
No interaction
(a) Previous Transformer-style HOI detectors
CNN
Structure-aware Transformer
Detection
Interaction Proposal Network
human-glove
human-bat
human-bat
human-bat
Inter-interaction semantic structure
Intra-interaction spatial structure
<human, glove, wear>
<human, bat, hold>
No interaction
(b) Structure-aware Transformer over Interaction Proposals (STIP)

Figure 5.1: Comparison between existing Transformer-style HOI detectors and our STIP. (a) Existing Transformer-style HOI detectors directly transform the parametric interaction queries into HOI predictions via vanilla Transformer in a one-stage fashion. (b) STIP adopts a two-phase solution, i.e., first producing interaction proposals via Interaction Proposal Network, and then mapping the non-parametric interaction queries (i.e., interaction proposals) into HOI predictions. Both the inter- and intra-interaction structures derived from interaction proposals are additionally exploited to boost HOI set prediction through a structure-aware Transformer.

5.1 Introduction

Human-Object Interaction (HOI) detection [210, 211] is intended to localize the interactive human-object pairs within an image and identify the interactions in between, yielding the HOI predictions in the form of $\langle human, object, interaction \rangle$ triplets. Practical HOI detection systems perform the human-centric scene understanding, and thus have a great potential impact for numerous applications, such as surveillance event detection [212, 213] and robot imitation learning [214]. In general, conventional HOI detectors [127, 129, 215, 216, 217, 218, 219, 220, 221, 222] tackle the HOI set prediction task in an **indirect** way, by formalizing it as surrogate regression and classification problems for human/object/interaction. Such an indirect approach needs subsequent post-processing by collapsing near-duplicate predictions and heuristic matching

[129, 218, 222], and thus cannot be trained in an end-to-end fashion, resulting in a sub-optimal solution. The intent to overcome the problem of sub-optimal solution leads to the development of recent state-of-the-art HOI detectors [128, 130, 132, 223] that follow the Transformer-based detector of DETR [22] to cast HOI detection as a **direct** set prediction problem, and embrace the "end-to-end" philosophy (Figure 5.1 (a)). In particular, a vanilla Transformer is commonly utilized to map the parametric interaction queries (i.e., the learnable positional embedding sequence) into a set of HOI predictions in a one-stage manner. However, these HOI detectors start the HOI set prediction from the parametric interaction queries with randomly initialized embeddings. That is, the correspondence between parametric interaction queries and output HOIs (commonly assigned by Hungarian algorithm for training) is **dynamic** in which the interaction query corresponding to each target HOI (e.g., "human hold bat") is unknown at the beginning of HOI set prediction. This can adversely hinder the exploration of prior knowledge (i.e., **inter-interaction** or **intra-interaction structure**) which would be very useful for relational reasoning among interactions in HOI set prediction.

Specifically, by inter-interaction structure, we refer to the holistic semantic dependencies among HOIs, which can be directly defined by considering whether or not two HOIs share the same human or object. Such structure implies the common sense knowledge that shall facilitate the prediction of one HOI by exploiting its semantic dependencies against other HOIs. Taking the input image in Figure 5.1 as an example, the existence of "human wear (baseball) glove" provides a strong indication for "(another) human hold bat". Moreover, the intra-interaction structure can be interpreted as the local spatial structure for each HOI, i.e., the layout of human and object, which acts as additional prior knowledge to steer model's attention over image areas for depicting the interaction.

In this work, we design a novel scheme based on a Transformer-style HOI detector, namely Structure-aware Transformer over Interaction Proposals (STIP). The design innovation is to decompose the one-stage solution of set prediction into two cascaded phases, i.e., first producing the interaction proposals (i.e., the probably interactive human-object pairs) and then performing HOI set prediction based on the interaction proposals (Figure 5.1 (b)). By taking the interaction proposals derived from Interaction Proposal Network (IPN) as non-parametric interaction queries, STIP naturally triggers the subsequent HOI set prediction with more reasonable interaction queries, leading to **static** query-HOI correspondence that is capable of boosting HOI set prediction. As a beneficial by-product, the predicted interaction proposals offer a fertile ground for constructing a structured understanding across interaction proposals or between human & object within each interaction proposal. A particular form of Transformer, i.e., structure-aware Transformer, is designed accordingly to encode the inter-interaction or intra-interaction structure for enhancing HOI predictions.

In sum, we have made the following contributions: (1) The proposed two-phase implementation of the Transformer-style HOI detector is shown capable of seamless incorporation of potential interactions among HOI proposals to overcome the problem associated with one-stage approach; (2) The exquisitely designed structure-aware Transformer is shown able to facilitate additional exploitation opportunity for utilizing inter-interaction and intra-interaction structure for enhanced performance of the vanilla Transformer; (3) The proposed structure-aware Transformer approach has been properly analyzed and verified through extensive experiments over V-COCO and HICO-DET datasets to validate its potential in solving the problems associated with one-stage approach to achieve desirable HOI detection.

5.2 Related Work

The task of Human-Object Interaction (HOI) detection has been primordially defined [210, 211] and recent developments of HOI detectors can be briefly divided into two categories: the two-stage methods and one-stage approaches.

Two-stage Methods. The first category schemes [126, 127, 215, 216, 217, 219, 220, 221, 224, 225, 226, 227, 228, 229, 230] mainly adopt two-stage paradigm, i.e., first detect humans/objects via off-the-shelf modern object detectors (e.g., Faster R-CNN [231]) and then carry out interaction classification. A number of schemes have been proposed to strengthen the HOI feature learning in the second stage for interaction classification. Generally, similar to prior works for visual relationship detection [58, 232, 233], HOI features are typically derived from three perspectives [126, 215, 216]: appearance/visual features of humans and objects, spatial features (e.g., the pairwise bounding boxes of human-object pair), and linguistic feature (e.g., the semantic embeddings of human/object labels). Various approaches [127, 219, 220, 229, 234, 235] further capitalize on message passing mechanism to perform relational reasoning over instance-centric graph structure, aiming to enrich HOI features with global contextual information among human and object instances. The authors in [221] devise contextual attention mechanisms to facilitate the mining of contextual cues. Moreover, the information about human pose [227, 230, 236], body parts [237] or detailed 3D body shape [238] can also be exploited to enhance HOI feature representation. In [228, 239], additional knowledge from external sources and language domain are further exploited to boost HOI feature learning. Most recently, the ATL scheme [224] constructs the affordance feature bank across multiple HOI datasets and injects affordance feature into object representations when inferring interactions.

One-stage Approaches. The second category schemes mainly construct one-stage HOI detectors [128, 129, 130, 131, 132, 218, 222, 223] by directly predicting HOI triplets, which are potentially faster and simpler than two-stage HOI detectors. UnionDet [129] is one of the first attempts that directly detect the union regions of human-object pairs in a one-stage manner. Other schemes [218, 222] formulate HOI detection as a keypoint detection problem, and thus enable a one-stage solution for this task. Most recently, inspired by the success of Transformer-based object detectors (e.g., DETR [22]), there has been a steady momentum of breakthroughs that push the limits of HOI detection by using Transformer-style architecture. In particular, the authors in [132, 223] employ a single interaction Transformer decoder to predict a set of HOI triplets, and the whole architecture can be optimized in an end-to-end fashion with Hungarian loss. However, the authors in [128, 130] design two parallel Transformer decoders for detecting interactions and instances, and the outputs are further associated to produce final HOI predictions.

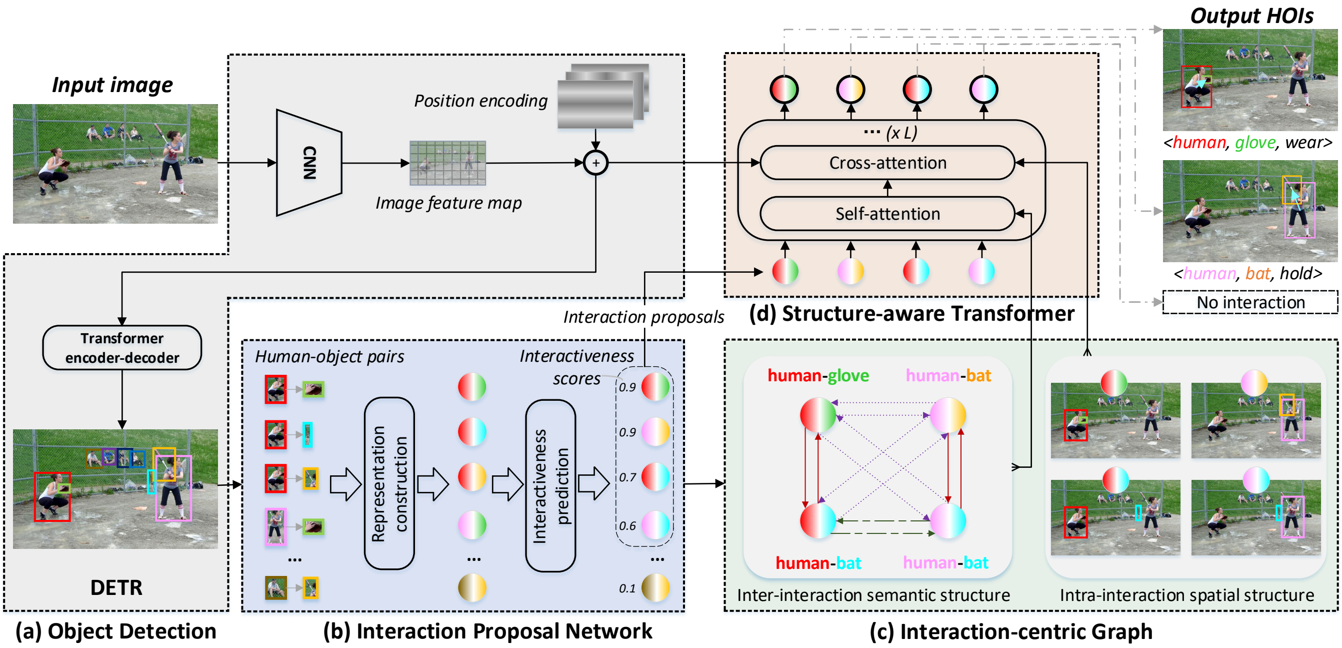


Figure 5.2: An overview of our proposed STIP framework. (a) Given an input image, we adopt an off-the-shelf DETR to detect the human and object instances within this image. (b) Based on the detected human and object instances, the Interaction Proposal Network (IPN) constructs all possible human-object pairs and then predicts the interactiveness score of each human-object pair. The most interactive human-object pairs with the highest interactiveness scores are taken as the output interaction proposals. (c) Next, by taking all interaction proposals as graph nodes and exploiting semantic connectivity as edges, we build an interaction-centric graph that unfolds rich inter-interaction semantic structure and intra-interaction spatial structure. (d) Finally, a structure-aware Transformer is utilized to transform the non-parametric interaction queries (i.e., interaction proposals) into a set of HOI predictions by additionally guiding relational reasoning with the inter- or intra-interaction structure derived from an interaction-centric graph.

This Scheme. The proposed STIP can also be considered as a Transformer-style architecture that tackles HOI detection as a set prediction problem, which eliminates the post-processing and enables the architecture to be end-to-end trainable. Unlike existing Transformer-style methods [128, 130, 132, 223] that perform HOI set prediction in a one-stage manner, the proposed STIP decomposes this process into two phases: the proposed scheme first produces interaction proposals as high-quality interaction queries and then takes them as non-parametric queries to trigger the HOI set prediction. Moreover, this STIP scheme goes beyond the conventional relational reasoning via vanilla Transformer by leveraging a structure-aware Transformer to exploit the rich inter- or intra-interaction structure, thereby boosting the performance of the HOI detection.

5.3 Approach

In this work, we devise the Structure-aware Transformer over Interaction Proposals (STIP) that casts HOI detection as a set prediction problem in a two-phase fashion. Meanwhile, this scheme boosts HOI set prediction with the prior knowledge of inter- and intra-interaction structures. Figure 5.2 depicts an overview of the proposed STIP. The whole framework consists of four main components, i.e., DETR for object detection, interaction proposal network for producing interaction proposals, interaction-centric graph construction, and structure-aware Transformer for HOI set prediction. Specifically, an off-the-shelf DETR [22] is first adopted to detect humans and objects within the input image. Next, based on the detection results, we design the Interaction Proposal Network (IPN) to select the most interactive human-object pairs as interaction proposals. After that, we take all selected interaction proposals as graph nodes to construct an interaction-centric graph to reveal the inter-interaction semantic structure and intra-interaction spatial structure. The selected interaction proposals are further taken as non-parametric queries to trigger the HOI set prediction via a structure-aware Transformer by exploiting the structured prior knowledge derived from an interaction-centric graph to strengthen relational reasoning.

5.3.1 Interaction Proposal Network

Conditioned on the detected human and object instances from DETR, the Interaction Proposal Network (IPN) targets for producing interaction proposals, i.e., the probably interactive human-object pairs. Concretely, we first construct all possible human-object pairs with pairwise connectivity between detected humans and objects. For each human-object pair, the IPN further predicts the probability of interaction existing in between (i.e., “interactiveness" score) through a multi-layer perceptron (MLP). Only the top-K human-object pairs with the highest interactiveness scores are finally emitted as the output interaction proposals.

Human-Object Pairs Construction. Here we connect each pair of detected human and object instances, yielding all possible human-object pairs within the input image. Each human-object pair can be represented from three perspectives, i.e., the appearance feature, spatial feature, and linguistic feature of human and object. In particular, the appearance feature is directly represented as the concatenation of human and object instance features derived from DETR (i.e., the 256-dimensional region feature before final prediction heads). By defining the normalized center coordinates of human and object bounding boxes as (c^h_x, c^h_y) and (c^o_x, c^o_y), we measure the spatial feature as the concatenation of all geometric properties, i.e., $[dx, dy, dis, arctan(\frac{dy}{dx}), A_h, A_o, I, U]$, where $dx = c^h_x - c^o_x, dy = c^h_y - c^o_y, dis = \sqrt{dx^2 + dy^2}$. A_h, A_o, I, U denote the areas of the human, object, their intersection, and union boxes, respectively. The linguistic feature is achieved by encoding the label name of the object (one-hot vector) into a 300-dimensional vector. The final representation of each human-object pair is calculated as the concatenation of appearance, spatial, and linguistic features.

Interactiveness Prediction. The interactiveness prediction module in IPN takes

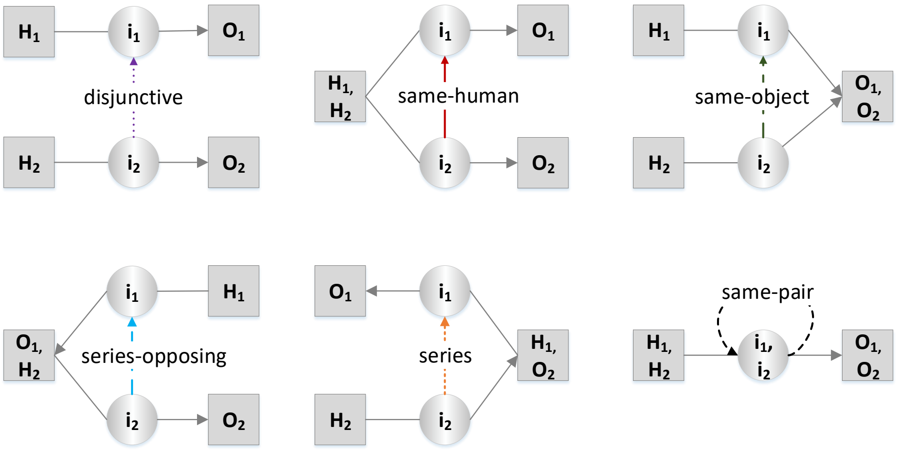


Figure 5.3: Definition of six kinds of inter-interaction semantic dependencies ⟨HOI(i_2)→HOI(i_1)⟩ between interaction HOI(i_1) and HOI(i_2) (square: human/object instance, circle: interaction).

the feature of each human-object pair as input, and learns to predict the probability of whether interactions exist between this pair, i.e., interactiveness score. We frame this sub-task of interactiveness prediction as a binary classification problem, and implement this module as MLP coupled with Sigmoid activation. During training, for each input image, we sample at most K human-object pairs, which consist of positive and negative pairs. Note that if both IoUs of predicted human and object bounding boxes in one human-object pair w.r.t ground-truths are larger than 0.5, we treat this pair as a positive sample, otherwise it is a negative sample. One natural way to fetch negative pairs is to use random sampling strategy. Instead, here we employ the hard mining strategy [240] to sample negative pairs with high predicted interactiveness scores, aiming to facilitate the learning of interactiveness prediction. After feeding all the N sampled human-object pairs in a mini-batch into the interactiveness prediction module, we optimize this module with focal loss [120] (FL):

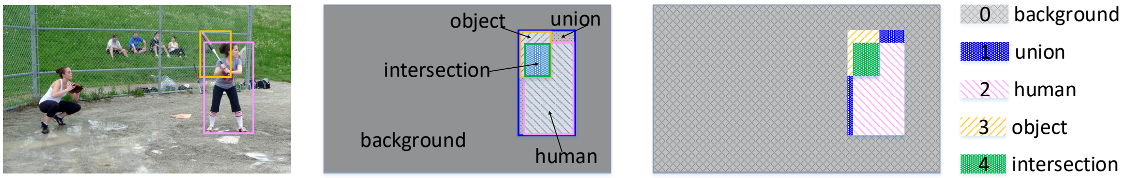
$$L_{proposal} = \frac{1}{\sum_{i=1}^{N} z_i} \sum_{i=1}^{N} FL(\hat{z}_i, z_i), \tag{5.1}$$

where $z_i \in \{0, 1\}$ indicates whether interactions exist in ground-truth and $\hat{z}_i$ is the predicted interactiveness score. At inference, only the top-K human-object pairs with the highest interactiveness scores are taken as interaction proposals.

5.3.2 Interaction-Centric Graph

Based on all the selected interaction proposals of each input image via IPN, we next present how to construct an interaction-centric graph that fully unfolds the rich prior knowledge of inter- and intra-interaction structures. Technically, we take each interaction proposal as one graph node, and the interaction-centric complete graph is thus built by densely connecting every two nodes as graph edges.

Inter-interaction Semantic Structure. Intuitively, there exists a natural semantic structure among interactions within the same image. For example, when we



(a) Interaction proposal (b) Spatial structure (c) Assigned label for each location

Figure 5.4: Definition of intra-interaction spatial structure for each interaction: (a) interaction proposal in an image; (b) the spatial structure, i.e., the layout of each component in this interaction; (c) the assigned label for each location in an image.

find the interaction of “human hold mouse" in an image, it is very likely that the mentioned “human” is associated with another interaction of “human look-at screen." This motivates us to exploit such common sense knowledge implied in the inter-interaction semantic structure to boost relational reasoning among interactions for HOI detection. Formally, we express the directional semantic connectivity as $\langle \mathrm{HOI}(i_2) \rightarrow \mathrm{HOI}(i_1) \rangle$, which denotes the relative semantic dependency of interaction proposal $\mathrm{HOI}(i_1)$ against interaction proposal $\mathrm{HOI}(i_2)$. Six kinds of inter-interaction semantic dependencies are thus defined according to whether two interaction proposals share the same human or object instance, as shown in Figure 5.3.

Concretely, if $\mathrm{HOI}(i_1)$ and $\mathrm{HOI}(i_2)$ do not share any human/object instance, we classify their dependency as “*disjunctive*" (class 0). If $\mathrm{HOI}(i_1)$ and $\mathrm{HOI}(i_2)$ only share the same human/object instance, we set the label of dependency as “*same-human*" (class 1) or “*same-object*" (class 2). When the human/object instance of $\mathrm{HOI}(i_1)$ is exactly the object/human instance of $\mathrm{HOI}(i_2)$, the dependency is classified as “*series-opposing*" (class 3) and “*series*" (class 4), respectively. If both of the human and object instances of $\mathrm{HOI}(i_1)$ and $\mathrm{HOI}(i_2)$ are same, the label of this dependency is “*same-pair*" (class 5).

Intra-interaction Spatial Structure. The inter-interaction semantic structure over the whole interaction-centric graph only unfolds the holistically semantic dependencies across all interaction proposals, while leaving the local spatial structure of human/object within each interaction proposal unexploited. Therefore, we characterize each graph node with an intra-interaction spatial structure, which can be interpreted as the layout of each component in the corresponding interaction proposal (see Figure 5.4). Specifically, we first identify the spatial location of each component (i.e., *background*, *union*, *human*, *object*, and *intersection*) for this interaction over the whole image, and then assign layout label $l_{ij} \in \{0, 1, 2, 3, 4\}$ to each location in this image according to the corresponding component.

5.3.3 Structure-aware Transformer

With the K interaction proposals and the interaction-centric graph, we next present how to integrate the prior knowledge of inter- and intra-interaction structures into relational reasoning for HOI set prediction in STIP. In particular, a structure-aware Transformer is devised to contextually encode all interaction proposals with the additional guidance of inter- and intra-interaction structures via structure-aware self-attention and cross-attention modules, yielding structure-aware HOI features for predicting HOI triplets.

Preliminary. We first briefly recall the vanilla Transformer that capitalizes on the attention mechanism, which aims to transform a sequence of queries $\boldsymbol{q} = (\boldsymbol{q}_1, ..., \boldsymbol{q}_m)$ plus a set of key-value pairs $(\boldsymbol{k} = (\boldsymbol{k}_1, ..., \boldsymbol{k}_n), \boldsymbol{v} = (\boldsymbol{v}_1, ..., \boldsymbol{v}_n))$ into the output sequence $\boldsymbol{o} = (\boldsymbol{o}_1, ..., \boldsymbol{o}_m)$. Each output element $\boldsymbol{o}_i$ is computed by aggregating all values weighted with attention: $\boldsymbol{o}_i = \sum_j \alpha_{ij}(\boldsymbol{W}_v \boldsymbol{v}_j)$, where each attention weight α_{ij} is normalized with softmax ($\alpha_{ij} = \frac{exp(e_{ij})}{\sum_j exp(e_{ij})}$). Here the primary attention weight e_{ij} is measured as the scaled dot-product between each key $\boldsymbol{k}_j$ and query $\boldsymbol{q}_i$:

$$e_{ij} = \frac{(\boldsymbol{W}_q \boldsymbol{q}_i)^T (\boldsymbol{W}_k \boldsymbol{k}_j)}{\sqrt{d_{key}}}. \tag{5.2}$$

Note that d_{key} is the dimension of keys, and $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v$ are learnable embedding matrices.

Structure-aware Self-attention. Existing Transformer-type HOI detectors perform relational reasoning among interactions via the self-attention module in vanilla Transformer for HOI set prediction. However, the relational reasoning process in vanilla Transformer is triggered by the parametric interaction queries, and leaves the prior knowledge of inter-interaction structure under-exploited. As an alternative, our structure-aware Transformer starts HOI set prediction from the non-parametric queries (i.e., the selected interaction proposals), and further upgrades the conventional relation reasoning with inter-interaction semantic structure through the structure-aware self-attention module.

Specifically, by taking the K interaction proposals $\boldsymbol{q}$ as interaction queries, keys, and values, the structure-aware self-attention module conducts the inter-interaction structure-aware reasoning among interactions to strengthen the HOI representation of each interaction. Inspired by relative position encoding in [241], we supplement each key $\boldsymbol{q}_j$ with the encodings of its inter-interaction semantic dependency with regard to query $\boldsymbol{q}_i$, which is measured as the concatenation of $\boldsymbol{q}_j$ and the corresponding semantic dependency label $d_{ij} \in \{0, 1..., 5\}$. In this way, we incorporate the inter-interaction semantic structure into the learning of attention weight by modifying Eq. (5.2) as:

$$e_{ij}^{self} = \frac{(\boldsymbol{W}_q \boldsymbol{q}_i)^T (\boldsymbol{W}_k \boldsymbol{q}_j + \boldsymbol{\psi}(\boldsymbol{q}_j, \boldsymbol{E}_{dep}(d_{ij})))}{\sqrt{d_{key}}}, \tag{5.3}$$

where $\boldsymbol{E}_{dep}$ denotes the embedding matrix of semantic dependency label and $\boldsymbol{\psi}$ is implemented as a 2-layer MLP to encode the inter-interaction semantic dependency. Accordingly, the output intermediate HOI features $\hat{\boldsymbol{q}}$ of structure-aware self-attention module are endowed with the holistically semantic structure among interactions.

Structure-aware Cross-attention. Next, based on the intermediate HOI features $\hat{\boldsymbol{q}}$, a structure-aware cross-attention module (see Figure 5.5) is utilized to further enhance HOI features by exploiting contextual information between interactions and the original image feature map in DETR. Formally, we take the K intermediate HOI features $\hat{\boldsymbol{q}} = (\hat{\boldsymbol{q}}_1, ..., \hat{\boldsymbol{q}}_K)$ as queries, and the image feature map $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ as keys

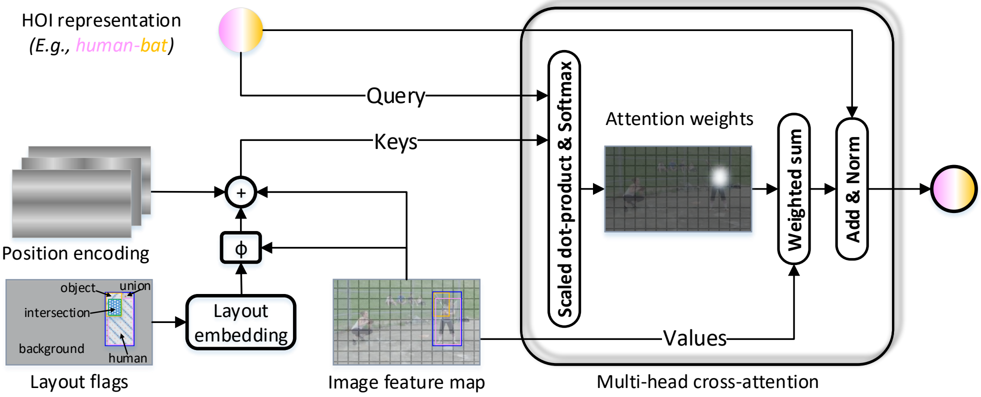


Figure 5.5: Structure-aware cross-attention module.

and values. The structure-aware cross-attention module performs intra-interaction structure-aware reasoning over the image feature map to strengthen the HOI feature of each interaction. Similar to the structure-aware self-attention module, each key $\boldsymbol{x}_j$ is supplemented with the encodings of the intra-interaction spatial structure with regard to query $\hat{\boldsymbol{q}}_i$ (i.e., the concatenation of $\boldsymbol{x}_j$ and its assigned layout label $l_{ij} \in \{0, 1, 2, 3, 4\}$). The learning of attention weight in structure-aware cross-attention module is thus integrated with the intra-interaction spatial structure, which is measured as:

$$e_{ij}^{cross} = \frac{(\boldsymbol{W}_{\hat{q}}\hat{\boldsymbol{q}}_i)^T(\boldsymbol{W}_{\hat{k}}\boldsymbol{x}_j + \boldsymbol{pos}_j + \boldsymbol{\phi}(\boldsymbol{x}_j, \boldsymbol{E}_{lay}(l_{ij})))}{\sqrt{d_{key}}}, \tag{5.4}$$

where $\boldsymbol{pos}_j$ is the position encoding, $\boldsymbol{E}_{lay}$ is the embedding matrix of the layout label, and we implement $\boldsymbol{\phi}$ as a 2-layer MLP to encode the intra-interaction spatial structure.

5.3.4 Training Objective

During training, we feed the final output HOI representations of the structure-aware Transformer into the interaction classifier (implemented as a 2-layer MLP) to predict the interaction classes of each interaction proposal. The objective of interaction classification is measured via focal loss:

$$L_{cls} = \frac{1}{\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}} \sum_{i=1}^{N}\sum_{c=1}^{C} FL(\hat{y}_{ic}, y_{ic}), \tag{5.5}$$

where C is the number of interaction classes, $y_{ic} \in \{0, 1\}$ indicates whether the labels of i-th proposal contain the c-th interaction class, and $\hat{y}_{ic}$ is the predicted probability of c-th interaction class. Accordingly, the overall objective of our STIP integrates the interactiveness prediction objective in Eq. (5.1) and interaction classification objective in Eq. (5.5):

$$L_{STIP} = L_{proposal} + L_{cls}. \tag{5.6}$$

5.4 Experiments

In this section, we empirically evaluate our STIP on two widely adopted HOI detection benchmarks, i.e., V-COCO [211] and HICO-DET [126].

5.4.1 Datasets and Experimental Settings

V-COCO is a popular dataset for benchmarking HOI detection, which is a subset of MS-COCO [242] covering 29 action categories. This dataset consists of 2,533 training images, 2,867 validation images, and 4,946 testing images. Following the settings in [130], we adopt Average Precision (AP_{role}) over 25 interactions as the evaluation metric. Two kinds of AP_{role}, i.e., $AP_{role}^{\#1}$ and $AP_{role}^{\#2}$, are reported under two scenarios with different scoring criteria for object occlusion cases. Specifically, in the scenario of $AP_{role}^{\#1}$, the model should manage to infer the occluded object correctly by predicting the 2D location of its bounding box as [0,0,0,0], meanwhile precisely localizing the corresponding human bounding box and recognizing the interaction in between. In contrast, for the scenario of $AP_{role}^{\#2}$, there is no need to infer the occluded object.

HICO-DET is a larger HOI detection benchmark, which contains 37,536 and 9,515 images for training and testing, respectively. The whole dataset covers 600 categories of $\langle human, object, interaction\rangle$ triplets, covering the same 80 object categories as in MS-COCO [242] and 117 verb categories. We follow [126] and report mAP in two different settings ($Default$ and $Known\, Object$). Here the $Default$ setting represents that the mAP is calculated over all testing images, while $Known\, Object$ measures the AP of each object solely over the images containing that object class. For each setting, we report the AP over three different HOI category sets, i.e., **Full** (all 600 HOI categories), **Rare** (138 HOI categories where each one contains less than 10 training samples), and **Non-Rare** (462 HOI categories where each one contains 10 or more training samples).

Implementation Details. For fair comparison with state-of-the-art baselines, we adopt the same object detector DETR pre-trained over MS-COCO (backbone: ResNet-50) and all learnable parameters in DETR are frozen during training as in [130]. On the HICO-DET dataset, we additionally report the results by fine-tuning DETR on HICO-DET and the performances by further jointly fine-tuning the object detector and HOI detector. In the experiments, we select the top-32 probably interactive human-object pairs as the output interaction proposals of the Interaction Proposal Network. Our proposed structure-aware Transformer consists of 6 stacked layers (structure-aware self-attention plus cross-attention modules). The whole architecture is trained over 2 Nvidia 2080ti GPUs with the AdamW optimizer. The mini-batch size is 8 and we set the initial learning rate as 5×10^{-5}. The maximum training epoch number is 30.

5.4.2 Performance Comparisons

V-COCO. Table 5.1 summarizes the performance comparisons in terms of $AP_{role}^{\#1}$ and $AP_{role}^{\#2}$ on V-COCO. In general, the results across all metrics under the same back-

Method	Backbone	Feature	$AP_{role}^{\#1}$	$AP_{role}^{\#2}$
One-stage methods				
UnionDet [129]	R50	A	47.5	56.2
IPNet [222]	HG-104	A	51.0	-
GGNet [131]	HG-104	A	54.7	-
HOITrans [132]	R50	A	52.9	-
AS-Net [128]	R50	A	53.9	-
HOTR [130]	R50	A	55.2	64.4
QPIC [223]	R50	A	58.8	61.0
Two-stage methods				
InteractNet [216]	R50-FPN	A	40.0	48.0
GPNN [219]	R101	A	44.0	-
TIN [227]	R50	A+S+P	48.7	-
DRG [127]	R50-FPN	A+S+L	51.0	-
FCMNet [217]	R50	A+S+L+P	53.1	-
ConsNet [228]	R50-FPN	A+S+L	53.2	-
IDN [226]	R50	A+S	53.3	60.3
STIP (Ours)	R50	A	**65.1**	**69.7**
STIP (Ours)	R50	A+S+L	**66.0**	**70.7**

Table 5.1: Performance comparison on V-COCO dataset. The letters in the Feature column indicate the input features: **A** (Appearance/Visual features), **S** (Spatial features [215]), **L** (Linguistic feature of label semantic embeddings), **P** (Human pose feature).

bone (ResNet-50, R50 in short) consistently demonstrate that our STIP exhibits better performances against existing techniques, including both one-stage methods (e.g., UnionDet, AS-Net, HOTR, and QPIC) and two-stage methods (e.g., FCMNet, ConsNet, and IDN). The results generally highlight the key advantage of two-phase HOI set prediction and the exploitation of inter- and intra-interaction structures. In particular, the conventional two-stage HOI detectors (e.g., GPNN, TIN, DRG) commonly construct instance-centric graphs to mine contextual information among instances. Instead, recent Transformer-style HOI detectors (e.g., HOITrans, AS-Net, HOTR, QPIC) fully capitalize on a vanilla Transformer to perform relational reasoning among instances/interactions, thereby leading to performance boosts. However, when only using appearance features (A), the $AP_{role}^{\#1}$ and $AP_{role}^{\#2}$ of HOTR and QPIC are still lower than our STIP, which not only takes interaction proposals as non-parametric interaction queries to trigger HOI set prediction, but also leverages a structure-aware Transformer to exploit the prior knowledge of inter-interaction and intra-interaction structures. For our STIP, further performance improvement is attained when utilizing more kinds of features (e.g., spatial and linguistic features).

HICO-DET. We further evaluate our STIP on the more challenging HICO-DET dataset. Table 5.2 reports the mAP scores over three different HOI category sets for each setting (Default/Known Object) in comparison with the state-of-the-art methods. Here we include three different training settings, i.e., pre-train object detector on MS-COCO, fine-tune object detector on HICO-DET, and jointly fine-tune object detector and HOI detector on HICO-DET, for a fair comparison. Similar to the observations on V-COCO, our STIP achieves consistent performance gains against existing HOI detectors across all the metrics for each training setting. The results basically demonstrate

Method	Backbone	Feature	Default			Known Object		
			Full	Rare	Non-Rare	Full	Rare	Non-Rare
Object detector pre-trained on MS-COCO								
InteractNet [216]	R50-FPN	A	9.94	7.16	10.77	-	-	-
GPNN [219]	R101	A	13.11	9.41	14.23	-	-	-
UnionDet [129]	R50	A	14.25	10.23	15.46	19.76	14.68	21.27
TIN [227]	R50	A+S+P	17.22	13.51	18.32	19.38	15.38	20.57
IPNet [222]	R50-FPN	A	19.56	12.79	21.58	22.05	15.77	23.92
DRG [127]	R50-FPN	A+S+L	19.26	17.74	19.71	23.40	21.75	23.89
FCMNet [217]	R50	A+S+L+P	20.41	17.34	21.56	22.04	18.97	23.12
ConsNet [228]	R50-FPN	A+S+L	22.15	17.12	23.65	-	-	-
IDN [226]	R50	A+S	23.36	22.47	23.63	26.43	25.01	26.85
HOTR [130]	R50	A	23.46	16.21	25.60	-	-	-
AS-Net [128]	R50	A	24.40	22.39	25.01	27.41	25.44	28.00
STIP (Ours)	R50	A	28.11	25.85	28.78	31.23	27.93	32.22
STIP (Ours)	R50	A+S+L	**28.81**	**27.55**	**29.18**	**32.28**	**31.07**	**32.64**
Object detector fine-tuned on HICO-DET								
DRG [127]	R50-FPN	A+S+L	24.53	19.47	26.04	27.98	23.11	29.43
ConsNet [228]	R50-FPN	A+S+L	24.39	17.10	26.56	-	-	-
IDN [226]	R50	A+S	26.29	22.61	27.39	28.24	24.47	29.37
HOTR [130]	R50	A	25.10	17.34	27.42	-	-	-
STIP (Ours)	R50	A	29.76	26.94	30.61	32.84	29.05	33.85
STIP (Ours)	R50	A+S+L	**30.56**	**28.15**	**31.28**	**33.54**	**30.93**	**34.31**
Jointly fine-tune object detector & HOI detector on HICO-DET								
UnionDet [129]	R50	A	17.58	11.72	19.33	19.76	14.68	21.27
PPDM [218]	HG104	A	21.73	13.78	24.10	24.58	16.65	26.84
GGNet [131]	HG104	A	29.17	22.13	30.84	33.50	26.67	34.89
HOITrans [132]	R50	A	23.46	16.91	25.41	26.15	19.24	28.22
AS-Net [128]	R50	A	28.87	24.25	30.25	31.74	27.07	33.14
QPIC [223]	R50	A	29.07	21.85	31.23	31.68	24.14	33.93
STIP (Ours)	R50	A	31.60	27.75	32.75	34.41	30.12	35.69
STIP (Ours)	R50	A+S+L	**32.22**	**28.15**	**33.43**	**35.29**	**31.43**	**36.45**

Table 5.2: Performance comparison on HICO-DET dataset. The letters in Feature column indicate the input features: **A** (Appearance/Visual features), **S** (Spatial features [215]), **L** (Linguistic feature of label semantic embeddings), **P** (Human pose feature).

the advantage of triggering HOI set prediction with the non-parametric interaction proposals and meanwhile exploiting the holistically semantic structure among interaction proposals & the local spatial structure within each interaction proposal.

5.4.3 Experimental Analysis

Ablation Study. To examine the impact of each design in STIP, we conduct ablation studies by comparing different variants of STIP on V-COCO and HICO-DET datasets in Table 5.3. Note that all experiments on HICO-DET here are conducted under the training setting of object detector fine-tuned on HICO-DET. We start from the basic model (**Base**), which utilizes a basic interaction proposal network (randomly sampling negative samples for training, without the hard mining strategy). The generated interaction proposals in Base model are directly leveraged for interaction classification, without any Transformer-style structure for boosting HOI prediction. Next, we extend Base model by leveraging the hard mining strategy to select the hard negative human-object pairs with higher interactiveness scores for training interaction proposal

Method	V-COCO		HICO-DET (Default)		
	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$	Full	Rare	Non-Rare
Base	52.49	58.25	21.74	18.09	22.83
+HM	58.45	62.64	24.16	19.45	25.57
+HM+TR	63.50	68.07	28.62	26.09	29.38
+HM+TR^{SS}	64.99	69.94	29.65	26.52	30.59
+HM+TR^{SC}	65.04	69.76	29.74	27.07	30.54
+HM+TR^{SS+SC} (**STIP**)	**66.04**	**70.65**	**30.56**	**28.15**	**31.28**

Table 5.3: Performance contribution of each component in our STIP. **HM**: Hard Mining strategy for training interaction proposal network. **TR**: vanilla TRansformer. **TR^{SS}**: TRansformer with only Structure-aware Self-attention that exploits inter-interaction structure. **TR^{SC}**: TRansformer with only Structure-aware Cross-attention that exploits intra-interaction structure.

# of selected interaction proposals (K)	V-COCO		HICO-DET (Default)		
	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$	Full	Rare	Non-Rare
8	64.20	69.11	29.03	28.16	29.29
16	65.68	70.63	30.18	28.66	30.64
<u>32</u>	**66.04**	**70.65**	30.56	28.15	**31.28**
64	65.93	70.50	**30.72**	**28.96**	31.24
100	65.78	70.45	30.40	27.89	31.14

Table 5.4: Performance comparison by using different numbers of selected interaction proposals (K) for interaction-centric graph construction in our STIP.

network, yielding **Base+HM** which achieves better performances. After that, by additionally involving a vanilla Transformer to perform relational reasoning among interaction proposals, another variant of our model (**Base+HM+TR**) leads to performance improvements across all metrics. Furthermore, we upgrade the vanilla Transformer with structure-aware self-attention that exploits the holistically semantic structure among interaction proposals, and this ablated run (**Base+HM+TR^{SS}**) outperforms Base+HM+TR. Meanwhile, the vanilla Transformer can be upgraded with structure-aware cross-attention that exploits the locally spatial structure within each interaction proposal, and **Base+HM+TR^{SC}** also exhibits better performances. These observations basically validate the merit of exploiting the structured prior knowledge, i.e., inter-interaction or intra-interaction structure, for HOI detection. Finally, when jointly upgrading the vanilla Transformer with structure-aware self-attention and structure-aware cross-attention (i.e., our **STIP**), the highest performances are attained.

Effect of Selected Interaction Proposal Number K for Interaction-centric Graph Construction. Recall that the interaction proposal network in our STIP selects only the top-K human-object pairs with the highest interactiveness scores as the output interaction proposals for constructing the interaction-centric graph. Such K selected interaction proposals are also taken as non-parametric interaction queries to trigger HOI set prediction in the structure-aware Transformer. Here we vary K from 8 to 100 to explore the relationship between the performance and the select interaction proposal number K. As shown in Table 5.4, the best performances across most met-

# of layers	V-COCO		HICO-DET (Default)		
(L)	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$	Full	Rare	Non-Rare
0	58.45	62.64	24.16	19.45	25.57
1	64.83	69.57	29.21	26.37	30.06
2	65.55	70.39	30.02	28.11	30.59
4	66.02	70.61	30.47	29.28	30.83
<u>6</u>	**66.04**	**70.65**	30.56	28.15	**31.28**
8	65.44	70.11	**30.93**	**29.78**	31.27

Table 5.5: Performance comparison with different layer numbers of the structure-aware Transformer in our STIP.

rics are attained when K is set as 32. In particular, enlarging the number of selected interaction proposals (until $K = 32$) can generally lead to performance boosts on two datasets. Once K is larger than 64, the performances slightly decrease. We speculate that the increase in selected interaction proposals results in more invalid proposals, which may affect the overall stability of relational reasoning among interaction proposals. Accordingly, we empirically set the number of selected interaction proposals K as 32.

Effect of Layer Number L in Structure-aware Transformer. To explore the effect of layer number L in the structure-aware Transformer, we show the performances on two benchmarks by varying this number from 0 to 8. The best performances across most metrics are achieved when the layer number is set to $L = 6$. Specifically, in the extreme case of $L = 0$, no self-attention and cross-attention module is utilized, and the model degenerates to a Base+HM model that directly performs interaction classification over interaction proposals without any relational reasoning via Transformer-style structure. When increasing the layer number in the structure-aware Transformer, the performances are gradually increased in general. This basically validates the effectiveness of enabling relational reasoning among interaction proposals through the structure-aware Transformer. In practice, the layer number L is generally set to 6.

Time Analysis. We evaluate the inference time of our STIP on a single Nvidia 2080ti GPU by constructing each batch with a single testing image. Specifically, for each input batch, object detection via DETR, interaction proposal generation through interaction proposal network, HOI set prediction with structure-aware Transformer, and the other processing (e.g., data loading) takes 41.9ms, 7.8ms, 20.4ms, and 3.8ms, respectively. Consequently, the overall inference stage of STIP finishes in 73.9ms on average, which is comparable to existing one-stage Transformer-style HOI detectors (e.g., the inference time of AS-Net [128] is 71ms).

5.4.4 Additional Analysis

Training Time Analysis. In this part, we include the detailed training time comparison between our proposed STIP and several existing Transformer-style HOI detectors [128, 130, 132, 223]. As shown in Table 5.6, existing Transformer-style HOI detectors commonly suffer from slow convergence as in DETR [22], which usually require over 100 training epochs. We speculate that this may be the result of HOI set prediction driven by the parametric interaction queries with randomly initialized embeddings.

Method	Batch Size	Epochs (HICO-DET dataset)
AS-Net [128]	64	90
HOTR [130]	16	100
QPIC [223]	16	150
HOITrans [132]	16	250
STIP (Ours)	8	30

Table 5.6: Training time comparison among Transformer-style HOI detectors. All runs are trained with the same backbone (ResNet-50) and optimizer (AdamW) for a fair comparison.

Feature	V-COCO		HICO-DET (Default)		
	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$	Full	Rare	Non-Rare
A	65.09	69.66	29.76	26.94	30.61
A+S	65.64	70.47	30.31	27.85	31.04
A+L	65.60	70.33	30.01	26.50	31.06
A+S+L	**66.04**	**70.65**	**30.56**	**28.15**	**31.28**

Table 5.7: An ablation study on the use of different HOI features. The letters in Feature column indicate the input features: **A** (Appearance/Visual features), **S** (Spatial features), **L** (Linguistic feature of label semantic embeddings).

Instead, our STIP starts HOI set prediction from non-parametric interaction queries (i.e., high-quality interaction proposals), thereby leading to more efficient Transformer learning with only 30 epochs.

Ablation Study on HOI Features. Here we conduct additional ablation studies to examine how HOI detection performance is affected when capitalizing on different types of HOI feature representations in our Interaction Proposal Network. Table 5.7 details the performances by exploiting different HOI features in STIP. In particular, the use of only appearance feature (A) in general achieves superior performances by clearly outperforming existing HOI detectors (see Table 5.1 and 5.2). As expected, by integrating the appearance feature with additional spatial or linguistic features, consistent performance gains are attained. The results indicate that both spatial and linguistic cues are complements to the visual appearance cues of humans and objects. Finally, the combination of all three types of HOI features reaches the highest performance, which basically demonstrates the complementarity in between.

Cross-attention Visualization. In order to better qualitatively evaluate the structure-aware cross-attention module in our structure-aware Transformer, we visualize the attended image regions according to the learned cross-attention weights for HOTR [130], Base+HM+TR, and our STIP in Figure 5.6. Note that Base+HM+TR is a degraded version of our STIP by using a vanilla Transformer for HOI set prediction. Specifically, as shown in Figure 5.6 (a-1) and (a-2), both HOTR and Base+HM+TR always focus on similar regions even when predicting different human-object pairs for the same image. We speculate that this may be the result of solely performing cross-attention learning in vanilla Transformer without any prior knowledge, where the estimated cross-attention can be easily overwhelmed with the inherent salient regions in images. As an alternative, our structure-aware Transformer in STIP facilitates cross-attention learning with the additional guidance of intra-interaction structure, and thus accurately steers cross-attention over image areas for depicting the target interaction (see Figure 5.6 (a-3)). Similarly, in Figure 5.6 (b)-(e), our STIP manages to attend over the relevant regions for recognizing the corresponding target interactions.

5.5 Conclusion and Discussion

In this chapter, we have presented STIP, a new end-to-end Transformer-style model for human-object interaction detection. Instead of performing HOI set prediction derived from parametric interaction queries in a one-stage manner, the proposed STIP capitalizes on a two-phase solution for HOI detection by first producing interaction proposals and then taking them as non-parametric interaction queries to trigger HOI set prediction. Furthermore, by going beyond the commonly adopted vanilla Transformer, a novel structure-aware Transformer is designed to exploit two kinds of structured prior knowledge, i.e., inter- and intra-interaction structures, to further boost HOI set prediction. We validate the proposed scheme and analysis through extensive experiments conducted on V-COCO and HICO-DET datasets. More importantly, the proposed STIP achieves new state-of-the-art results on both benchmarks.

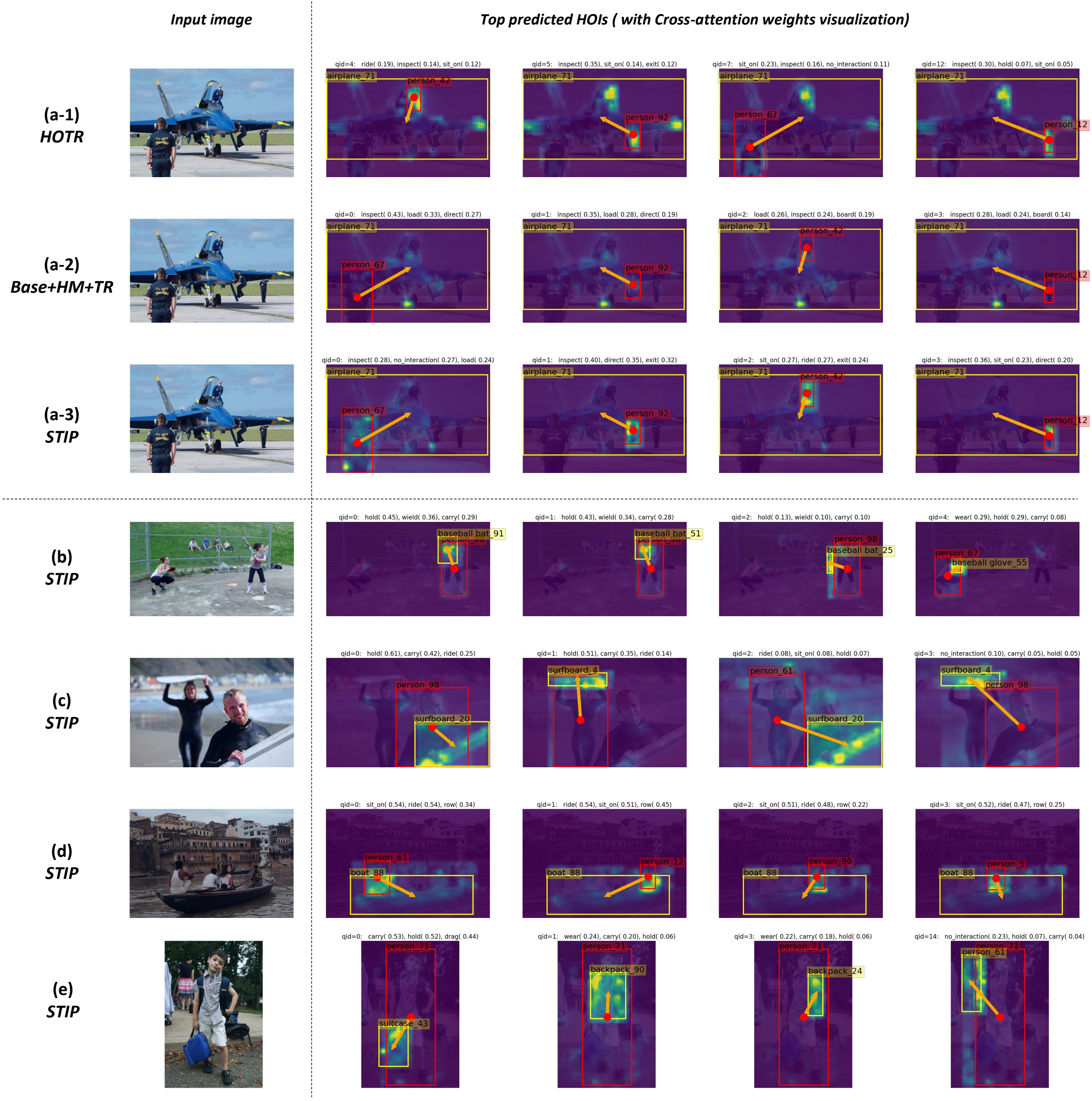

Input image
Top predicted HOIs (with Cross-attention weights visualization)
(a-1)
HOTR
qid=4: ride(0.19), inspect(0.14), sit_on(0.12)
airplane_71
person_42
qid=5: inspect(0.35), sit_on(0.14), exit(0.12)
airplane_71
person_92
qid=7: sit_on(0.23), inspect(0.16), no_interaction(0.11)
airplane_71
person_67
qid=12: inspect(0.30), hold(0.07), sit_on(0.05)
airplane_71
person_12
(a-2)
Base+HM+TR
qid=0: inspect(0.43), load(0.33), direct(0.27)
airplane_71
person_67
qid=1: inspect(0.35), load(0.28), direct(0.19)
airplane_71
person_92
qid=2: load(0.26), inspect(0.24), board(0.19)
airplane_71
person_42
qid=3: inspect(0.28), load(0.24), board(0.14)
airplane_71
person_12
(a-3)
STIP
qid=0: inspect(0.28), no_interaction(0.27), load(0.24)
airplane_71
person_67
qid=1: inspect(0.40), direct(0.35), exit(0.32)
airplane_71
person_92
qid=2: sit_on(0.27), ride(0.27), exit(0.24)
airplane_71
person_42
qid=3: inspect(0.36), sit_on(0.23), direct(0.20)
airplane_71
person_12
(b)
STIP
qid=0: hold(0.45), wield(0.36), carry(0.29)
baseball bat_91
person_98
qid=1: hold(0.43), wield(0.34), carry(0.28)
baseball bat_51
person_98
qid=2: hold(0.13), wield(0.10), carry(0.10)
person_98
baseball bat_25
qid=4: wear(0.29), hold(0.29), carry(0.08)
person_67
baseball glove_55
(c)
STIP
qid=0: hold(0.61), carry(0.42), ride(0.25)
person_98
surfboard_20
qid=1: hold(0.51), carry(0.35), ride(0.14)
surfboard_4
qid=2: ride(0.08), sit_on(0.08), hold(0.07)
person_61
surfboard_20
qid=3: no_interaction(0.10), carry(0.05), hold(0.05)
surfboard_4
person_98
(d)
STIP
qid=0: sit_on(0.54), ride(0.54), row(0.34)
person_61
boat_88
qid=1: ride(0.54), sit_on(0.51), row(0.45)
boat_88
person_12
qid=2: sit_on(0.51), ride(0.48), row(0.22)
boat_88
person_90
qid=3: sit_on(0.52), ride(0.47), row(0.25)
boat_88
person_5
(e)
STIP
qid=0: carry(0.53), hold(0.52), drag(0.44)
person_71
suitcase_43
qid=1: wear(0.24), carry(0.20), hold(0.06)
person_71
backpack_90
qid=3: wear(0.22), carry(0.18), hold(0.06)
person_71
backpack_24
qid=14: no_interaction(0.23), hold(0.07), carry(0.04)
person_71
person_61

Figure 5.6: The cross-attention visualization of testing samples on the HICO-DET dataset. We highlight the attended image regions according to the cross-attention weights in the last layer of Transformer. (a-1), (a-2), and (a-3) show the cross-attention visualization results of the same image for HOTR [130], Base+HM+TR and our proposed STIP (using the same pre-trained DETR). (b)-(e) showcase more cross-attention visualization results of our STIP. Particularly, for each image, we present the top-4 predicted human-object pairs. For each human-object pair (identified by 'qid'), we also list the top-3 predicted interactions and their classification scores on the top of each image.

Chapter 6

Video Scene Graph Generation using Spatio-temporal Transformer

Our previous works study the image SGG task. In this chapter[1], we tackle the challenging video SGG task by proposing a novel end-to-end Transformer-based architecture.

Video SGG has been an emerging research topic, which aims to interpret a video as a temporally-evolving graph structure by representing video objects as nodes and their relations as edges. Existing approaches predominantly follow a multi-step scheme, including frame-level object detection, relation recognition and temporal association. Although effective, these approaches neglect the mutual interactions between independent steps, resulting in a sub-optimal solution. We present a novel end-to-end framework for video scene graph generation, which naturally unifies object detection, object tracking, and relation recognition via a new Transformer structure, namely Temporal Propagation Transformer (TPT). Particularly, TPT extends the existing Transformer-based object detector (e.g., DETR) along the temporal dimension by involving a query propagation module, which can additionally associate the detected instances by identities across frames. A temporal dynamics encoder is then leveraged to dynamically enrich the features of the detected instances for relation recognition by attending to their historic states in previous frames. Meanwhile, the relation propagation strategy is devised to emphasize the temporal consistency of relation recognition results among adjacent frames. Extensive experiments conducted on VidHOI and Action Genome benchmarks demonstrate the superior performance of the proposed TPT over the state-of-the-art methods.

6.1 Introduction

Scene graph generation (SGG) is one of the fundamental problems in multimedia and computer vision fields, which aims to describe unstructured natural images/videos in the form of abstract graph structure (namely scene graph). Each node in the scene

[1] ***Yong Zhang**, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. "End-to-End Video Scene Graph Generation with Temporal Propagation Transformer.", IEEE Transactions on Multimedia (TMM), 2023.*

graph denotes a target object instance, while an edge between two nodes reflects their visual relations. Recent advances in deep learning have successfully pushed the limits of SGG in images (i.e., image SGG) [42, 62, 63, 74, 134, 243, 244]. The learned high-quality scene graphs in images benefit a series of downstream vision applications, ranging from image retrieval [4], image captioning [9, 10], visual question answering [16, 38] to image generation [95]. Nevertheless, the extension of such SGG from image to video is not trivial, since a video is an information-intensive media with complexities along both spatial and temporal dimensions, not to mention that the object instances might suffer from motion blur or occlusions. For the task of video SGG, a fundamental problem is how to exploit the spatio-temporal coherence across frames to boost the generation of a temporally-evolving graph structure.

Compared to image SGG, the task of video SGG is still far less explored by the research community. One straightforward way to tackle this task is to directly employ image-based SGG techniques [42, 62, 63, 134] over individual frames, as benchmarked in [54]. Nevertheless, this way neglects temporal context information among adjacent frames, which is crucial for relation recognition in videos. To alleviate this limitation, recent studies [100, 101, 102, 103, 104, 105, 245, 246, 247] specifically focus on how to exploit the spatio-temporal context within videos to boost video SGG. These methods can be categorized into two main directions, i.e., 3D-based methods [100, 101] and spatio-temporal contextualization methods [102, 103, 104, 105]. Generally, both directions are framed in a multi-step scheme. That is, they first perform frame-wise object detection by using an off-the-shelf pre-trained object detector (e.g., Faster R-CNN [21]). Next, the 3D-based methods [15], [16] (Figure 6.1 (a)) associate the detected instances corresponding to the identical visual target via tracking algorithms, and extract the pooled visual features along the trajectories for each target via a pre-trained 3D CNN backbone (e.g., I3D [248]). The relation recognition is then performed pair-wisely over the detected instances inside each frame. The video scene graph is finally generated based on the estimated relations. In contrast, the spatio-temporal contextualization methods (Figure 6.1 (b)) devise deliberate architectures like spatial-temporal Transformers [102, 103, 104] or graph neural networks [105] over the detected instances, to trigger the context mining both inside a single frame and across multiple consecutive frames. Moreover, a post-processing of object tracking is required to construct instance-level temporal associations. However, most of the existing techniques utilize specialized modules for object detection, relation recognition and instance temporal association independently, resulting in a separate training scheme. This inevitably breaks the cooperative interactions among these modules for different sub-tasks, and thus leads to a sub-optimal solution.

In this work, we propose a unified video SGG architecture (Figure 6.1 (c)), namely Temporal Propagation Transformer (TPT), which elegantly exploits crucial spatio-temporal contexts in videos for boosting video SGG in an end-to-end fashion. This design jointly tackles the three sub-tasks of object detection, instance temporal association and relation recognition in video SGG through a novel monolithic spatio-temporal Transformer architecture. Specifically, we firstly extend an off-the-shelf Transformer-based object detector (e.g., DETR [22, 24]) along the temporal dimension by equipping it with a shared query propagation module (QPM). QPM upgrades vanilla latent ob-

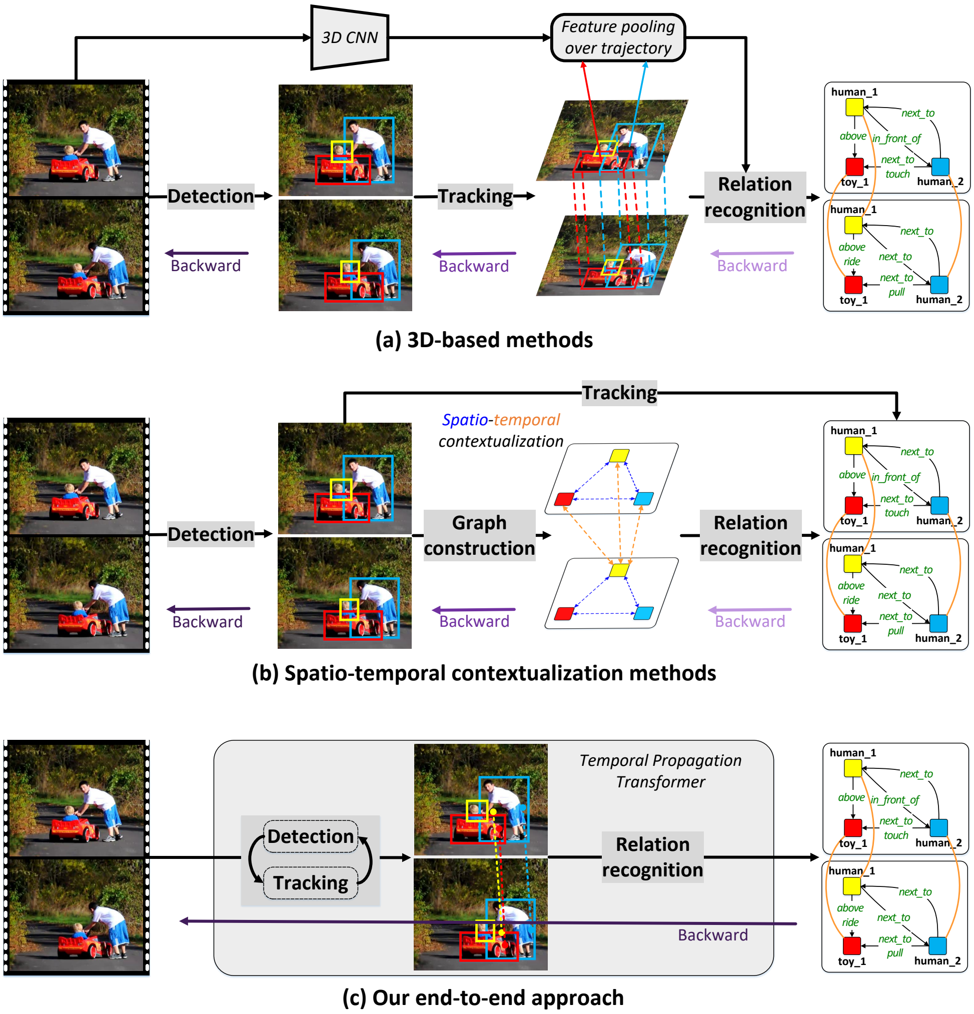


Figure 6.1: Comparison between existing video SGG techniques and our end-to-end solution. (a) **3D-based methods** typically obtain object detections via an off-the-shelf detector, and then associate detection results into trajectories via tracking. The temporal-aware visual features are thus extracted through 3D CNN backbones for relation recognition. (b) **Spatio-temporal contextualization methods** first achieve detection results similarly, and then construct a graph structure to mine contextual cues along both spatial and temporal dimensions to boost relation recognition. A post-processing of object tracking is further employed to associate instances temporally. (c) **Our end-to-end approach** directly predicts a temporally-evolving scene graph from the input video, by unifying object detection, tracking, and relation recognition into a monolithic framework.

ject queries and makes them aware of previous detection outputs, thereby enabling the joint detection and tracking of visual targets within a video. Next, conditioned on predicted object instances on each frame, a relation recognition module is designed in a classic two-stage paradigm, i.e., first producing relation proposals (i.e., the probably interactive object pairs) and then performing relation classification. More importantly, thanks to the inherently established temporal associations, TPT can easily explore the spatio-temporal coherence within videos to facilitate video SGG. Particularly, we devise a temporal dynamics encoder (TDE) to enrich object representations with their historic representations in previous frames via an attention mechanism. Moreover, we employ the relation propagation (RP) strategy to augment the set of current relation proposals with previous relation recognition results. The whole architecture of TPT can be jointly optimized in an end-to-end manner.

In summary, we have made the following contributions: (1) As far as we know, the proposed TPT is the first end-to-end Transformer-based framework for video SGG; (2) TPT elegantly associates video targets across frames via the latent object queries, which significantly ease the exploitation of spatio-temporal coherence in videos for facilitating video SGG; (3) We fully validate the effectiveness of TPT through extensive experiments on VidHOI and Action Genome datasets, and TPT has achieved state-of-the-art video SGG performances.

6.2 Related Work

Image Scene Graph Generation. The concept of scene graph representation is first introduced in [4]. Next, a series of innovations [42, 62, 63, 74, 134, 243, 244, 249] have been proposed for image SGG task. Most studies adopt a two-stage paradigm [42, 62, 63, 134, 243, 244] that decouples this task into sequential object detection and relation recognition. The sub-task of object detection is usually implemented with an off-the-shelf detector such as Faster R-CNN [21]. For relation recognition sub-task, they commonly exploit spatial context through message passing [42, 62, 63, 134, 244], or leverage external knowledge [250], contrastive loss [61] or casual inference [243], to strengthen relation recognition results. More recently, one-stage image SGG models [74, 77] based on Transformer structure are devised to pursue a joint optimization of object detection and relation recognition through an end-to-end learning philosophy. However, the extension of the end-to-end Transformer structure in an image to a video is very challenging due to the complex spatio-temporal context.

Video Scene Graph Generation. The task of video SGG is first proposed by [54], which has constructed a human-annotated dataset (i.e., Action Genome) for this task. Recently, a more challenging dataset (named VidHOI [100]) is built with dense spatio-temporal scene graph annotations by covering more diverse object and predicate categories. The videos in VidHOI are naturally captured in daily life that mainly focus on human actions. Existing video SGG techniques [100, 101, 102, 103, 104, 105, 245, 246, 247] generally follow a multi-stage pipeline that sequentially performs frame-wise object detection, tracking, and relation recognition. Since the former two sub-tasks are usually tackled via pre-trained modules, these works often pay more attention on improving relation recognition. Therefore the solutions of the video SGG task share a similar spirit with the approaches in video relation detection [251, 252] and video human-object interaction detection [100, 103, 219, 253]. Specifically, Ji et.al [54] benchmark video SGG by directly migrating image SGG techniques. To additionally exploit temporal information for relation recognition in videos, TRACE [101] and ST-HOI [100] propose to extract visual features from temporal-aware 3D CNNs, and ST-HOI [100] leverages extra human pose features. Similar to the practice in image

SGG task of mining semantic context via message passing, spatio-temporal Transformers [102, 103, 104] and graph neural networks [105, 219] are utilized to strengthen relation representations by fusing spatio-temporal context. Different from existing spatio-temporal Transformer methods [102, 103, 104] that only employ Transformer networks to refine relation feature in relation recognition sub-task, our TPT presents a principled end-to-end framework that capitalizes on a unified Transformer structure to perform each sub-task (i.e., object detection, tracking, and relation recognition) throughout the whole pipeline.

Transformer in Vision. In recent years, Transformer structure [40] has achieved revolutionary success in various computer vision tasks, such as image classification [50, 202], object detection [22, 24], human-object interaction detection [130, 132], video object tracking [254, 255, 256] and segmentation [28, 29]. Particularly, DETR [22, 24] presents an elegant end-to-end object detection approach that eliminates burdensome hand-designed components (e.g., anchor boxes, non-maximum suppression), by directly mapping learnable object queries to set predictions. Such a new object detection paradigm has inspired high-quality Transformer-based object tracking methods, rather than hinging on previous dominating *tracking-by-detection* paradigm [257, 258]. For example, Trackformer [255] and MOTR [254] extend object query mechanism in DETR along the temporal dimension, in which existing tracks are propagated frame-by-frame as additional track queries for tracking objects in a video. Similarly, here we also trigger instance temporal associations through latent object queries. However, instead of constructing additional track queries during propagation [254, 255], we maintain a fixed size of object queries, and make each query continuously track an object once it has detected the object. This leads to a simple yet effective way to exploit temporal consistency across frames for relation recognition.

6.3 Approach

In this chapter, we present a new unified video SGG framework, named Temporal Propagation Transformer (TPT), to jointly performs object detection, tracking and relation recognition in an end-to-end fashion. Such design targets for elegantly exploiting spatio-temporal context in videos to facilitate video SGG. As shown in Figure 6.2, the whole framework extends a Transformer-based object detector (e.g., DETR [22, 24]) along both spatial and temporal dimensions. Specifically, given the first frame of input video, an off-the-shelf Transformer-based object detector (consisting of a CNN backbone, Transformer encoder and instance decoder) is utilized to map a set of object queries into object predictions. Next, along the temporal dimension, the following video frames are fed into the shared object detector one by one, but the corresponding object queries are progressively constructed via the devised query propagation module (QPM). This way naturally triggers temporal associations of object instances by their identities. Meanwhile, with regard to the spatial dimension for each frame, we adopt a two-stage pipeline for relation recognition. That is, we first employ the relation proposal network to select the most interactive object pairs based on the detection results of each frame, and then enhance their representations via the relation decoder for relation classification. Most specifically, since the second frame, we devise a temporal

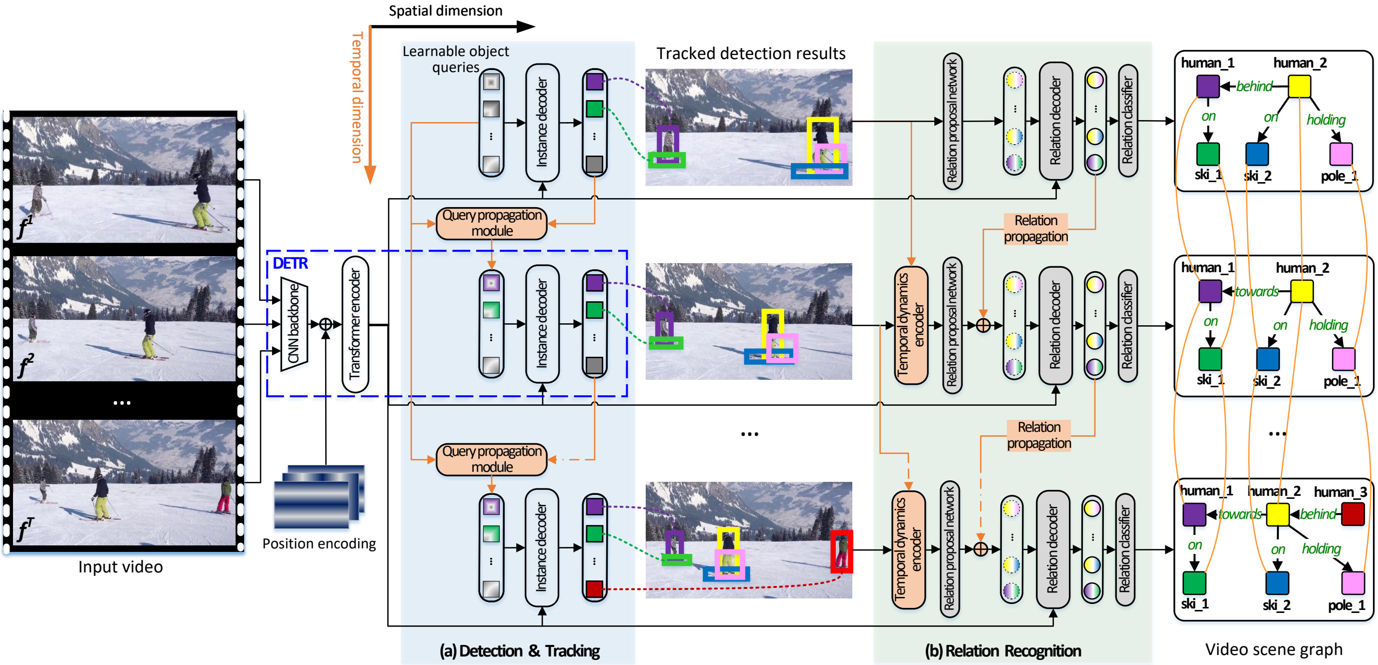


Figure 6.2: An overview of our Temporal Propagation Transformer (TPT), which extends upon a Transformer-based object detector (i.e., DETR [22]) along both spatial and temporal dimensions. Given a sequence of video frames, the Transformer encoder in DETR first extracts frame-wise features. (a) Next, a shared instance decoder in DETR maps object queries into detection results. Note that since the second frame, the input object queries of each frame are constructed by the query propagation module, which upgrades vanilla learnable object queries with previous detection outputs. (b) Based on the detection results of each frame, a relation proposal network is further utilized to select the most interactive object pairs, and the object pair representations are enhanced through the relation decoder for relation classification. Particularly, since the second frame, the temporal dynamics encoder and the relation propagation strategy are devised to enhance object and relation representations with video temporal context for boosting relation recognition. Finally, a sequence of frame-level scene graphs is generated as the output video scene graph.

dynamics encoder (TDE) and the relation propagation (RP) strategy to fully exploit video temporal context to boost relation recognition. Finally, a sequence of frame-level scene graphs with learned temporal associations is produced as the output video scene graph.

6.3.1 Preliminary: DETR

The vanilla Transformer-based object detector, i.e., DETR [22, 24], frames the task of object detection in a still image as a direct set prediction problem. Given the image features extracted by a CNN and Transformer encoder, an instance decoder transforms a fixed set of learnable object queries $\mathcal{Q} = \{\boldsymbol{q}_1, ..., \boldsymbol{q}_M\}$ into object predictions $\hat{\mathcal{Y}} = \{\hat{y}_1 = (\hat{l}_1, \hat{\boldsymbol{b}}_1), ..., \hat{y}_M = (\hat{l}_m, \hat{\boldsymbol{b}}_M)\}$ in parallel, where $\hat{l}_i$ and $\hat{\boldsymbol{b}}_i$ denote the predicted object label and bounding box coordinates of each object respectively. During training, the optimal assignment $\pi(\cdot)$ between M object predictions and N ground truth objects

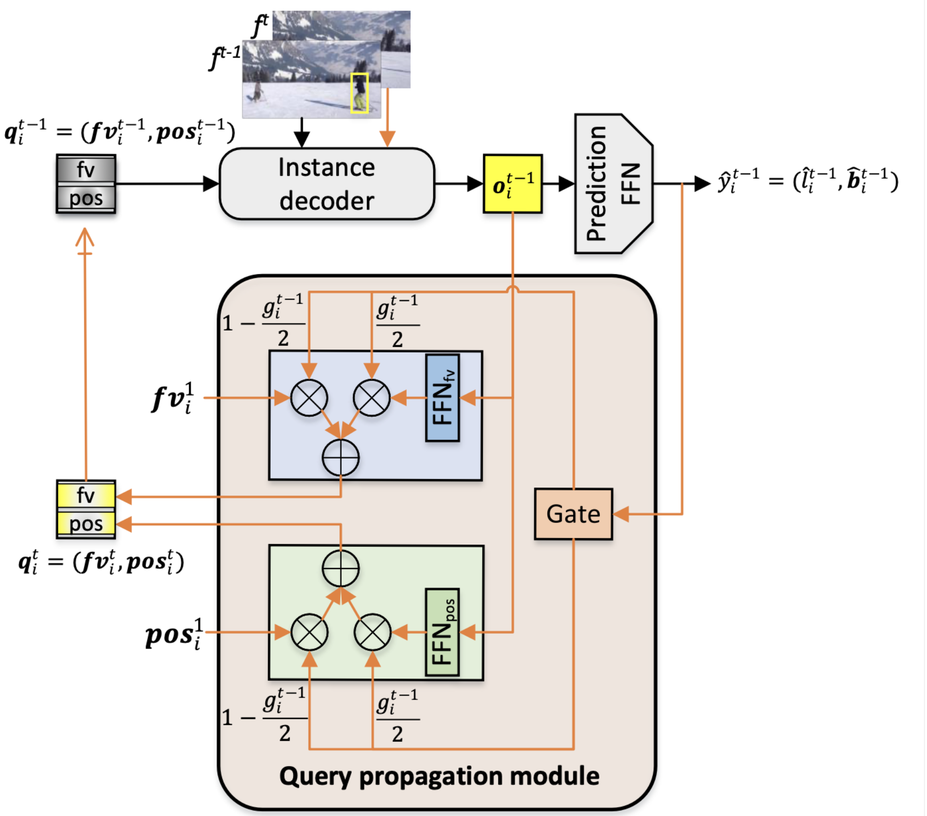


Figure 6.3: An illustration of Query Propagation Module (QPM). QPM constructs an object query $\boldsymbol{q}_i^t$ by jointly exploiting the initial prototypical query $\boldsymbol{q}_i^1$ and the detection output representation $\boldsymbol{o}_i^{t-1}$ in the previous frame.

$\mathcal{Y} = \{y_1 = (l_1, \boldsymbol{b}_1), ..., y_N = (l_N, \boldsymbol{b}_N)\}$ is achieved using Hungarian bipartite matching algorithm, and $\pi(i) \in \{1, ..., M\}$ represents the index of assigned prediction for the ground truth object y_i. Then, depending on all the matched pairs, Hungarian loss is computed by combining the negative log-likelihood for class prediction and a box loss for coordinates prediction.

6.3.2 Object Detection with Instance Temporal Association

In an effort to associate object instances across consecutive video frames that share the same identity, we extend a Transformer-based detector DETR with a new query propagation module along the temporal dimension. Such design upgrades DETR as an autoregressive model that can jointly detect and track multiple visual targets in a video.

Query Propagation Module (QPM). Each object query of the Transformer-based object detector can be regarded as a 'prototype' that learns to specialize in detecting objects with certain traits [22], e.g., large objects on the left of the image. However, simply applying this Transformer-based object detector over individual frames will neglect the rich temporal context information among adjacent frames, thereby might resulting in temporal discontinuity of detection results across frames. To alleviate this issue, we design Query Propagation Module (QPM) to augment the initial prototypical query with the detection results of previous frames, aiming to pursue the temporal coherence of detection results within videos. Figure 6.3 illustrates the detailed architecture of QPM.

Formally, let $video = [f^1, f^2, ..., f^T]$ denote the input video, where the superscripts represent the frame index. For the first frame f^1, QPM detects objects from scratch, and we denote the corresponding initial/prototypical object queries as $\mathcal{Q}^1 = \{\boldsymbol{q}_1^1, ..., \boldsymbol{q}_M^1\}$. Please note that a query consists of a feature vector $\boldsymbol{fv}_i^t$ and position encoding $\boldsymbol{pos}_i^t$: $\boldsymbol{q}_i^1 = (\boldsymbol{fv}_i^1, \boldsymbol{pos}_i^1)$. Next, since the second frame ($t \geq 2$), depending on the detection outputs of the Transformer decoder on the previous frame f^{t-1} (i.e., $\mathcal{O}^{t-1} = \{\boldsymbol{o}_1^{t-1}, ..., \boldsymbol{o}_M^{t-1}\}$), we construct each object query $\boldsymbol{q}_i^t = (\boldsymbol{fv}_i^t, \boldsymbol{pos}_i^t)$ for current frame f^t as:

$$\begin{aligned} g_i^{t-1} &= \mathbb{1}\{\boldsymbol{o}_i^{t-1} \text{ predicts an object}\} \\ &= \begin{cases} \mathbb{1}\{\hat{y}_i^{t-1} \text{ matches ground truth}\}, & \text{at training} \\ \mathbb{1}\{score(\hat{y}_i^{t-1}) > \sigma_{det}\}, & \text{at inference} \end{cases}, \end{aligned} \tag{6.1}$$

$$\boldsymbol{fv}_i^t = (1 - \frac{g_i^{t-1}}{2}) \cdot \boldsymbol{fv}_i^1 + \frac{g_i^{t-1}}{2} \cdot FFN_{fv}(\boldsymbol{o}_i^{t-1}), \tag{6.2}$$

$$\boldsymbol{pos}_i^t = (1 - \frac{g_i^{t-1}}{2}) \cdot \boldsymbol{pos}_i^1 + \frac{g_i^{t-1}}{2} \cdot FFN_{pos}(\boldsymbol{o}_i^{t-1}). \tag{6.3}$$

In Eq. (6.1), $\mathbb{1}\{\cdot\}$ denotes the indicator function, and we set it as 1 if $\boldsymbol{o}_i^{t-1}$ predicts an object (the corresponding prediction $\hat{y}_i^{t-1}$ is matched with a ground-truth object at training; or the detection confidence score $score(\hat{y}_i^{t-1})$ is greater than a threshold σ_{det} at inference), otherwise 0. Eq. (6.2) and Eq. (6.3) achieve the constructed query $\boldsymbol{q}_i^t$ by fusing the prototypical query $\boldsymbol{q}_i^1$ and its detection result in the previous frame via gate structure. Here we implement FFN_{fv} and FFN_{pos} as 2-layer feed-forward networks (FFN) with ReLU activations.

Adapted bipartite matching. Following DETR [22], the optimal bipartite assignment $j = \pi^1(i)$ (i.e., from ground truth y_i^1 to prediction $\hat{y}_j^1$ in frame f^1) is determined by the Hungarian algorithm, and we adopt the same pairwise costs between ground truths and predictions. Since the second frame, in order to force queries tracking particular objects, the bipartite assignment $\pi^t(i)$ for f^t is obtained by:

$$\pi^t(i) = \begin{cases} \pi^{t-1}(ID_{t\to t-1}(i)), & \text{if object } y_i^t \text{ exists in } f^{t-1}, \\ \text{using Hungarian algorithm,} & \text{otherwise,} \end{cases} \tag{6.4}$$

where $\pi^{t-1}(\cdot)$ is the assignment for f^{t-1}, and $ID_{t\to t-1}(i)$ is set as the index of the object in f^{t-1} sharing the same identity as y_i^t in f^t. For the case of 'otherwise' (i.e., newly appeared objects), we perform the Hungarian algorithm over unassigned predictions to search for the minimum cost assignment. An illustration of the adapted bipartite matching is shown in Figure 6.4.

6.3.3 Relation Recognition

After detecting the objects in each frame, the task of video SGG needs to further estimate the visual relations/interactions between detected objects. For each frame,

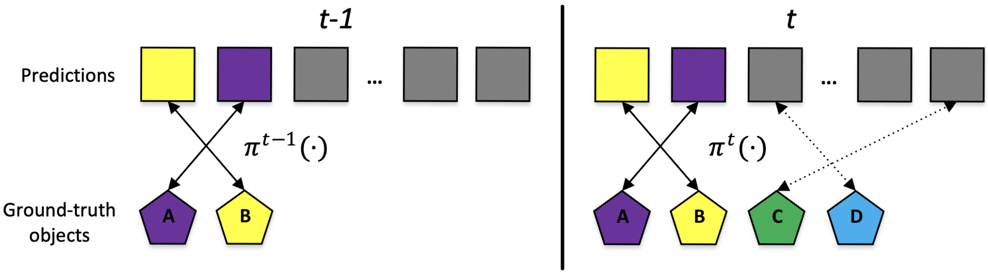


Figure 6.4: An illustration of the adapted bipartite matching. To obtain the optimal bipartite assignment $\pi^t(\cdot)$ for frame f^t, for objects that exist in the previous frame (A & B), $\pi^t(\cdot)$ inherits the assignments in $\pi^{t-1}(\cdot)$; for newly appeared objects (C & D), the Hungarian algorithm is performed over unassigned predictions (grey squares) to search for the minimum cost assignment.

conditioned on the detected object instances, we adopt a two-stage design for relation recognition, i.e., first producing relation proposals, and then performing relation classification. Moreover, we integrate the two-stage design with two new modules, i.e., Temporal Dynamics Encoder (TDE) and Relation Propagation strategy (RP), that additionally exploit the inherent dynamics and consistency across frames to strengthen relation recognition.

Temporal Dynamics Encoder (TDE). For each frame f^t, we denote its final detection outputs derived from instance decoder as $\mathcal{O}^t = \{\boldsymbol{o}_1^t, ..., \boldsymbol{o}_M^t\}$ ($\boldsymbol{o}_i^t \in \mathbb{R}^{256}$). As shown in Figure 6.5, TDE enriches each output instance representation by attending over its historic representations along the established trajectory. In this way, we measure the strengthened instance representation $\tilde{\boldsymbol{o}}_i^t$ via attention mechanism:

$$\tilde{\boldsymbol{o}}_i^t = \sum_{\tau=max(1,t-T_{his})}^{t} \alpha_i^{t,\tau} \cdot \boldsymbol{o}_i^{\tau}, \tag{6.5}$$

where $T_{his} \in \mathbb{N}$ is the max timespan for retrieving historic representations. The attention weight $\alpha_i^{t,\tau}$ is normalized with Softmax: $\alpha_i^{t,\tau} = \frac{e_i^{t,\tau} \cdot g_i^{\tau}}{\sum_\tau e_i^{t,\tau} \cdot g_i^{\tau}}$, where g_i^τ is obtained by Eq. (6.1) to mask empty detections in trajectory, and the primary attention weight $e_i^{t,\tau}$ is measured as:

$$e_i^{t,\tau} = \frac{(\boldsymbol{W}_1 \cdot \boldsymbol{o}_i^t)^T (\boldsymbol{W}_2 \cdot \boldsymbol{o}_i^\tau + PE(t-\tau))}{\sqrt{d_{hidden}}}, \tag{6.6}$$

where PE is the relative position encodings [241] of $\mathbb{R}^{(T_{his}+1) \to 256}$ to encode time interval information, $d_{hidden} = 256$ is the hidden dimension, and $\boldsymbol{W}_1, \boldsymbol{W}_2$ are parametric matrices. In practice, TDE is simply implemented as a standard 2-layer Transformer encoder [40].

Relation Proposal Network. Next, we construct all possible subject-object pairs based on the detection results. Suppose N_{det} object instances are detected in a video frame, there would be $N_{det} * (N_{det} - 1)$ subject-object pairs in total. Each pair is represented as the concatenation of the representations of two involved instances: $\boldsymbol{pair}_{i \to j}^t = cat[\tilde{\boldsymbol{o}}_i^t, \tilde{\boldsymbol{o}}_j^t] \in \mathbb{R}^{512}$. Taking the subject-object pair $\boldsymbol{pair}_{i \to j}^t$ as inputs, the

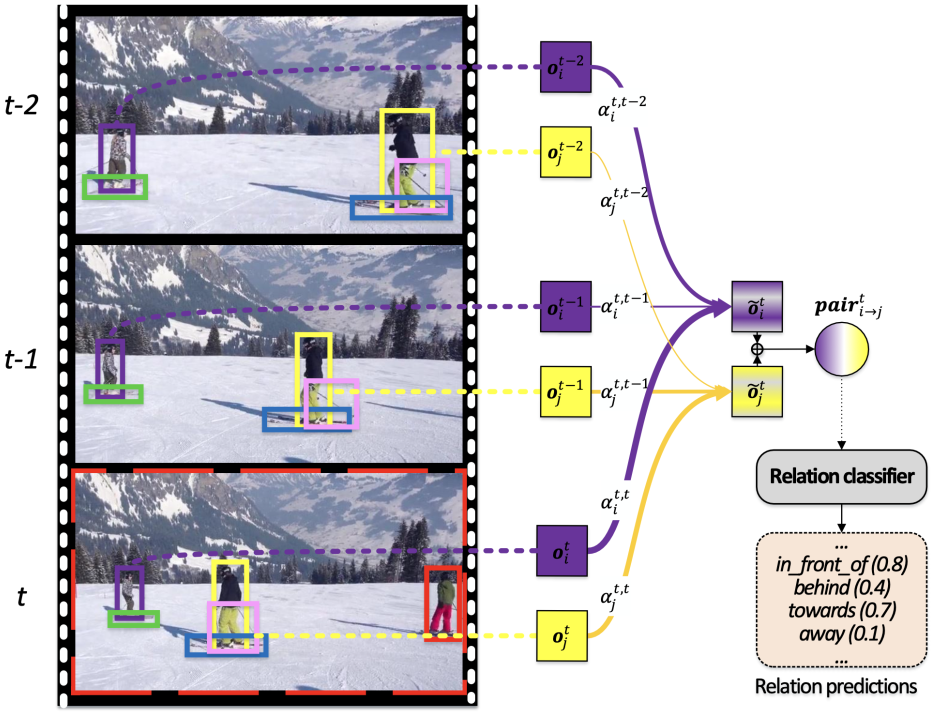


Figure 6.5: An illustration of Temporal Dynamics Encoder (TDE). Before predicting the relations of 'person_i' (purple box) to 'person_j' (yellow box) in frame t, TDE enriches each instance representation (i.e., $\boldsymbol{o}_i^t$ and $\boldsymbol{o}_j^t$) by attending over their historic states. Next, we perform relation recognition based on the enhanced instance representations ($\tilde{\boldsymbol{o}}_i^t$ and $\tilde{\boldsymbol{o}}_j^t$).

relation proposal network predicts an interactiveness score $\hat{z}_{i\to j}^t \in [0,1]$, which indicates the possibility that relations exist in this pair. We implement this relation proposal network as a 2-layer FFN with Sigmoid output activation. During training, for each frame, we sample at most K pairs, including both positive and negative training samples. Note that if both IoUs (intersection over union) of predicted bounding boxes with regard to ground truths are larger than 0.5, we consider such pair as positive, otherwise it is negative. We optimize this module with focal loss [120] (FL):

$$L_{rel\_prop} = \frac{1}{N_{pair}} \sum_{n=1}^{N_{pair}} FL(\hat{z}_n, z_n), \tag{6.7}$$

where N_{pair} is the number of sampled pairs in a training batch, and $z_n \in \{0,1\}$ is the ground truth label. At inference, only the top-K subject-object pairs of each frame with the highest interactiveness scores are retained as candidate proposals for further relation classification.

Relation Propagation Strategy (RP). Considering that video events (e.g., the visual relations between two objects) are temporally consistent across consecutive frames, we further design the relation propagation strategy to explicitly mine the temporal context of relations from previous frames. Concretely, since the second frame, we supplement the outputs of the relation proposal network with the predicted top-K_{RP} interactive object pairs from the previous frame ($K_{RP} = K/2$ in our implementation).

Formally, the complete set of relation proposals in the current frame ($\mathcal{P}^t$) is the union of two sets:

$$\mathcal{P}^t = \bar{\mathcal{P}}^t \cup \mathcal{P}(triplets^{t-1}[: K_{RP}]), \tag{6.8}$$

where $\bar{\mathcal{P}}^t$ is the output set of the relation proposal network (i.e., top-K subject-object pairs with the highest interactiveness scores in the current frame), and $\mathcal{P}(triplets^{t-1}[: K_{RP}])$ represents the set of object pairs among top-K_{RP} output relation triplets (ranked by triplet scores, see Section 6.3.4) in the previous frame. This way directly constructs a complete set of relation proposals by fully preserving prior information, which is denoted as $\mathcal{P}^t = \{\boldsymbol{pair}_1^t, ..., \boldsymbol{pair}_{K'}^t\}, K' \leq K + K_{RP}$.

Relation Decoder (RD). Before relation classification, we employ a relation decoder as in the standard Transformer-based decoder [40], targeting for enhancing K' object pair representations $\mathcal{P}^t = \{\boldsymbol{pair}_1^t, ..., \boldsymbol{pair}_{K'}^t\}$ via attention mechanism. Specifically, the pair representations are firstly transformed by a fully-connected (FC) layer ($\mathbb{R}^{512 \to 256}$). They are further fed into the self-attention layer to perform relational reasoning among object pairs, leading to context-aware intermediate pair representations $\hat{\mathcal{P}}^t = \{\hat{\boldsymbol{pair}}_1^t, ..., \hat{\boldsymbol{pair}}_{K'}^t\}$. Next, a cross-attention layer enriches each pair representation $\hat{\boldsymbol{pair}}_k^t$ by attending over primary image features, which are shared with the object detector. Finally, the relation decoder produces the enhanced pair representations $\tilde{\mathcal{P}}^t = \{\tilde{\boldsymbol{pair}}_1^t, ..., \tilde{\boldsymbol{pair}}_{K'}^t\}$.

Relation Classifier. Given the output pair representations from relation decoder $\tilde{\boldsymbol{pair}}^t$, the sub-task of relation classification is formulated as a multi-label classification problem. We implement the relation classifier as a 2-layer FFN with final Sigmoid activation. The objective of the relation classifier is measured by focal loss:

$$L_{rel\_cls} = \frac{1}{N_{pair} \cdot C_{rel}} \sum_{n=1}^{N_{pair}} \sum_{c=1}^{C_{rel}} FL(\hat{r}_{ic}, r_{ic}), \tag{6.9}$$

where C_{rel} is the number of relation/predicate categories, $r_{ic} \in \{0, 1\}$ indicates whether the i-th pair has the c-th ground truth relation, and $\hat{r}_{ic}$ is the predicted probability.

6.3.4 Training and Inference

Training. At the training stage, we adopt a 4-step training scheme to optimize our TPT progressively, which shares a similar spirit of curriculum learning [259].

(i) In the first step, we train a Transformer-based detector for object detection in individual video frames. It is initialized from a COCO-pre-trained model and fine-tuned using frame-level box annotations. The objective is measured as $L_{step1} = \sum_i^{N_{bs}} L_{obj\_det}(f_i)$, where N_{bs} denotes batch size, and $L_{obj\_det}(\cdot)$ denotes the Hungarian loss [22] for a single image f_i of the Transformer-based detector.

(ii) In the second step, we fine-tune the Transformer-based detector together with the QPM module, to further build temporal associations of detected instances. The learning rate of the Transformer-based detector is set as 1/10 of the one of QPM for stable learning. This step is trained with sampled video clips of length T_{clip}, and the training loss is measured as $L_{step2} = \frac{1}{T_{clip}} \sum_{t=1}^{T_{clip}} L_{obj\_det}(f^t)$ under the bipartite matching described in Section 6.3.2.

(iii) In the third step, we freeze all parameters of the architecture learned in the previous step, and only optimize the modules for relation recognition. This step is also trained with video clips, and the objective is the integration of the relation proposal loss in Eq. (6.7) and relation classification loss in Eq. (6.9): $L_{step3} = L_{rel\_prop} + L_{rel\_cls}$.

(iv) In the final step, we jointly fine-tune the whole framework with the overall loss ($L_{step4} = L_{step2} + L_{step3}$), which learns to generate video scene graphs from raw videos in an end-to-end fashion.

Inference. At inference, each video frame is fed into the proposed video SGG model one by one, and we produce the frame-level scene graphs in an online fashion. Specifically, the object predictions are retained if their confidence scores are greater than a threshold (0.5), and two objects in consecutive frames are temporally associated if they are predicted by the same object query slot. For relation recognition on each frame, we compute the ranking score for each ⟨*subject-predicate-object*⟩ relation triplet as $s_{triplet} = s_{subject} \cdot s_{predicate} \cdot s_{object}$, where $s_{subject}$, $s_{predicate}$ and s_{object} represent the confidence scores for subject, predicate and object respectively.

6.4 Experiments

We evaluate our proposed TPT on two video SGG benchmarks: VidHOI [100] and Action Genome [54]. Both datasets focus on recognizing human-object/human-human relations, which are pivotal for human-centric scene understanding.

Implementation Details. We adopt Deformable-DETR [24] as the Transformer-based object detector, and leverage the same standard data augmentation pipeline. The CNN backbone of Deformable-DETR is ResNet-50 [39], and it uses $M = 300$ object queries. The whole architecture is trained over 8 Nvidia 2080ti GPUs with the AdamW optimizer. Initial learning rates are set as 10^{-4} for parameters trained from scratch, and 10^{-5} for other parameters initialized from pre-trained models. The maximum training epoch number is 10, with learning rates dropping by 10 times after 6 epochs. In experiments, we set video clip length $T_{clip} = 3$ and $T_{his} = 2$ for the temporal dynamics encoder due to GPU memory limit. Moreover, we select top-16 (i.e., K=16) object pairs from the relation proposal network for final relation recognition.

6.4.1 Experiments on VidHOI

We comprehensively evaluate our approach on the VidHOI benchmark [100] for the video SGG task. This dataset is constructed upon VidOR [260] by sampling keyframes at 1FPS frequency. Slightly different from [100], we keep intermediate video keyframes even though no relations are annotated, in order to avoid inconsistent temporal jittery. The resulting VidHOI consists of 6,366 videos with 208,686 keyframes for training, and 756 videos with 24,667 keyframes for testing. The whole dataset has ~1.5M annotated relation triplets, covering 78 object categories and 50 predicate categories. Specifically,

Method	RD	QPM	TDE	RP	*Detection* (mAP)			*Oracle* (mAP)		
					Full	Temporal	Spatial	Full	Temporal	Spatial
Base	-	-	-	-	7.0	2.9	9.4	27.9	9.9	38.9
Base*	✓	-	-	-	7.2	3.1	9.7	30.4	12.2	41.2
+TDE	✓	-	✓	-	7.2	3.2	9.7	30.7	12.7	41.2
+QPM	✓	✓	-	-	7.3	3.2	9.8	30.8	12.8	41.4
+QPM+TDE	✓	✓	✓	-	7.5	3.4	9.9	31.1	13.3	41.6
+QPM+RP	✓	✓	-	✓	7.5	3.2	10.0	31.0	13.2	41.5
TPT (*w/o jointly fine-tune*)	✓	✓	✓	✓	7.7	3.4	10.2	**31.4**	**13.6**	**41.9**
TPT	✓	✓	✓	✓	**7.9**	**3.5**	**10.4**	-	-	-

Table 6.1: Ablation studies on VidHOI. Note that the *Detection* mode uses predicted object trajectories, while the *Oracle* mode uses ground truth trajectories. TPT performance under *Oracle* is omitted since there is no need to fine-tune with detection and tracking when the ground truth trajectories are provided. (RD: relation decoder, QPM: query propagation module, TDE: temporal dynamics encoder, RP: relation propagation)

the predicate categories include 25 temporal predicates (e.g., *pull*, *lift*, occupying ∼5% of the dataset) and 25 spatial predicates (e.g., *next to*, *behind*).

Evaluation Metrics. Following [100], we compute mean Average Precision (mAP) as evaluation metrics over five relation triplet sets, i.e., ***Full*** (all 557 triplet categories), ***Temporal*** (207 triplet categories with temporal predicates), ***Spatial*** (350 triplet categories with spatial predicates), ***Non-rare*** (242 triplet categories with no less than 25 instances) and ***Rare*** (315 triplet categories with less than 25 instances). Similarly, we report results under two evaluation modes: the ***Oracle*** mode which assumes ground truth trajectories known at inference, and the ***Detection*** mode which uses predicted trajectories instead. Note that, we implement the Oracle mode by obtaining the optimally matched predictions from the instance decoder as ground truth object representations.

Ablation Studies. We conduct ablation studies by comparing different variants of our proposed model. The results are shown in Table 6.1. The ***Base*** model simply stacks a relation proposal network plus the relation classifier over a pre-trained DETR. We also provide a stronger baseline (***Base****), which exploits the relation decoder to enhance the relation feature by attending on low-level image features. The performance improvements (e.g., the Full/Temporal/Spatial mAP is increased from 7.0/2.9/9.4 to 7.2/3.1/9.7 in Detection mode) show that such feature enhancement can benefit both spatial and temporal relation recognition. Particularly, we add the TDE module over Base*, i.e., ***+TDE***, which requires additional trackers to construct temporal associations at inference. The obtained performance boost over Base* is trivial, perhaps because the used non-differentiable tracker is not jointly trained with the whole framework.

Both of the two baselines (*Base* and *Base**) learn video scene graphs from static frames without considering any temporal information, and they also require a potential tracking algorithm to build instance temporal associations. Instead, we propose to build temporal associations by extending the pre-trained DETR with a query propagation module. In this way, ***+QPM*** slightly boosts up the performances of *Base** by

Method	Backbone	Size	Time	*Detection* (mAP)					*Oracle* (mAP)				
				Full	Temporal	Spatial	Non-Rare	Rare	Full	Temporal	Spatial	Non-Rare	Rare
2D model [261]	ResNet-50	76.0M	-	2.6	1.5	2.7	4.7	1.7	14.1	8.3	18.6	22.9	11.3
3D model [100]	ResNet-50	92.5M	-	2.6	1.6	2.9	4.9	1.9	14.4	7.7	20.9	23.0	12.6
ST-HOI [100]	ResNet-50	155.5M	-	3.1	1.9	3.3	5.9	2.1	17.6	14.4	25.0	27.2	17.3
SERVO-HOI [247]	ResNeXt-101	-	-	4.8	-	-	6.8	**4.1**	21.1	-	-	29.2	19.5
TUTOR [246]	ResNet-50	-	-	-	-	-	-	-	26.9	**21.3**	32.2	37.1	**23.5**
STTran [262]	ResNet-50	51.1M	91.8ms	7.2	2.8	9.8	12.6	2.9	30.9	13.2	41.4	43.8	21.1
Base*	ResNet-50	47.2M	162.5ms	7.2	3.1	9.7	13.3	2.3	30.4	12.2	41.2	43.7	20.4
TPT	ResNet-50	50.2M	213.2ms	**7.9**	**3.5**	**10.4**	**13.9**	2.6	**31.4**	13.6	**41.9**	**44.3**	21.5

Table 6.2: Comparison with state-of-the-art video SGG methods on VidHOI. Please note that the *Detection* mode uses predicted object trajectories, while the *Oracle* mode uses ground truth trajectories. We report the model sizes (i.e., number of parameters) and inference time of our implemented models and some previous models with open-sourced codes.

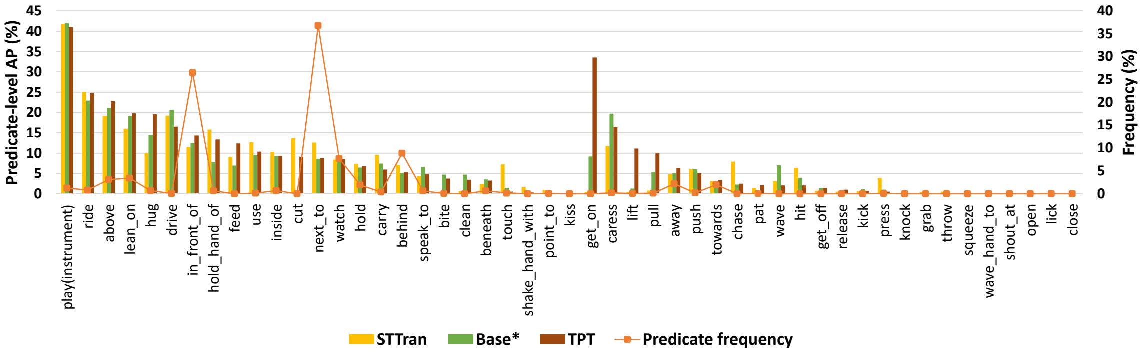


Figure 6.6: Predicate-level AP comparison among STTran, Base* and the proposed TPT. The predicates of VidHOI can be divided into two groups [100]: Spatial predicates (*play(instrument)* → *kiss*) and Temporal predicates (*get_ on* → *close*).

jointly unifying object detection, tracking and relation recognition. More importantly, such a unified framework provides a fertile ground to exploit cooperative interactions among these sub-tasks for boosting video SGG. Concretely, by additionally employing the relation propagation strategy (***+QPM+RP***) or encoding instance temporal dynamics (***+QPM+TDE***), the upgraded model leads to significant performance improvements respectively. These results clearly highlight the effectiveness of our RP and TDE, which manage to exploit video temporal consistency and dynamics for facilitating relation recognition. When RP and TDE are leveraged together (i.e., ***TPT*** without jointly fine-tuning), we attain the performance improvements as expected. Finally, after jointly fine-tuning the whole framework (i.e., the full ***TPT***, the highest performances are attained. These observations basically validate the merits of our proposed unified framework that enables flexible spatio-temporal modeling for pursuing high-quality video SGG.

Comparison with State-of-the-arts. Table 6.2 presents the performance comparison in terms of mAP metrics under both Oracle and Detection modes. Specifically, we compare TPT with several existing video SGG methods, including: i) the basic 2D model [261] that extracts visual features from a single target frame; ii) the 3D model [100] that leverages temporal-aware features from a 3D backbone; iii) ST-HOI [100] that uses improved temporal feature pooling and additional human pose knowledge; iv) the concurrent SERVO-HOI [247] model that particularly addresses the long-tail distribution issue; v) the concurrent TUTOR [246] model that structurizes a video into tubelet tokens for learning spatio-temporal visual semantics. Overall, we observe that our proposed TPT clearly outperforms these methods over most metrics. This is partially due to a weaker object detector and a feature backbone trained from scratch in the 2D model, 3D model and ST-HOI (implementation details of these models can be found here[2]). Moreover, all these methods adopt pre-trained object detectors and optional feature backbones for trajectory feature pooling, and their independent optimization may also lead to sub-optimal solutions. It is worth noting that ST-HOI and TUTOR surpass TPT in recognizing temporal predicates in the Oracle mode. This is because they have leveraged additional human pose information or the detailed object ground truth trajectories around the target frame.

To make the comparison fairer, here we additionally report the results of our Base* model using the same underlying detector (with AP_{50}=26.5 on VidHOI). We also implement STTran [262], a recent state-of-the-art spatio-temporal contextualization video SGG method that employs Transformer architecture over Faster R-CNN detection results (AP_{50}=26.6). In general, the results under the same backbone (ResNet-50) show that our proposed TPT exhibits better performances than existing methods across different evaluation modes consistently. Thanks to our unified design that flexibly incorporates temporal information, both performances for predicting spatial and temporal relation triplets are improved. This demonstrates the outstanding spatial-temporal modeling ability of our TPT for the video SGG task.

In addition, the proposed TPT performs very well in recognizing Non-rare categories, which surpasses all existing methods. In terms of Rare categories, TPT outperforms previous baseline methods (i.e., 2D model, 3D model, ST-HOI, STTran and Base*), while some concurrent methods (e.g., SERVO-HOI [247], TUTOR [246]) can achieve even better results. This is because these latest methods might have particular designs to tackle the long-tail distribution issue, e.g., class-weighted training objective in SERVO-HOI. Please note that TPT could also apply these techniques to improve the recognition of Rare categories. We show the detailed performance comparison on predicate-level APs (computed by averaging APs of all triplets of the same predicate) in Figure 6.6. The proposed TPT is better at recognizing spatial relations such as *lean_on, above*, and temporal predicates such as *get_on, lift, pull.* However, the imbalanced distribution of predicates has actually made the video SGG task even more challenging. All three methods show large performance variances in recognizing different relations, and they still fail to recognize infrequent temporal predicates such as *lick* and *close.*

Evaluation of Temporal Association. We also evaluate the temporal association quality in video SGG, which is often neglected in prior works. We think the high-quality temporal associations of video scene graphs are also essential, especially when applying the generated video scene graphs for downstream tasks that need spatio-temporal reasoning. For example, given two successive relation triplets 'human hold basketball' and 'human throw basketball', whether the two mentioned 'human' are the same player or not reflects different video events (i.e., assisting or shooting).

[2] https://github.com/coldmanck/VidHOI

Video SGG model	Tracker	IDF1 (↑)	IDs (↓)
STTran [262]	IOUTracker	53.0	6407
	SORT [258]	52.3	4239
	BYTE [257]	53.8	3061
Base*	IOUTracker	52.3	7184
	SORT [258]	54.3	4343
	BYTE [257]	54.3	2670
TPT	-	**59.9**	**2586**

Table 6.3: Evaluation of temporal association quality in video SGG.

Here we follow the traditional multi-object tracking task [263, 264] and leverage the IDF1 (i.e., identity F1 score) and IDs (i.e., number of identity switches) metrics to evaluate the temporal association quality of video scene graphs. Specifically, we compare our TPT with STTran [262] and the Base* model. We adopt three popular tracking algorithms, i.e., IOUTracker, SORT [258] and BYTE [257], to construct temporal associations for STTran [262] and Base* model in post-processing. IOUTracker greedily associates a detection result to the existing trajectory that has the same category label and the maximum IoU overlapping score. In contrast, SORT utilizes the Hungarian algorithm for optimal bipartite matching between detections and trajectories. It also employs the Kalman filter [265] to incorporate motion prediction. BYTE is a recent state-of-the-art multi-object tracking algorithm. Instead of neglecting low-confidence detections, BYTE associates all detection results to extend trajectories. This makes BYTE exhibit superiority in eliminating object missing and fragmented trajectories issues.

Table 6.3 summarizes the performance comparison on temporal association quality. In particular, the basic IOUTracker results in relatively large IDs. This is due to its greedy association strategy which is sensitive to object overlapping and fast-moving issues. SORT and BYTE trackers that employ the Hungarian algorithm for data association can effectively mitigate these limitations. Hence, they achieve lower IDs and better IDF1. In our TPT, temporal associations are inherently determined by query propagation. This enables each object query continually to search for a particular visual object, eliminating hand-designed components such as Hungarian matching, and Kalman filtering. The results show the proposed TPT has achieved the best IDF1 and IDs scores. Such superior performances validate the effectiveness of the end-to-end philosophy for constructing temporal associations in video SGG.

Analysis of Inference Time. To analyze inference time, we evaluate video SGG models on a single NVIDIA RTX 2080Ti GPU. The results are shown in Table 6.2. Specifically, since STTran processes multiple video frames in a batch, it achieves the highest inference speed (i.e., 91.8ms per frame). Though both Base* and TPT process videos in a frame-by-frame manner, TPT has temporal propagation modules to achieve unified object detection, tracking and relation recognition. Hence, TPT requires relatively more inference time. Please note that we sample keyframes in videos at 1FPS frequency. TPT's inference speed (i.e., 213.2ms / 4.7FPS) is sufficient for real-time prediction. The inference times of previous methods are unavailable since they are not open-sourced or use pre-computed detection boxes.

Method	RD	QPM	TDE	RP	*SGDET*			*PREDCLS*		
					R@10	R@20	R@50	R@10	R@20	R@50
Base	-	-	-	-	29.3	37.1	47.4	83.6	96.8	**99.9**
Base*	✓	-	-	-	30.5	38.2	48.9	85.2	97.1	**99.9**
+TDE	✓	-	✓	-	30.5	38.1	49.1	85.3	97.3	**99.9**
+QPM	✓	✓	-	-	30.7	37.8	49.7	85.3	97.3	**99.9**
+QPM+TDE	✓	✓	✓	-	31.0	38.3	50.7	85.5	**97.4**	**99.9**
+QPM+RP	✓	✓	-	✓	30.7	38.0	50.1	85.4	97.3	**99.9**
TPT (*w/o jointly fine-tune*)	✓	✓	✓	✓	31.1	38.6	51.2	**85.6**	**97.4**	**99.9**
TPT	✓	✓	✓	✓	**32.0**	**39.6**	**51.5**	-	-	-

Table 6.4: Ablation studies on Action Genome. Please note that the *SGDET* mode uses predicted objects, while the *PREDCLS* mode uses ground truth object labels and locations. TPT performance under *PREDCLS* is omitted since there is no need to fine-tune with detection and tracking when the ground truth object labels and locations are provided. (RD: relation decoder, QPM: query propagation module, TDE: temporal dynamics encoder, RP: relation propagation)

Method	Backbone	Size	AP_{50}	*SGDET*			*PREDCLS*		
				R@10	R@20	R@50	R@10	R@20	R@50
GPNN [219]	ResNet-101	-	20.7	-	32.2	42.1	-	62.3	68.1
HORT [103]	ResNet-101	-	20.7	-	37.2	47.8	-	71.7	76.2
VRD [41]	ResNet-101	-	24.6	19.1	28.8	40.5	59.6	78.5	99.2
Motif Freq [62]	ResNet-101	-	24.6	22.8	34.3	46.4	73.4	92.4	99.6
MSDN [137]	ResNet-101	-	24.6	23.1	34.7	46.5	74.9	92.7	99.0
VCTREE [134]	ResNet-101	-	24.6	23.9	35.3	46.8	75.5	92.9	99.3
RelDN [61]	ResNet-101	-	24.6	24.1	35.4	46.8	75.7	93.0	99.0
GPS-Net [68]	ResNet-101	-	24.6	24.4	35.7	47.3	76.0	93.6	99.5
STTran [262]	ResNet-101	70.0M	24.6	24.6	36.2	48.8	77.9	94.2	99.1
Li *et al.* [245]	ResNet-101	-	24.6	25.7	37.9	50.1	78.5	95.1	99.2
TPT	ResNet-50	50.2M	26.6	29.6	37.3	49.2	85.5	97.3	**99.9**
TPT	ResNet-101	69.1M	**27.6**	**32.0**	**39.6**	**51.5**	**85.6**	**97.4**	**99.9**

Table 6.5: Comparison with state-of-the-art video SGG methods on Action Genome. Please note that the *SGDET* mode uses predicted objects, while the *PREDCLS* mode uses ground truth object labels and locations. Results of previous methods are mainly from [103, 262]. We report model sizes for those with open-sourced implementation. (Most of the previous models are originally designed for image SGG, while their modified versions by [103, 262] for video SGG on Action Genome are not open-sourced.)

6.4.2 Experiments on Action Genome

In addition, we evaluate our TPT on the Action Genome dataset [54]. This dataset provides frame-level scene graph labels for Charades videos [266], including 234,253 sampled keyframes derived from ∼10K videos for training and testing. The number of total annotated relation triplets is ∼1.7M, covering 35 object categories and 25 predicate categories. However, as pointed out in [100], this dataset suffers from limitations such as incomplete/incorrect labels, and the annotated visual relations only focus on a single “actor" even though other people may exist. Please also note that the keyframes in this dataset are non-uniformly sampled, and the ground truth object trajectories are unavailable. In practice, considering that at most one instance is annotated in almost all video frames for each object category, we simply associate instances of the same category as the ground truth trajectories.

Evaluation Metrics. Following previous works [54, 103, 104, 262], we compute Recalls among top predictions (i.e., R@10/20/50) for evaluation, since relations are sparsely annotated in Action Genome. Moreover, we report recall metrics without constraining each subject-object pair to predict only one predicate, since we consider relation recognition as a multi-label classification problem. Particularly, we adopt two evaluation protocols: the predicate classification (PREDCLS) mode and the scene graph detection (SGDET) mode. PREDCLS assumes that the object information (i.e., bounding boxes and labels) is given, and it only evaluates the relation recognition between subject-object pairs. Instead, SGDET generates scene graphs from raw video frames using detected objects.

Ablation Studies. We conduct ablation studies on Action Genome to validate the effectiveness of each design in our TPT by comparing different variants (i.e., ***Base***, ***Base****, ***+TDE***, ***+QPM***, ***+QPM+TDE***, ***+QPM+RP***), as defined in Section 6.4.1. The results are summarized in Table 6.4. Similar to the observations on VidHOI, the performance boosts of each component on Action Genome are attained. This again validates the efficacy of our devised modules or strategies for achieving high-quality video SGG.

Comparing to State-of-the-arts. Next, we compare our results with state-of-the-art video SGG approaches on Action Genome. The results are summarized in Table 6.5. By using a weaker backbone ResNet-50, the proposed TPT has already achieved impressive performances that surpass most existing methods across both SGDET and PREDCLS modes. When upgrading with the stronger ResNet-101 backbone as in existing methods, our TPT achieves better detection performances (AP_{50} score from 26.6 to 27.6). Clear performance boosts on relation recognition are obtained in SGDET mode. While in terms of PREDCLS which uses given object information, the performance gains also increase a bit because of better instance representations. These results generally demonstrate the superiority of the proposed TPT over existing approaches.

6.4.3 Qualitative Analysis

In this section, we analyze video SGG models by visualizing the prediction results. Figure 6.7 showcases video scene graphs generated via STTran, our Base^* (without using tracking algorithms for temporal association) and the proposed TPT model. We take one VidHOI test video (Figure 6.7 (a)) and one Action Genome test video (Figure 6.7 (b)) for demonstration. Please note that only the top-10 predicted relation triplets and their involved objects are shown in each frame-level scene graph.

Firstly, we see that visual targets are well tracked by object query indices in our TPT. Note that the numbers after bounding box category names in Figure 6.7 are the indices of queries that predict the objects. While in Base^*, the same visual targets across frames are usually predicted by different object queries especially when video content changes drastically. As such, detected instances of Base^* require potential tracking algorithms for associating like STTran.

A very important finding here is that, compared with STTran and Base*, the topologies of the temporally-evolving frame-level scene graphs via TPT are much more consistent. That is, TPT seems to focus on several salient objects (i.e., 'person_269', 'person_251', 'sofa_187', 'table_166' in Figure 6.7 (a), and 'person_271', 'bed_139', 'laptop_50' in Figure 6.7 (b)), and continuously monitors how their interactions change over time. In contrast, the set of involved objects in top predicted relations of STTran and Base* exhibit much more variations over time. We attribute such topology consistency property to the devised query and relation propagation mechanisms. These mechanisms inherently enforce dependencies of objects and their pairwise interactions across consecutive frames. We think this property aligns with human's visual perception, and will be advantageous for supporting downstream video spatio-temporal reasoning.

Moreover, the proposed TPT demonstrates excellent performance in recognizing object relations. Note that green relation predicates on edges in Figure 6.7 indicate correct predictions. For example, when examining relations between 'person_251' and 'table_166' in Figure 6.7 (a), except for correctly predicting a spatial predicate 'next_to', the model also manages to recognize 'person_251 towards table_166' before frame *t+2* and 'person_251 away table_166' thereafter. This indicates that the proposed TPT can effectively exploit temporal context for predicting challenging temporal relations.

Finally, we summarize patterns for failure cases: 1) Wrongly or low-quality detection results, e.g., 'guitar_39' in frame t of Figure 6.7 (a), and 'person_271' with incomplete bounding boxes in frame t and *t+1* of Figure 6.7 (b). We believe improving underlying object detection can lead to higher-quality video SGG. 2) Ambiguous predicates prediction. For example, we tend to use more accurate 'sit-on' instead of 'in_front_of' to describe the relations between 'person_269' and 'sofa_187' in the video of Figure 6.7 (a). For this issue, we believe incorporating human pose knowledge can help, since the orientation and pose of the person provide crucial cues for describing human-object interactions.

6.5 Conclusion

In this chapter, we have presented TPT, a novel end-to-end Transformer-based architecture to generate temporally-evolving graph structure for videos. Particularly, we first upgrade a Transformer-based detector with devised query propagation mechanism to jointly perform object detection and tracking. Over such a flexible framework, we devise the temporal dynamics encoder and the relation propagation strategy to fully exploit video spatio-temporal context for boosting relation recognition. We validate the proposed approach on two challenging benchmarks, and have achieved superior performance improvements against existing video SGG methods that adopt separate stage-by-stage designs.

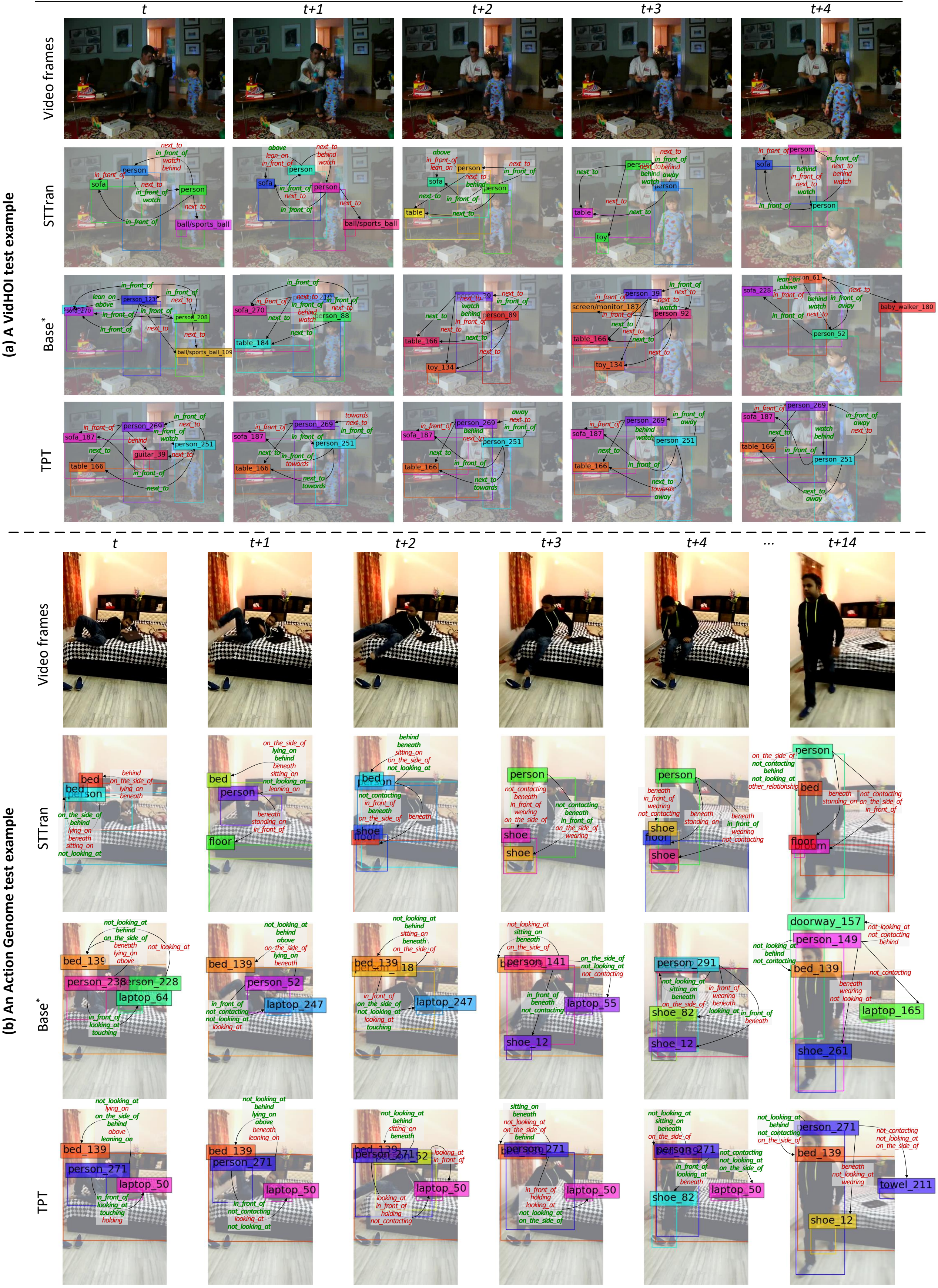


Figure 6.7: Video scene graphs generated via STTran, our Base* and TPT model, from the input video clip in the topmost row of (a) & (b). Only top-10 relation triplets and involved objects are shown. Predicates (on connection edges by order) matched with ground truths are highlighted in bold green. Particularly, for Base* and TPT, query indices that predict the objects are presented after category names, e.g., 'person_251'. (Better viewed in color.)

Chapter 7

Conclusion and Outlook

7.1 Dissertation Summary

Making computers perceive, understand and utilize immense unstructured visual data is highly needed in building next-generation human-like AI. Human's visual cognition system is highly structured – we identify objects from raw perceptual input and understand visual scenes in terms of their composition as objects and relations. In this thesis, we seek to enable computers to construct visual intelligence like humans. To this end, we envision an intermediate system to first transform visual data into structured representations that are ready for use in intelligent applications. In such a way, visual data can be perceived, understood and fused into the intelligence-building process.

Concretely, we adopt the scene graph for structured visual representation, which describes complex visual semantics by encoding objects as nodes and their relations as edges into a symbolic graph structure. Scene graphs have various advantages including alignment with human visual perception, bridging multi-modal / multi-source information, and could be easily inspected and interpreted. However, the current state of research is not capable of generating scene representations for visual data with practical accuracy and efficiency. In this dissertation, we have presented novel algorithms and approaches for pursuing high-quality structured visual parsing, mainly on scene graph generation (SGG). We have explored highlighting salient elements in scene graphs (Chapter 3), pursuing language-supervised and open-vocabulary SGG (Chapter 4), improving human-centered visual structure parsing using Transformer (Chapter 5), and tackling the challenging video SGG task through a monolithic spatio-temporal Transformer (Chapter 6). These technologies have been validated to be effective through extensive experiments and applied in downstream applications (Chapter 3). Overall, our work in this thesis has strongly promoted the research for semantic representation, comprehension and applications of visual data in a structural way.

7.2 Open Problems in Future Work

The generation and application of structured visual representation in terms of scene graphs is still an area of active research. Many directions demand further research in future work. To utilize scene graph representations in life-critical fields like autonomous driving and healthcare, it is essential to guarantee that the SGG technology is sufficiently developed and robust to function with a manageable error rate. For other applications such as robotics and augmented reality, models must encompass a wider array of object/relation concepts and exhibit unbiased performance across unbalanced distributions [3]. They are also more likely to involve generating scene graphs from 3D visual data (e.g., point clouds from LIDAR) and video sequences. In applications demanding deeper reasoning, such as multimedia dialogue and visual question answering, we may need more comprehensive structured visual representation (e.g., including additional object attributes, high-order interactions and events) and develop more reliable symbolic reasoning methods. Here, we discuss a few particularly interesting future research directions in more detail.

3D SGG. 3D visual data (i.e., point clouds, RGB-D images) can greatly enhance the understanding and interpretation of complex, real-world scenes. Scene graphs generated from 3D data can capture more accurate object positions, sizes, orientations and spatial relationships between objects, enabling more accurate and comprehensive understanding of the scene's structure and layout. This is particularly important for tasks such as autonomous navigation, robotic manipulation, or human-robot interaction, where understanding the context is crucial. However, the task of 3D SGG introduces several challenges that are unique to 3D data compared to 2D images or video frames. First, 3D data require methods capable of handling 3D geometry and topology for tasks like object recognition, segmentation, and relationship modeling. Developing algorithms that can efficiently process 3D data and capture complex geometric and topological relationships is challenging. Moreover, 3D data like point clouds often have sparse and irregular structures, with varying point densities across different areas of the scene. This sparsity and irregularity can make it challenging to detect objects and relationships accurately and consistently. Finally, processing large-scale 3D data can be computationally intensive, which makes it difficult to generate scene graphs efficiently, especially for real-time applications. Increasingly more works [55, 56, 113, 114] focus on this field, but more research efforts are required to tackle aforementioned challenges.

Video SGG. In real-world deployments, such as autonomous driving and robotics planning, temporal information across continually-collected visual frames is essential and non-negligible. Hence, the analysis of spatio-temporal scene graphs from videos offers a wider range of application scenarios. Video SGG begins to attract more and more attention in recent years, such as [54, 100, 101, 102, 103, 104, 104, 105, 107, 108] and our work in Chapter 6. There are still various open challenges in this field. Firstly, existing works suffer from the seriously long-tailed distribution, such that the predicted video scene graphs bias to recognize simple uninformative relations. Secondly, videos involve more complex semantic interactions that have not been explored in previous research, such as action-action relations (e.g., causality), object-object temporal co-references and high-order events, which require more complex graphical structures and extraction mechanisms [3]. In addition, developing efficient video SGG algorithms that can meet real-time inference requirements using limited computing resources (i.e., embedded systems in edge computing and robotics) is also an important research direction.

Open-world SGG. In real-world scenarios, we encounter open-world concepts such as unbounded types of objects and relations. Nevertheless, most developed SGG approaches are restricted to recognizing limited objects and relations pre-defined in training datasets. Though recent works [190, 201, 267] attempt to address this challenge by exploiting pre-training knowledge from open-world training corpora. The developed methods can achieve open-world object detection, but they are still incapable of recognizing open-vocabulary relations in the open world. This is because objects usually have explicit visual presentations that exhibit relatively small intra-class variation, while relations are more ambiguous. For example, the relation "on" can represent different meanings such as "above with surface touching", "wearing" or "being carried by somebody". Even if some pre-trained models have seen a large corpus, it is still challenging for them to distinguish relation concepts. Inspired by the amazingly good multi-modal large model like GPT-4 [268], we can expect that this challenge can be solved by these large models.

Reasoning over scene graphs. This refers to the process of drawing inferences, making decisions, or solving problems using a structured, abstract representation of a visual scene. Since scene graphs are symbolic, we naturally ask whether conventional symbolic reasoning (e.g., first-order logic [147, 269]) could be used. Symbolic reasoning operates on scene graphs makes sense of the visual information and performs high-level tasks, which can be potentially applied in tasks like video anomaly detection, video event extraction, knowledge discovery over cross-media knowledge bases etc. The merit of symbolic reasoning is to align with human cognition, as humans tend to reason about the world using symbols and relationships. Such alignment can facilitate human-computer interaction and help make AI systems more interpretable and user-friendly. Challenges in symbolic reasoning over scene graphs include: 1) accurate and robust scene graph generation is a prerequisite for effective symbolic reasoning. This imposes higher demands for SGG techniques, and symbolic reasoning must be able to make reasonable inferences despite the imperfections in the scene graph (e.g., incomplete or ambiguous information about the scene). 2) Symbolic reasoning can become intractable within large-scale scene graphs due to combinatorial explosion. 3) Effective symbolic reasoning often requires integrating commonsense knowledge about the world, while it is an ongoing challenge to incorporate such knowledge. Neural-symbolic methods [16, 147, 148, 270] – the combination of symbolic methods with deep-learning methods, is also a trending research direction for scene graph reasoning.

Publications

Here is a list of my publications during the Ph.D. study:

- **Yong Zhang**, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. “Boosting Scene Graph Generation with Visual Relation Saliency." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) (2022).
- **Yong Zhang**, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. “Exploring Structure-aware Transformer over Interaction Proposals for Human-Object Interaction Detection." CVPR 2022.
- **Yong Zhang**, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. “Learning to Generate Language-supervised and Open-vocabulary Scene Graph using Pre-trained Visual-Semantic Space." CVPR 2023.
- **Yong Zhang**, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. “End-to-End Video Scene Graph Generation with Temporal Propagation Transformer.", IEEE Transactions on Multimedia (TMM), 2023.
- Panwen Hu, **Yong Zhang**, Rui Huang. “A Multi-purpose Automatic Editing System based on Lecture Semantics for Remote Education.", IEEE Transactions on Multimedia (TMM), 2023. (under review)

Bibliography

[1] Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, Mei-Ling Shyu, and SS Iyengar. Multimedia big data analytics: A survey. *ACM computing surveys (CSUR)*, 2018. 1

[2] Chang Wen Chen. Internet of video things: Next-generation iot with visual sensors. *IEEE Internet of Things Journal*, 2020. 1

[3] Alireza Zareian. *Learning Structured Representations for Understanding Visual and Multimedia Data*. Columbia University, 2021. 1, 88

[4] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015. 1, 2, 7, 8, 11, 15, 16, 34, 36, 68, 70

[5] Thomas N Kipf et al. *Deep learning with graph-structured representations*. 2020. 2

[6] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020. 2, 8, 11, 34

[7] Yongzhi Li, Duo Zhang, and Yadong Mu. Visual-semantic matching by exploring high-order attention and distraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12795, 2020. 2, 8, 11, 34

[8] Brigit Schroeder and Subarna Tripathi. Structured query-based image retrieval using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 2, 8, 11, 34

[9] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2, 8, 12, 34, 68

[10] Xiangyang Li and Shuqiang Jiang. Know more say less: Image captioning based on scene graphs. *IEEE Transactions on Multimedia*, 2019. 2, 8, 12, 68

[11] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 10323–10332, 2019. 2, 8, 12

[12] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2, 8, 12, 34

[13] Victor Milewski, Marie-Francine Moens, and Iacer Calixto. Are scene graphs good enough to improve image captioning? *arXiv preprint arXiv:2009.12313*, 2020. 2, 8, 12, 34

[14] Cheng Zhang, Wei-Lun Chao, and Dong Xuan. An empirical study on leveraging scene graphs for visual question answering. *arXiv preprint arXiv:1907.12133*, 2019. 2, 5, 8, 12

[15] Drew A Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. *arXiv preprint arXiv:1907.03950*, 2019. 2, 8, 12

[16] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2, 8, 12, 68, 89

[17] Vinay Damodaran, Sharanya Chakravarthy, Akshay Kumar, Anjana Umapathy, Teruko Mitamura, Yuta Nakashima, Noa Garcia, and Chenhui Chu. Understanding the role of scene graphs in visual question answering. *arXiv preprint arXiv:2101.05479*, 2021. 2, 5, 8, 12

[18] Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, et al. Gaia: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020. 2, 6, 8, 13, 34

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012. 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 2015. 5

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 2016. 5, 9, 19, 36, 68, 70

[22] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 5, 9, 51, 52, 54, 63, 68, 71, 72, 73, 74, 77

[23] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, 2016. 5, 9

[24] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5, 9, 68, 71, 72, 78

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 5, 9

[26] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014. 5, 8

[27] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 5

[28] Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. Seqformer: a frustratingly simple model for video instance segmentation. *arXiv preprint arXiv:2112.08275*, 2021. 5, 71

[29] Jialun Pei, Tianyang Cheng, He Tang, and Chuanbo Chen. Transformer-based efficient salient instance segmentation networks with orientative query. *IEEE Transactions on Multimedia*, pages 1–1, 2022. doi: 10.1109/TMM.2022.3141891. 5, 71

[30] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 5

[31] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *ECCV*, 2020. 5, 34

[32] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 5, 11, 27

[33] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 5, 11, 29

[34] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2018. 5, 27, 30, 31

[35] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020. 5, 26, 30, 31

[36] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 5

[37] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *CVPR*, 2017. 5, 16

[38] Tianwen Qian, Jingjing Chen, Shaoxiang Chen, Bo Wu, and Yu-Gang Jiang. Scene graph refinement network for visual question answering. *IEEE Transactions on Multimedia*, pages 1–1, 2022. doi: 10.1109/TMM.2022.3169065. 5, 68

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 28, 78

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 6, 71, 75, 77

[41] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*. Springer, 2016. 6, 83

[42] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017. 6, 8, 10, 14, 17, 19, 21, 22, 23, 26, 34, 36, 41, 42, 43, 44, 46, 47, 68, 70

[43] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Sketching image gist: Human-mimetic hierarchical scene graph generation. In *ECCV*, 2020. 6, 8, 11, 12, 14, 16, 17, 18, 23, 24

[44] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, 2005. 6

[45] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 2003. 6

[46] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010. 6

[47] Jun Yang, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, 2007. 6

[48] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 2015. 7

[49] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 7

[50] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7, 71

[51] Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor. An image is worth 16x16 words, what is a video worth? *arXiv preprint arXiv:2103.13915*, 2021. 7

[52] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022. 7

[53] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 2017. 7, 8, 13, 14, 17, 23, 34, 36, 41

[54] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 7, 8, 13, 68, 70, 78, 83, 84, 88

[55] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2020. 7, 8, 88

[56] Chaoyi Zhang, Jianhui Yu, Yang Song, and Weidong Cai. Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9705–9715, 2021. 7, 8, 88

[57] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 8, 9, 17

[58] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 8, 9, 52

[59] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5678–5686, 2017. 8, 9

[60] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE conference on computer vision and Pattern recognition*, 2017. 8, 9

[61] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *CVPR*, 2019. 8, 9, 11, 17, 26, 36, 44, 70, 83

[62] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018. 8, 10, 14, 17, 19, 23, 24, 26, 28, 29, 30, 31, 34, 36, 42, 43, 44, 45, 47, 68, 70, 83

[63] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *ECCV*, 2018. 8, 11, 17, 26, 68, 70

[64] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive relational networks for mapping images to scene graphs. In *CVPR*, 2019. 8, 11, 19

[65] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. *arXiv preprint arXiv:2104.00308*, 2021. 8, 11, 13, 14, 17, 23, 24, 26, 34, 36

[66] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *ECCV*, 2018. 8, 11, 17

[67] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. Counterfactual critic multi-agent training for scene graph generation. In *ICCV*, 2019. 8, 11, 17

[68] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *CVPR*, 2020. 8, 11, 17, 26, 44, 83

[69] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly supervised visual semantic parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 8, 11, 36, 45

[70] Boris Knyazev, Harm de Vries, Cătălina Cangea, Graham W Taylor, Aaron Courville, and Eugene Belilovsky. Graph density-aware losses for novel compositions in scene graph generation. *arXiv preprint arXiv:2005.08230*, 2020. 8, 11

[71] Jingyi Zhang, Yong Zhang, Baoyuan Wu, Yanbo Fan, Fumin Shen, and Heng Tao Shen. Dual resgcn for balanced scene graphgeneration. *arXiv preprint arXiv:2011.04234*, 2020. 8, 11

[72] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. *NeurIPS*, 2018. 8, 11

[73] Hengyue Liu, Ning Yan, Masood S Mortazavi, and Bir Bhanu. Fully convolutional scene graph generation. *arXiv preprint arXiv:2103.16083*, 2021. 8, 11, 43, 44

[74] Rongjie Li, Songyang Zhang, and Xuming He. Sgtr: End-to-end scene graph generation with transformer. In *CVPR*, 2022. 8, 11, 43, 44, 68, 70

[75] Xin Lin, Changxing Ding, Yibing Zhan, Zijian Li, and Dacheng Tao. Hl-net: Heterophily learning network for scene graph generation. In *CVPR*, 2022. 8, 14, 34, 36, 43, 44

[76] Xin Lin, Changxing Ding, Jing Zhang, Yibing Zhan, and Dacheng Tao. Ru-net: Regularized unrolling network for scene graph generation. In *CVPR*, 2022. 8, 14, 34, 36, 43, 44

[77] Suprosanna Shit, Rajat Koner, Bastian Wittmann, Johannes Paetzold, Ivan Ezhov, Hongwei Li, Jiazhen Pan, Sahand Sharifzadeh, Georgios Kaissis, Volker Tresp, et al. Relationformer: A unified framework for image-to-graph generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 422–439. Springer, 2022. 8, 11, 70

[78] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020. 8, 13, 14, 17, 18, 19, 23, 24, 26, 28, 29, 30, 31, 34, 36, 41, 43, 44, 47

[79] Jing Yu, Yuan Chai, Yue Hu, and Qi Wu. Cogtree: Cognition tree loss for unbiased scene graph generation. *arXiv preprint arXiv:2009.07526*, 2020. 8, 11, 13

[80] He Huang, Shunta Saito, Yuta Kikuchi, Eiichi Matsumoto, Wei Tang, and Philip S Yu. Addressing class imbalance in scene graph parsing by learning to contrast and score. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 8, 13

[81] Gengcong Yang, Jingyi Zhang, Yong Zhang, Baoyuan Wu, and Yujiu Yang. Probabilistic modeling of semantic ambiguity for scene graph generation. *arXiv preprint arXiv:2103.05271*, 2021. 8, 13

[82] Xiaoguang Chang, Teng Wang, Changyin Sun, and Wenzhe Cai. Biasing like human: A cognitive bias framework for scene graph generation. *arXiv preprint arXiv:2203.09160*, 2022. 8

[83] Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1581–1590, 2021. 8

[84] Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, and Yueting Zhuang. Visually-prompted language model for fine-grained scene graph generation in an open world. *arXiv preprint arXiv:2303.13233*, 2023. 8

[85] Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. Learning to generate scene graph from natural language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1823–1834, 2021. 8, 13, 36, 40, 41, 43, 44, 45

[86] Jing Shi, Yiwu Zhong, Ning Xu, Yin Li, and Chenliang Xu. A simple baseline for weakly-supervised scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16393–16402, 2021. 8, 13, 36, 41, 43, 45

[87] Keren Ye and Adriana Kovashka. Linguistic structures as weak supervision for visual scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8289–8299, 2021. 8, 13, 36, 41, 43, 45

[88] Jianming Lv, Qinzhe Xiao, and Jiajie Zhong. Avr: Attention based salient visual relationship detection. *arXiv preprint arXiv:2003.07012*, 2020. 8, 9, 10, 14, 17, 18

[89] Fan Yu, Haonan Wang, Tongwei Ren, Jinhui Tang, and Gangshan Wu. Visual relation of interest detection. In *ACM MM*, 2020. 8, 10, 14, 17, 18

[90] Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Not all relations are equal: Mining informative labels for scene graph generation. *CVPR*, 2022. 8

[91] Wenbin Wang, Ruiping Wang, and Xilin Chen. Topic scene graph generation by attention distillation from caption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15900–15910, 2021. 8

[92] Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Not all relations are equal: Mining informative labels for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15596–15606, 2022. 8

[93] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. In *Proceedings of the AAAI conference on artificial intelligence*, 2019. 8, 10, 14

[94] Motoharu Sonogashira, Masaaki Iiyama, and Yasutomo Kawanishi. Towards open-set scene graph generation with unknown objects. *IEEE Access*, 10:11574–11583, 2022. 8, 35

[95] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 8, 12, 68

[96] Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D Hager, Federico Tombari, and Christian Rupprecht. Semantic image manipulation using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 8, 12

[97] Sarthak Garg, Helisa Dhamo, Azade Farshad, Sabrina Musatian, Nassir Navab, and Federico Tombari. Unconditional scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16362–16371, 2021. 8

[98] Azade Farshad, Sabrina Musatian, Helisa Dhamo, and Nassir Navab. Migs: Meta image generation from scene graphs. *BMVC*, 2021. 8

[99] Qiushuo Zheng, Hao Wen, Meng Wang, and Guilin Qi. Visual entity linking via multi-modal learning. *Data Intelligence*, 4(1):1–19, 2022. 8, 13

[100] Meng-Jiun Chiou, Chun-Yu Liao, Li-Wei Wang, Roger Zimmermann, and Jiashi Feng. St-hoi: A spatial-temporal baseline for human-object interaction detection in videos. In *Proceedings of the 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval*, pages 9–17, 2021. 8, 68, 70, 78, 79, 80, 83, 88

[101] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13688–13697, 2021. 8, 68, 70, 88

[102] Ning Wang, Guangming Zhu, Liang Zhang, Peiyi Shen, Hongsheng Li, and Cong Hua. Spatio-temporal interaction graph parsing networks for human-object interaction recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4985–4993, 2021. 8, 68, 70, 71, 88

[103] Jingwei Ji, Rishi Desai, and Juan Carlos Niebles. Detecting human-object relationships in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8106–8116, 2021. 8, 68, 70, 71, 83, 84, 88

[104] Shengyu Feng, Subarna Tripathi, Hesham Mostafa, Marcel Nassar, and Somdeb Majumdar. Exploiting long-term dependencies for generating dynamic scene graphs. In *arXiv preprint arXiv:2112.09828*, 2021. 8, 68, 70, 71, 84, 88

[105] Ning Wang, Guangming Zhu, Liang Zhang, Peiyi Shen, Hongsheng Li, and Cong Hua. Spatio-temporal interaction graph parsing networks for human-object interaction recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4985–4993, 2021. 8, 68, 70, 71, 88

[106] Kaifeng Gao, Long Chen, Yulei Niu, Jian Shao, and Jun Xiao. Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs. *CVPR*, 2022. 8

[107] Wenqing Wang, Yawei Luo, Zhiqing Chen, Tao Jiang, Lei Chen, Yi Yang, and Jun Xiao. Taking a closer look at visual relation: Unbiased video scene graph generation with decoupled label learning. *arXiv preprint arXiv:2303.13209*, 2023. 8, 88

[108] Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K Roy Chowdhury. Unbiased scene graph generation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 8, 88

[109] Anoop Cherian, Chiori Hori, Tim K Marks, and Jonathan Le Roux. (2.5+ 1) d spatio-temporal scene graphs for video question answering. *arXiv preprint arXiv:2202.09277*, 2022. 8

[110] Wanze Xie, Junshen K Chen, and Alan Zelun Luo. Towards compositional action recognition with spatio-temporal graph neural network. 2020. 8, 13

[111] Masoud Pourreza, Mohammadreza Salehi, and Mohammad Sabokrou. Anograph: Learning normal scene contextual graphs to detect video anomalies. *arXiv preprint arXiv:2103.10502*, 2021. 8, 13

[112] Saeid Amiri, Kishan Chandan, and Shiqi Zhang. Reasoning with scene graphs for robot planning under partial observability. *IEEE Robotics and Automation Letters*, 7(2):5560–5567, 2022. 8, 13, 34, 35

[113] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7515–7525, 2021. 8, 88

[114] Changsheng Lv, Mengshi Qi, Xia Li, Zhengyuan Yang, and Huadong Ma. Revisiting transformer for point cloud-based 3d scene graph generation. *arXiv preprint arXiv:2303.11048*, 2023. 8, 88

[115] Helisa Dhamo, Fabian Manhardt, Nassir Navab, and Federico Tombari. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16352–16361, 2021. 8, 12

[116] Ege Özsoy, Evin Pınar Örnek, Ulrich Eck, Tobias Czempiel, Federico Tombari, and Nassir Navab. 4d-or: Semantic scene graphs for or domain modeling. *arXiv preprint arXiv:2203.11937*, 2022. 8

[117] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alexander G Hauptmann. A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 7

[118] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2015. 9

[119] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2015. 9

[120] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2017. 9, 40, 55, 76

[121] Wei Zhang, Ting Yao, Shiai Zhu, and Abdulmotaleb El Saddik. Deep learning–based multimedia analytics: a review. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2019. 9

[122] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *CVPR 2011*, 2011. 9

[123] Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. On exploring undetermined relationships for visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5128–5137, 2019. 9, 10

[124] Vincent S Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Re, and Li Fei-Fei. Scene graph prediction with limited labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 9, 10, 13

[125] Ranjay Krishna, Ines Chami, Michael Bernstein, and Li Fei-Fei. Referring relationships. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6867–6876, 2018. 9

[126] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 10, 52, 59

[127] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 10, 50, 52, 60, 61

[128] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021. 10, 51, 52, 53, 60, 61, 63, 64

[129] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, 2020. 10, 50, 51, 52, 60, 61

[130] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. *arXiv preprint arXiv:2104.13682*, 2021. 10, 51, 52, 53, 59, 60, 61, 63, 64, 66, 71

[131] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *CVPR*, 2021. 10, 52, 60, 61

[132] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. *CVPR*, 2021. 10, 51, 52, 53, 60, 61, 63, 64, 71

[133] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *CVPR*, 2019. 11, 17

[134] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, 2019. 11, 14, 17, 24, 26, 34, 36, 42, 43, 44, 47, 68, 70, 83

[135] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. *Advances in Neural Information Processing Systems*, 2017. 11

[136] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Exploring context and visual pattern of relationship for scene graph generation. In *CVPR*, 2019. 11, 17, 19, 23

[137] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *ICCV*, 2017. 11, 16, 18, 26, 83

[138] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *CVPR*, 2019. 11, 34, 36

[139] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. *arXiv preprint arXiv:2001.02314*, 2020. 11, 18, 34, 36

[140] Alireza Zareian, Zhecan Wang, Haoxuan You, and Shih-Fu Chang. Learning visual commonsense for robust scene graph generation. *arXiv preprint arXiv:2006.09623*, 2020. 11

[141] Sahand Sharifzadeh, Sina Moayed Baharlou, and Volker Tresp. Classification by attention: Scene graph classification with prior knowledge. *arXiv preprint arXiv:2011.10084*, 2020. 11, 34, 36

[142] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, 2015. 11, 13, 36, 40, 46

[143] Max H Quinn, Erik Conser, Jordan M Witte, and Melanie Mitchell. Semantic image retrieval via active grounding of visual situations. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 172–179. IEEE, 2018. 11

[144] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4654–4662, 2019. 11

[145] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 12, 16, 34

[146] Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, and Xuanjing Huang. Storytelling from an image stream using scene graphs. In *AAAI*, 2020. 12

[147] Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. Neuro-symbolic visual reasoning: Disentangling "visual" from "reasoning". In *International Conference on Machine Learning*, pages 279–290. PMLR, 2020. 12, 89

[148] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019. 12, 89

[149] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016. 13

[150] Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. Cross-media structured common space for multimedia event extraction. *arXiv preprint arXiv:2005.02472*, 2020. 13, 34

[151] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1405–1414, 2020. 13, 34

[152] Yufan Hu, Junyu Gao, and Changsheng Xu. Learning scene-aware spatio-temporal gnns for few-shot early action prediction. *IEEE Transactions on Multimedia*, 2022. 13

[153] NF Chen, Zhiyuan Du, and Khin Hua Ng. Scene graphs for interpretable video anomaly classification. In *Conference on Neural Information Processing Systems Workshop on Visually Grounded Interaction and Language*, 2018. 13

[154] Shaotian Yan, Chen Shen, Zhongming Jin, Jianqiang Huang, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. Pcpl: Predicate-correlation perception learning for unbiased scene graph generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 13

[155] Tao He, Lianli Gao, Jingkuan Song, Jianfei Cai, and Yuan-Fang Li. Learning from the scene and borrowing from the rich: Tackling the long tail in scene graph generation. *arXiv preprint arXiv:2006.07585*, 2020. 13

[156] Hai Wan, Yonghao Luo, Bo Peng, and Wei-Shi Zheng. Representation learning for scene graph completion via jointly structural and visual embedding. In *IJCAI*, 2018. 13

[157] Tzu-Jui Julius Wang, Selen Pehlivan, and Jorma Laaksonen. Tackling the unannotated: Scene graph generation with bias-reduced models. *arXiv preprint arXiv:2008.07832*, 2020. 13

[158] Yuan Yao, Ao Zhang, Xu Han, Mengdi Li, Cornelius Weber, Zhiyuan Liu, Stefan Wermter, and Maosong Sun. Visual distant supervision for scene graph generation. *arXiv preprint arXiv:2103.15365*, 2021. 13

[159] Xingchen Li, Long Chen, Wenbo Ma, Yi Yang, and Jun Xiao. Integrating object-aware and interaction-aware knowledge for weakly supervised scene graph generation. In *ACM MM*, 2022. 14, 34, 36, 41, 43, 44, 45

[160] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *ECCV*, 2018. 14, 34, 36

[161] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE transactions on image processing*, 2015. 16

[162] Xin Xu, Shiqin Wang, Zheng Wang, Xiaolong Zhang, and Ruimin Hu. Exploring image enhancement for salient object detection in low light images. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2021. 16

[163] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 20

[164] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 20

[165] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *ECML*, 2001. 21

[166] HT Lin and L Li. Ordinal regression by extended binary classifications. *NeurIPS*, 2007. 21

[167] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, 2016. 21

[168] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank-consistent ordinal regression for neural networks. *arXiv preprint arXiv:1901.07884*, 2019. 21

[169] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 24

[170] Kaihua Tang. A scene graph generation codebase in pytorch, 2020. `https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch`. 25, 26, 42

[171] Yehao Li, Yingwei Pan, Ting Yao, Jingwen Chen, and Tao Mei. Scheduled sampling in vision-language pretraining with decoupled encoder-decoder network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8518–8526, 2021. 26

[172] Yehao Li, Jiahao Fan, Yingwei Pan, Ting Yao, Weiyao Lin, and Tao Mei. Uni-eden: Universal encoder-decoder network by multi-granular vision-language pre-training. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2022. 26

[173] Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Hongyang Chao, and Tao Mei. Coco-bert: Improving video-language pre-training with contrastive cross-modal matching and denoising. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5600–5608, 2021. 26, 36

[174] Yingwei Pan, Yehao Li, Jianjie Luo, Jun Xu, Ting Yao, and Tao Mei. Auto-captions on gif: A large-scale video-sentence dataset for vision-language pre-training. *arXiv preprint arXiv:2007.02375*, 2020. 26

[175] Fangxiang Feng, Xiaojie Wang, Ruifan Li, and Ibrar Ahmad. Correspondence autoencoders for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2015. 27

[176] Chengyuan Zhang, Jiayu Song, Xiaofeng Zhu, Lei Zhu, and Shichao Zhang. Hcmsl: Hybrid cross-modal similarity learning for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2021. 27

[177] Yehao Li, Yingwei Pan, Jingwen Chen, Ting Yao, and Tao Mei. X-modaler: A versatile and high-performance codebase for cross-modal analytics. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3799–3802, 2021. 27

[178] Jie Wu, Haifeng Hu, and Yi Wu. Image captioning via semantic guidance attention and consensus selection strategy. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2018. 27

[179] Anqi Wang, Haifeng Hu, and Liang Yang. Image captioning with affective guiding and selective attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2018. 27

[180] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2621–2629, 2019. 27

[181] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Pointing novel objects in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12497–12506, 2019. 27

[182] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2014. 27

[183] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 27

[184] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 28

[185] Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, and Yu-Wing Tai. Reflective decoding network for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8888–8897, 2019. 30, 31

[186] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10971–10980, 2020. 31

[187] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICLR*, 2021. 35, 36, 37

[188] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2021. 35, 37

[189] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022. 35, 37, 38, 42, 44, 45

[190] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *Advances in Neural Information Processing Systems*, 2022. 35, 37, 89

[191] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 35, 37

[192] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 35, 37

[193] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *ICLR*, 2021. 35, 37

[194] Jiayuan Mao. Scenegraphparser, 2019. `https://github.com/vacancy/SceneGraphParser` (Access date: 2022-8-11). 36, 46

[195] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 36, 40, 43

[196] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 2021. 36

[197] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr–modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*, 2021. 36

[198] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 2019. 36

[199] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019. 36

[200] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 36, 37

[201] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Towards open-vocabulary scene graph generation with prompt-based finetuning. In *ECCV*, 2022. 37, 41, 46, 47, 89

[202] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 38, 43, 71

[203] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019. 38

[204] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 41

[205] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, 2017. 42, 43, 44

[206] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *CVPR*, 2019. 43, 44

[207] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 45

[208] Boris Knyazev, Harm de Vries, Cătălina Cangea, Graham W Taylor, Aaron Courville, and Eugene Belilovsky. Generative compositional augmentations for scene graph prediction. In *ICCV*, 2021. 47

[209] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *CVPR*, 2021. 47

[210] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015. 50, 52

[211] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 50, 52, 59

[212] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *PAMI*, 2008. 50

[213] Mihai Dogariu, Liviu-Daniel Stefan, Mihai Gabriel Constantin, and Bogdan Ionescu. Human-object interaction: Application to abandoned luggage detection in video surveillance scenarios. In *2020 13th International Conference on Communications (COMM)*, 2020. 50

[214] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 2009. 50

[215] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 50, 52, 60, 61

[216] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018. 50, 52, 60, 61

[217] Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *ECCV*, 2020. 50, 52, 60, 61

[218] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020. 50, 51, 52, 61

[219] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 50, 52, 60, 61, 70, 71, 83

[220] Hai Wang, Wei-shi Zheng, and Ling Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. In *ECCV*, 2020. 50, 52

[221] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *ICCV*, 2019. 50, 52

[222] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, 2020. 50, 51, 52, 60, 61

[223] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021. 51, 52, 53, 60, 61, 63, 64

[224] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, 2021. 52

[225] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 2021. 52

[226] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. *arXiv preprint arXiv:2010.16219*, 2020. 52, 60, 61

[227] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, 2019. 52, 60, 61

[228] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *ACM MM*, 2020. 52, 60, 61

[229] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, 2020. 52

[230] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for human-object interaction detection. In *ECCV*, 2020. 52

[231] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 52

[232] Li Mi and Zhenzhong Chen. Hierarchical graph attention network for visual relationship detection. In *CVPR*, 2020. 52

[233] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *CVPR*, 2019. 52

[234] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Exploiting scene graphs for human-object interaction detection. *arXiv preprint arXiv:2108.08584*, 2021. 52

[235] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. *arXiv preprint arXiv:2012.06060*, 2020. 52

[236] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *ICCV*, 2019. 52

[237] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *ICCV*, 2019. 52

[238] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020. 52

[239] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019. 52

[240] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016. 55

[241] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 57, 75

[242] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 59

[243] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020. 68, 70

[244] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Boosting scene graph generation with visual relation saliency. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2022. 68, 70

[245] Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13874–13883, 2022. 68, 70, 83

[246] Danyang Tu, Wei Sun, Xiongkuo Min, Guangtao Zhai, and Wei Shen. Video-based human-object interaction detection from tubelet tokens. In *Advances in Neural Information Processing Systems*, 2022. 68, 70, 80, 81

[247] Apoorva Agarwal, Rishabh Dabral, Arjun Jain, and Ganesh Ramakrishnan. Skew-robust human-object interactions in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5098–5107, 2023. 68, 70, 80, 81

[248] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 68

[249] Yiming Li, Xiaoshan Yang, Xuhui Huang, Zhe Ma, and Changsheng Xu. Zero-shot predicate prediction for scene graph parsing. *IEEE Transactions on Multimedia*, pages 1–1, 2022. doi: 10.1109/TMM.2022.3155928. 70

[250] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *European Conference on Computer Vision*, pages 606–623. Springer, 2020. 70

[251] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *Proceedings of the 25th ACM international conference on Multimedia*, 2017. 70

[252] Qianwen Cao and Heyan Huang. Attention guided relation detection approach for video visual relation detection. *IEEE Transactions on Multimedia*, pages 1–1, 2021. doi: 10.1109/TMM.2021.3109430. 70

[253] Bingjie Xu, Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Interact as you intend: Intention-driven human-object interaction detection. *IEEE Transactions on Multimedia*, 22(6):1423–1432, 2020. doi: 10.1109/TMM.2019.2943753. 70

[254] Fangao Zeng, Bin Dong, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. *arXiv preprint arXiv:2105.03247*, 2021. 71

[255] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 71

[256] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 71

[257] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2110.06864*, 2021. 71, 82

[258] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468. IEEE, 2016. 71, 82

[259] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48, 2009. 77

[260] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287, 2019. 78

[261] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478, 2019. 80

[262] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16372–16382, 2021. 80, 81, 82, 83, 84

[263] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015. 82

[264] Jonathon Luiten, Aljos A Os Ep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129(2):548–578, 2021. 82

[265] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the American Society of Mechanical Engineers*, 82:35–44, 1960. 82

[266] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 83

[267] Kaifeng Gao, Long Chen, Hanwang Zhang, Jun Xiao, and Qianru Sun. Compositional prompt tuning with motion cues for open-vocabulary video relation detection. In *International Conference on Learning Representations*, 2023. 89

[268] OpenAI. Gpt-4 technical report, 2023. 89

[269] Stuart J Russell. *Artificial intelligence a modern approach.* Pearson Education, Inc., 2010. 89

[270] Weixin Liang, Yanhao Jiang, and Zixuan Liu. Graphvqa: Language-guided graph neural networks for scene graph question answering. *NAACL-HLT 2021*, page 79, 2021. 89