Mastering Scene Understanding: Scene Graphs to the Rescue

Carlos Caetano*, Leo Sampaio Ferraz Ribeiro*, Camila Laranjeira†, Gabriel Oliveira dos Santos*,
Artur Barros*, Caio Petrucci*, Andreza Aparecida dos Santos*, João Macedo†,
Gil Carvalho*, Fabricio Benevenuto†, Jefersson A. dos Santos‡ and Sandra Avila*
*Instituto de Computação (IC), Universidade Estadual de Campinas (UNICAMP), Campinas, Brazil
†Department of Computer Science (DCC), Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
‡School of Computer Science, University of Sheffield, Sheffield, United Kingdom

***Abstract*—The evolution of scene understanding in computer vision has seen remarkable advancements, driven significantly by the development and utilization of scene graphs due to their powerful structural and semantic representation. This structured approach allows for better contextual understanding, facilitating tasks such as image captioning, image generation, image retrieval, human-object interaction, and visual question answering. This tutorial paper aims to comprehensively investigate the current scene graph research by discussing their generation methods, applications, standard datasets, and future development insights.**

I. Introduction

Many of the most discussed applications of computer vision rely on imagery containing a single object or concept. Classification models are always evaluated on ImageNet, and the image generation literature obsesses over people and animals. In this tutorial, we argue for a renewed interest in tasks under the pantheon of Scene Understanding and show promising paths that use the Scene Graph (SG) structure. It is known that, even for a "simple" task such as classification, when a single object is moved, added, or removed from a scene, the label changes [1]. It is paramount, then, to incorporate the information of these relationships within our learning frameworks.

Beyond recognition, SGs were adopted for the tasks of image captioning [2], image generation [3], image retrieval [4], human-object interaction [5] and visual question answering [6]; SGs were first formalized by Johnson et al. [4] as a graph that describes a scene by its individual objects and the relationships (predicate) between those objects (e.g., from simple relationships such as "to the left of" to more specific ones such as "holding" or "touching"). Thus, each SG can also be formulated as a set of visual relationship triplets (i.e., <subject, predicate, object>).

A SG captures detailed semantics by explicitly modeling objects, their attributes, and how objects are related to each other [7, 8]. The success of SGs as representations was such that the novel task of Scene Graph Generation (SGG) was created — as manual annotation of this fine-grained information is infeasible for large datasets. Formally, Jung et al. [9] describe the SGG problem as:

Definition 1. *Given an image I, the goal of SGG is to generate a visually grounded graph $G = (\mathcal{O}, \mathcal{R})$ that represents objects $\mathcal{O}$ and their semantic relationships $\mathcal{R}$ for object classes $\mathcal{C}$ and predicate classes $\mathcal{P}$. An object $o_i \in \mathcal{O}$ is described by a pair of a bounding box $b_i \in [0,1]^4$ and its class label $c_i \in \mathcal{C}$: $o_i = (b_i, c_i)$. A relationship $r_k \in \mathcal{R}$ is represented by a triplet of a subject $o_i \in \mathcal{O}$, an object $o_j \in \mathcal{O}$, and a predicate label $p_{ij} \in \mathcal{P} : r_k = (o_i, o_j, p_{ij})$, which represents relationship p_{ij} between subject o_i and object o_j.*

In this tutorial, we will present the diverse methodologies of SGG, exploring different paradigms and discussing their trade-offs. We will also present the commonly used datasets and evaluation metrics tailored to different research questions and objectives. Furthermore, we will explore the practical applications of SGGs across various domains, illustrating how they can be effectively employed to address real-world tasks. Finally, we will speculate on future trends, highlighting the methods that promise to shape the future of the research field.

II. Scene Graph Generation (SGG)

From the literature, it is possible to perceive two main approach types to SGG: (i) two-stage methods that detect objects and follow with pairwise relationship recognition and (ii) one-stage methods that perform both tasks simultaneously. Figure 1 illustrates the differences between these approaches. We review each within the following sections.

1) Two-stage SGG: The first stage involves finding regions that potentially have objects and classifying these findings. To that end, many studies [9, 11, 12, 13] employ the classical Faster RCNN detector [14] due to its high accuracy and its robustness in handling a wide variety of objects in different scales and aspect ratios.

The second stage comprises attribute and relationship inference. The former refers to object attributes such as color (e.g., yellow), state (e.g., standing), material (e.g., wooden), and others [7]; the latter comprises object pairs [15] and thus provides the predicate information between the objects. When all visual triplets are collected (i.e., <subject, predicate, object>), a SG is constructed.

Based on cognitive psychology research that human beings appear to learn gradually and hierarchically, Jin et al. [12] proposed a contextual augmentation method by employing slight perturbations in the position and size of objects. They produced diverse context descriptions and predicted the likelihood

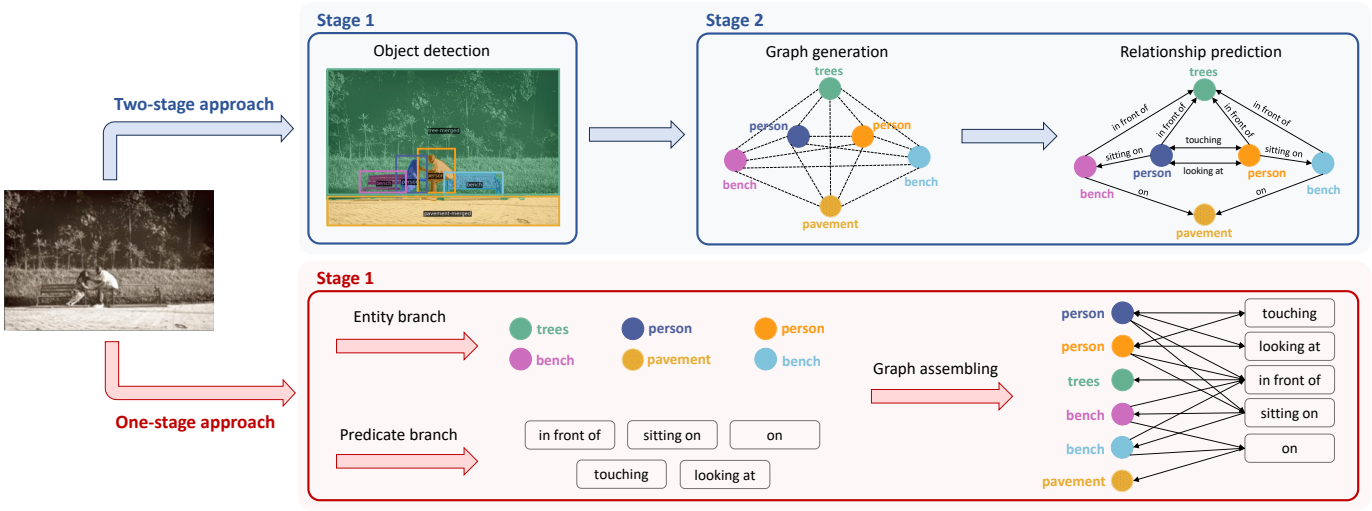

Fig. 1. Example of a two-stage and one-stage scene graph generation pipelines. Figure adapted from Li et al. [7], Yang et al. [10].

and possible predicates between two contextually described objects. Goel et al. [11] framework exploits relation labels based on their informativeness. They categorize the relations into explicit and implicit, based on whether the relation defines the relative spatial configuration between objects. They first train on implicit labels, and a second training stage includes a procedure that imputes missing implicit labels, followed by training on both the annotated and imputed labels.

More recently, Vision Transformers [16] took the place of CNNs as the best-performing models for vision. Given that, SGG works also employed it in their pairwise relationship recognition stage. Jung et al. [9] proposed the Selective Quad Attention Network that learns to select relevant object pairs and disambiguate them via diverse contextual interactions. An edge selection module removes irrelevant object pairs. By modeling the graph generation process as sequential decoding of adjacency lists, Kundu and Aakur [13] effectively models the interaction between detected entities using a simple, directed graph. Their two Transformer-based components sample the underlying interaction graph between the detected entities before reasoning over the sampled semantic structure.

One of the main disadvantages related to two-stage methods is that such approaches need to predict $\mathcal{O}(n^2)$ relation triplets, where n is the number of detected objects, which is computationally expensive [7, 17, 18]. On the other hand, one-stage methods only need to predict a sparse relation candidate set, which is computationally efficient.

2) One-stage SGG: Contrasting with aforementioned studies, one-stage paradigm directly detect object and infer relationships using anchor-free object detection models. They are known to be simple, fast, and easy to train [7].

To harvest the benefits of end-to-end detection methods, studies have extended the Detection Transformer (DETR) [19] to the SGG problem. Yang et al. [10] extended DETR's Hungarian matcher algorithm to be triplet compatible. The method extracts image features from a CNN backbone. Then, it feeds them along with queries and position encoding into a Transformer encoder-decoder. The queries learn a representation of SG triplets, and predictions are computed by three Feed-Forward Networks (FFNs). Cong et al. [20] approach employs an encoder-decoder that directly predicts triplets without inferring predicates between all object pairs. By concatenating the corresponding subject and object representation plus a spatial feature vector (reshaped heatmaps), the predicate probability is predicted by a multi-layer perceptron. Li et al. [18] also adopted the DETR decoder to produce entity nodes from a set of learnable entity queries. Three parallel Transformer decoders were used to predict the generator. Afterward, a differentiable graph assembling module infers the directed edges of a bipartite graph to form the relation triplets.

Liu et al. [17] presented a bottom-up representation of objects and relationships by modeling objects as points and relationships as vectors. Their approach is based on four-headed CNN streams, three of them for object prediction and another for relation prediction. On the contrary, Teng and Wang [21] introduced a top-down method that first uses a set of learnable triplet queries to generate entity and predicate candidate sets. After, the multi-head self-attention mechanism refines the relation detection results progressively.

Although the literature [7, 17, 18] points out that one-stage methods are computationally efficient when compared to two-stage approaches, one main disadvantage of on-stage paradigm is that such methods usually employ complex networks (e.g., attention-based) to learn the queries of entities and predicates. In contrast, two-stage methods have advantages in terms of explainability and interpretability since they usually employ simple classifiers to predict predicates [7].

III. Datasets and Evaluation Metrics

In this section, we list and describe the datasets and metrics used to evaluate the SGG performance.

A. Datasets

Visual Phrase [22] is a dataset containing visual relationships as phrases that encapsulate the relationship of scene composites. It contains 17 types of relationships between 8 objects, with a total of 2,769 images.

Scene Graph [4] contains 5,000 hand-made scene graphs that are image-based and focused on image retrieval. It contains more than 93,000 object instances, 110,000 attribute instances, and 112,000 relationship instances. This dataset prioritizes detailed annotations of images, with each sample containing an average of 21.9 relationships.

Visual Relationship Detection (VRD) [23] is a dataset constructed for benchmarking visual relationship prediction, with 5,000 images with 100 object categories and 70 predicates. In total, it features over 6,500 unique relationship types. This dataset specialty is the presence of rare predicates.

Visual Genome (VG) [24] is a large-scale dataset with 108,077 images, each containing several components: objects, attributes, relationships, region descriptions, question-answer pairs, and scene graphs. There is also a pre-processed version of the VG dataset to improve the quality of object annotations (**VG150** [25]), and to reduce duplicated relationships (**VrR-VG** [26]).

Open Images [27] is a dataset of 9 million images with rich annotations. It has 600 object classes, more than 374.000 visual relationship triplets, and over 15.4 million bounding boxes, making it one of the larger datasets for visual understanding tasks.

SpatialSense [28] was created via adversarial crowdsourcing to include negative examples and challenging reasoning questions. It features 17,498 relations across 11,569 images, involving 3,679 unique object classes, with 2,139 appearing only once, presenting a long-tail distribution of concepts.

SpatialVOC2K [29] includes 2,026 images with 9,804 unique object pairs and was designed for image-to-text generation, focusing on annotating spatial relations between object pairs. The dataset includes images of two or three object-bounding boxes and images with four or five bounding boxes labeled as "difficult" due to their small size.

Panoptic Scene Graph (PSG) [10] contains more than 49,000 samples from 133 classes. The panoptic approach addresses several limitations of bounding box-based labeling, such as imprecise object localization, redundant information, and the inability to capture the entire scene comprehensively.

B. Evaluation Metrics

Evaluation of SGG methods is broken down into three tasks rising in complexity: (i) predicate detection (uses GT object locations and labels) (ii) SG classification (uses GT object location), and (iii) SG detection (only the image is provided).

Recall@K (R@K) is the most common metric; it measures the fraction of times the correct relationship appears among the top K predictions. However, R@K can be influenced by reporting bias due to the class imbalance [30]. To address this issue Chen et al. [31] introduced mean Recall@K (mR@K). Unlike R@K, which may be constrained to one relationship per object pair, mR@K retrieves predicates separately and averages R@K scores across all, providing a balanced evaluation — especially for datasets with long-tail distributions.

IV. Applications

Having discussed the starting task of obtaining Scene Graphs, we focus on how they are employed in scene understanding tasks. We wish to highlight a thread in the history of SG applications. SGs are useful for image captioning, image retrieval, image generation, human-object interaction, and visual question-answering tasks. Most of these uses of SGs were challenged by larger, general methods such as CLIP [32], Stable Diffusion [33], and (Multimodal) LLMs [34]. In this section, we present the history of SGs within these fields but also contrast what has changed since these disruptive publications took the spotlight.

A. Image Captioning

Image captioning is a task in which, given an image, a text that describes its visual content is generated. It demands an understanding of visual objects, attributes, relationships, syntax, and semantics. The challenge of this task comes from the gap between the text and image modalities. To that end, SGs — which can be derived from both modalities — offer a structured representation of scenes and can be easily mapped into textual descriptions, bridging the gap while providing for accuracy and contextual appropriateness.

Jia et al. [35] surveyed works that include SGs in the pipeline of image captioning. They presented [36] as one of the first works to explicitly combine SGs with attention mechanisms. Li and Jiang [36] use a Region Proposal Network [37] to compute object proposals for a given image, which are used to generate visual feature representation and semantic relationship features and then forward them to an LSTM decoder with a hierarchical attention module generating the image caption. Yang et al. [38] incorporated language inductive bias into the encoder-decoder image captioning framework. In their method, an autoencoder is trained to reconstruct a sentence from an SG so that a Graph Neural Network (GNN) learns to map the SG into a dictionary in the bottleneck. Then, the learned dictionary is passed as a language before the decoder, guiding the caption generation. Similarly, Yang et al. [39] proposed a Transformer-based [40] GNN for embedding SGs. The output of the GNN is a series of object/attribute/relation embeddings fed into the decoder to generate the captions. Nguyen et al. [41] presented an approach that relies only on the SG labels without visual features. Their method leverages the spatial location of SG nodes and enhances them with human-object interaction labels. Then, analogous to previously mentioned works, they encode the SG using a GNN and feed a decoder. Yet, Semantic Propositional Image Caption Evaluation [42], a metric designed to evaluate image captions, also relies on SGs. Given a candidate caption and a set of reference captions, it converts them into SGs, and computes an F-score based on the overlap of these SGs. Thus, the semantic meaning the captions share is measured.

B. Image Retrieval

Text-based image retrieval is the task where a user provides a query text as input, and a system returns a ranked list of images, thereby attributing higher ranks to semantically relevant images. It is a well-known task for its wide use in web search engines. Searching collections of images is a natural target of SG usage; after all, the task tackled by Johnson et al. [4] on the study that formalized SGs for the first time.

In the seminal study of Johnson et al. [4], the authors are also tasked with SGG. To that end, they follow the two-stage approach. Conditional Random Fields (CRF) are used to ground query SGs to an image and the obtained probabilities are used to rank images. The follow-up study [43] introduced SG prediction from text queries, relieving users from creating SG queries. Both methods require the grounding probabilities to be computed for every image, amounting to infeasible computational requirements; this problem is considered by Qi et al. [44], where a learnable hashed embedding is employed to represent SGs.

A couple of years later, the study of Conser et al. [45] presented a critique of the method and data presented by Johnson et al. [4], they showed that the model does not make good use of relationships and that samples within the benchmark were biased toward valuing object co-presence only. In the work of Ramnath et al. [46], the CLEVR dataset was adapted to experiment with SGs, and the authors introduced a "Catalog Graph", where the entire searchable database is represented with a single graph. Studies that followed [47, 48] relied on embedding learning through the use of GNNs.

The image retrieval literature changed with the introduction of CLIP [32]. The use of SGs in the field has diminished, but a few recent studies are now integrating SGs within new models. Such is the case for methods from Cong et al. [49] and Liu et al. [50]; in the former, Graph and Image Transformers are used to encode each modality into a common feature space, and, in the latter, embeddings computed through GNNs are compared on object, relationship and scene levels. These promising results show that SGs are still relevant to models that seek compactness and specialization.

C. Image Generation

Image generation is a well-known task in the computer vision pantheon. Some of the definitions of this task allow for images to be generated simply to match a desired image distribution — often the one that yields the training set; this interpretation defined early studies with Generative Adversarial Networks (GANs) [51]. It is desirable, however, to allow users to have finer control over these synthetic creations, and naturally, GANs soon adapted, starting with the so-called Conditional GAN [52]. The literature flourished from this point onwards and not a long time passed until Johnson et al. [3] introduced a Scene-Graph-to-Image generator. SGs are natural candidates to guide image generation as they are a prime space for structurally defining a desired scene design. In this seminal study [3], the authors employed a GCN to process objects and relationships into vector representations that could, in turn, be used to drive bounding box and mask generators. Combining these object masks and boxes creates a "semantic layout" that is then input to a Cascaded Refinement Network to produce an image; the entire process is driven by adversarial training. This setup of generating a semantic layout first and translating this layout to an image was a staple of most SG-to-Image models until very recently.

The studies that followed each improved upon aspects of the typical setup. Ashual and Wolf [53] and Li et al. [54] introduced ways to encode the specific appearance of objects to add to the semantic layout representation and give users more control over the individual objects. Mittal et al. [55] showed that, by training the generative model to add objects to a scene iteratively, one could have a final system that allowed users to iterate on the synthetic image without losing current results. Tripathi et al. [56] introduced heuristics to determine object depth placement — a common issue when composing the semantic layout; their follow-up study [57] showed benefits from changing bounding boxes for eight extreme points prediction for each object. Herzig et al. [58] showed that introducing a canonical definition for SGs (and therefore making the image-SG pair one-to-one) improves generation robustness and generalization.

All aforementioned methods relied on adversarial training and the GAN framework; it is no surprise then that the image quality improvements brought on by diffusion-based methods [59] had a strong impact in the literature of SGs for image synthesis. The first study to incorporate diffusion into the mix came from Yang et al. [60]. Their model employed masked image reconstruction to learn SG representations to use as conditional inputs to a Stable Diffusion model. This is the first notable method not to make use of intermediary scene layouts; others followed. Liu and Liu [61] made use of T5 sentence embeddings to initialize node representations and designed a SG Transformer (SGFormer) to encode SGs. Finally, Wu et al. [62] showed that SGs often do not contain enough details to compose a complete image; they solved this problem through a SG hallucination mechanism that "completes" the graph through discrete diffusion and allows their method to be competitive on image quality even against the likes of DALL-E [63]. While impacted by the change in the image generation literature, the advantages of synthesizing with SGs (control, explainability) have kept the field afloat and rising.

D. Human-Object Interaction

Human-object interaction (HOI) is a task that has a similar definition to that of SGG, with both having as a goal the inference of triplets representing spatial or semantic relationships between a subject and an object. The particularities of HOI, enough to make it a separate task altogether, are twofold. Firstly, the subject is always a human, a non-monolithic entity with diverse interactions for each set of its constituent parts. Thus, it is common to resort to a decomposed view of the human body, with features such as pose [64] or body part bounding boxes [65] as part of the inference pipeline. Secondly, HOI involves high-level predicates with

predicate labels like verbs that represent human actions (riding, wearing, and feeding). The polysemy of such verbs, spatial and semantically, has been a topic of interest in the literature [66].

A recent survey on HOI [67] highlights the importance of graph modeling in two-stage methods, with Qi et al. [68] cited as one of the earliest methods to infer human-object relations as a structured graph. By representing humans and objects as nodes and producing a fully connected graph, Qi et al. [68] proposed a message passing algorithm to iteratively update stronger connections, producing an optimized graph structure and the respective edge labels in an end-to-end fashion. A more recent work [69] relies on a similar message-passing system. However, messages are conditioned on the spatial relationship of a given pair of nodes, i.e., communicating their relative locations. With the success of Transformer-based architectures, a trend of one-stage methods surfaced in recent HOI proposals, as outlined by Antoun and Asmar [67].

He et al. [70] leveraged the intrinsic relationship between SGG and HOI by proposing a hierarchical Transformer-based method to perform both tasks. The proposition is to encode visual and semantic features from the input image, then split the architecture into two decoding branches: SG inference and HOI prediction. These decoders are hierarchically connected, with the output of the SG decoder feeding into the HOI decoder via query transformations. The authors found that joint inference of both tasks significantly improves performance compared to models that handle each task individually.

E. Visual Question Answering

Within the *visual question answering* (VQA) task, the system is given an image and a related question in natural language; the expected output is the correct answer. Accomplishing this requires reasoning over visual elements and, at times, the inclusion of general knowledge (e.g., humans *wear* clothes or *ride* horses). The first notable study to investigate the usefulness of SGs for this task was done by Zhang et al. [6]. Other methods had incorporated SGs before they were used to support question answering through node selection [71]. Zhang et al. [6]'s study was the first to use SGs within the learning protocol by way of a GNN; they also demonstrated that SGs alone contain enough information to answer common questions, corroborating a previous finding [72].

Studies that followed adhered to a common framework: SGs are used with GNNs to aid in the VQA question classification tasks, object nodes are often represented by GloVe [73] word vectors [74, 75] and questions are encoded through recurrent networks [76] or Transformers [77]. As with other SG applications, the emergence of multimodal large language models had a significant impact, the current scenario being that models such as Gemini [34] and PaLI [78] are the SoTA. While progress has slowed in using SGs, there is a demonstrated utility in not relying on large models for all tasks. The method of Wang et al. [79] has shown the effectiveness of using SGs with both a visual and semantic (LM-based) object node representation combined with a knowledge graph and a node representing the question in a single graph. With these developments, we expect future literature on VQA methods to be split between LLMs and compact SG-based methods.

V. Conclusions and Directions

In this tutorial, we have delved into SGs' significant impact on various computer vision tasks. SGs provide a detailed and organized representation of objects and their relationships within an image, significantly improving tasks like image captioning, image generation, image retrieval, human-object interaction, and visual question answering. By enabling a more comprehensive understanding of visual data, SGs allow for more advanced reasoning and interaction capabilities in machine learning models.

While SGs offer significant advantages in computer vision, developing new methods for their generation comes with its own set of challenges. Two-stage methods, which offer better explainability due to simpler classifiers, struggle with high computational costs because they predict $\mathcal{O}(n^2)$ relation triplets. On the other hand, one-stage methods are more efficient in predicting a sparse set of relations, but they use complex networks like Transformers, making them harder to interpret. Balancing computational efficiency with interpretability remains crucial in exploring and developing new techniques for SGG.

Another main challenge is handling rare visual relationships between objects. Due to less common interactions, some relationships are poorly represented, making it hard for models to learn them well. This can lead to vague descriptions of these rare interactions. It is crucial to distinguish between significant relationships and common but less meaningful ones to improve graph reasoning, which is essential for unbiased SGG.

Acknowledgments

This work is partially funded by FAPESP 2023/12086-9, and the Serrapilheira Institute R-2011-37776. C. Caetano, L. Ribeiro, A. Barros, and C. Petrucci are also funded by FAPESP 2024/01210-3, 2022/14690-8, 2024/09372-2, 2024/09375-1, respectively. S. Avila is also funded by FAPESP 2020/09838-0, 2013/08293-7, H.IAAC 01245.003479/2024-10, and CNPq 316489/2023-9.

References

[1] A. López-Cifuentes, M. Escudero-Viñolo *et al.*, "Semantic-Aware Scene Recognition," *Pattern Recognition*, 2020.
[2] L. Gao, B. Wang, and W. Wang, "Image captioning with scene-graph based semantic concepts," in *ICMLC*, 2018.
[3] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *CVPR*, 2018.
[4] J. Johnson, R. Krishna *et al.*, "Image retrieval using scene graphs," in *CVPR*, 2015.
[5] Y. Zhang, Y. Pan *et al.*, "Exploring structure-aware transformer over interaction proposals for human-object interaction detection," in *CVPR*, 2022.
[6] C. Zhang, W.-L. Chao, and D. Xuan, "An empirical study on leveraging scene graphs for visual question answering," in *BMVC*, 2019.
[7] H. Li, G. Zhu *et al.*, "Scene graph generation: A comprehensive survey," *Neurocomputing*, 2024.
[8] X. Chang, P. Ren *et al.*, "A comprehensive survey of scene graphs: Generation and application," *IEEE TPAMI*, 2023.
[9] D. Jung, S. Kim *et al.*, "Devil's on the edges: Selective quad attention for scene graph generation," in *CVPR*, 2023.

[10] J. Yang, Y. Z. Ang *et al.*, "Panoptic scene graph generation," in *ECCV*, 2022.

[11] A. Goel, B. Fernando *et al.*, "Not all relations are equal: Mining informative labels for scene graph generation," in *CVPR*, 2022.

[12] T. Jin, F. Guo *et al.*, "Fast contextual scene graph generation with unbiased context augmentation," in *CVPR*, 2023.

[13] S. Kundu and S. N. Aakur, "Is-sgg: Iterative scene graph generation with generative transformers," in *CVPR*, 2023.

[14] S. Ren, K. He *et al.*, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015.

[15] T. Chen, W. Yu *et al.*, "Knowledge-embedded routing network for scene graph generation," in *CVPR*, 2019.

[16] A. Dosovitskiy, L. Beyer *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[17] H. Liu, N. Yan *et al.*, "Fully convolutional scene graph generation," in *CVPR*, 2021.

[18] R. Li, S. Zhang, and X. He, "Sgtr: End-to-end scene graph generation with transformer," in *CVPR*, 2022.

[19] N. Carion, F. Massa *et al.*, "End-to-end object detection with transformers," in *ECCV*, 2020.

[20] Y. Cong, M. Yang, and B. Rosenhahn, "Reltr: Relation transformer for scene graph generation," *IEEE TPAMI*, 2023.

[21] Y. Teng and L. Wang, "Structured sparse r-cnn for direct scene graph generation," in *CVPR*, 2022.

[22] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *CVPR*, 2011.

[23] C. Lu, R. Krishna *et al.*, "Visual relationship detection with language priors," in *ECCV*, 2016.

[24] R. Krishna, Y. Zhu *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, 2017.

[25] D. Xu, Y. Zhu *et al.*, "Scene graph generation by iterative message passing," in *CVPR*, 2017.

[26] Y. Liang, Y. Bai *et al.*, "Vrr-vg: Refocusing visually-relevant relationships," in *ICCV*, 2019.

[27] A. Kuznetsovaand, H. Rom *et al.*, "The open images dataset v4," *IJCV*, 2020.

[28] K. Yang, O. Russakovsky, and J. Deng, "Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition," in *ICCV*, 2019.

[29] A. Belz, A. Muscat *et al.*, "SpatialVOC2K: A multilingual dataset of images with annotations and features for spatial relations between objects," in *INLG*, 2018.

[30] K. Tang, Y. Niu *et al.*, "Unbiased scene graph generation from biased training," in *CVPR*, 2020.

[31] T. Chen, W. Yu *et al.*, "Knowledge-embedded routing network for scene graph generation," in *CVPR*, 2019.

[32] "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[33] R. Rombach, A. Blattmann *et al.*, "High-Resolution Image Synthesis with Latent Diffusion Models," in *CVPR*, 2022.

[34] G. Team, R. Anil *et al.*, "Gemini: A Family of Highly Capable Multimodal Models," *arXiv:2312.11805*, 2024.

[35] J. Jia, X. Ding *et al.*, "Image captioning based on scene graphs: A survey," *Expert Systems with Applications*, 2023.

[36] X. Li and S. Jiang, "Know more say less: Image captioning based on scene graphs," *IEEE Transactions on Multimedia*, 2019.

[37] R. Girshick, "Fast r-cnn," in *ICCV*, 2015.

[38] X. Yang, K. Tang *et al.*, "Auto-encoding scene graphs for image captioning," in *CVPR*, 2019.

[39] X. Yang, J. Peng *et al.*, "Transforming visual scene graphs to image captions," in *ACL*, 2023.

[40] A. Vaswani, N. Shazeer *et al.*, "Attention is all you need," *NeurIPS*, 2017.

[41] K. Nguyen, S. Tripathi *et al.*, "In defense of scene graphs for image captioning," in *ICCV*, 2021.

[42] P. Anderson, B. Fernando *et al.*, "Spice: Semantic propositional image caption evaluation," in *ECCV*, 2016.

[43] S. Schuster, R. Krishna *et al.*, "Generating semantically precise scene graphs from textual descriptions for improved image retrieval," in *Proceedings of the Fourth Workshop on Vision and Language*, 2015.

[44] M. Qi, Y. Wang, and A. Li, "Online cross-modal scene retrieval by binary representation and semantic graph," in *MM*, 2017.

[45] E. Conser, K. Hahn *et al.*, "Revisiting visual grounding," in *NAACL*, 2019.

[46] S. Ramnath, A. Saha *et al.*, "Scene graph based image retrieval – a case study on the CLEVR dataset," *arXiv:1911.00850*, 2019.

[47] B. Schroeder and S. Tripathi, "Structured query-based image retrieval using scene graphs," in *CVPR Workshops*, 2020.

[48] S. Wang, R. Wang *et al.*, "Cross-modal scene graph matching for relationship-aware image-text retrieval," in *WACV*, 2020.

[49] Y. Cong, W. Liao *et al.*, "SPAN: Learning similarity between scene graphs and images with transformers," *arXiv:2304.00590*, 2024.

[50] Y. Liu, X. Yuan *et al.*, "SEMScene: Semantic-consistency enhanced multi-level scene graph matching for image-text retrieval," *TOMM*, 2024.

[51] I. J. Goodfellow, J. Pouget-Abadie *et al.*, "Generative adversarial nets," in *NeurIPS*, 2014.

[52] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv:1411.1784*, 2014.

[53] O. Ashual and L. Wolf, "Specifying object attributes and relations in interactive scene generation," in *ICCV*, 2019.

[54] Y. Li, T. Ma *et al.*, "PasteGAN: A Semi-Parametric Method to Generate Image from Scene Graph," in *NeurIPS*, 2019.

[55] G. Mittal, S. Agrawal *et al.*, "Interactive image generation using scene graphs," *Journal of King Saud University*, 2019.

[56] S. Tripathi, A. Bhiwandiwalla *et al.*, "Heuristics for image generation from scene graphs," in *ICLR Workshop*, 2019.

[57] S. Tripathi, S. N. Sridhar *et al.*, "Compact scene graphs for layout composition and patch retrieval," in *CVPR Workshops*, 2019.

[58] R. Herzig, A. Bar *et al.*, "Learning canonical representations for scene graph to image generation," in *ECCV*, 2020.

[59] R. Rombach, A. Blattmann *et al.*, "High-Resolution Image Synthesis with Latent Diffusion Models," *arXiv:2112.10752*, 2022.

[60] L. Yang, Z. Huang *et al.*, "Diffusion-based scene graph to image generation with masked contrastive pre-training," *arXiv:2211.11138*, 2022.

[61] J. Liu and Q. Liu, "R3CD: Scene graph to image generation with relation-aware compositional contrastive control diffusion," in *AAAI*, 2024.

[62] S. Wu, H. Fei *et al.*, "Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion," in *NeurIPS*, 2024.

[63] A. Ramesh, M. Pavlov *et al.*, "Zero-shot text-to-image generation," in *ICML*, 2021.

[64] Y.-L. Li, X. Liu *et al.*, "Detailed 2d-3d joint representation for human-object interaction," in *CVPR*, 2020.

[65] P. Zhou and M. Chi, "Relation parsing neural network for human-object interaction detection," in *ICCV*, 2019.

[66] X. Zhong, C. Ding *et al.*, "Polysemy deciphering network for robust human–object interaction detection," *IJCV*, 2021.

[67] M. Antoun and D. Asmar, "Human object interaction detection: Design and survey," *Image and Vision Computing*, 2023.

[68] S. Qi, W. Wang *et al.*, "Learning human-object interactions by graph parsing neural networks," in *ECCV*, 2018.

[69] F. Z. Zhang, D. Campbell, and S. Gould, "Spatially conditioned graphs for detecting human-object interactions," in *ICCV*, 2021.

[70] T. He, L. Gao *et al.*, "Toward a unified transformer-based framework for scene graph generation and human-object interaction detection," *IEEE TIP*, 2023.

[71] S. Ghosh, G. Burachas *et al.*, "Generating Natural Language Explanations for Visual Question Answering using Scene Graphs and Visual Attention," *arXiv:1902.05715*, 2019.

[72] Q. Wu, D. Teney *et al.*, "Visual question answering: A survey of methods and datasets," *CVIU*, vol. 163, 2017.

[73] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *EMNLP*, 2014.

[74] V. Damodaran, S. Chakravarthy *et al.*, "Understanding the Role of Scene Graphs in Visual Question Answering," *arXiv:2101.05479*, 2021.

[75] S. V. Nuthalapati, R. Chandradevan *et al.*, "Lightweight Visual Question Answering using Scene Graphs," in *CIKM*, 2021.

[76] H. Li, X. Li *et al.*, "Joint Learning of Object Graph and Relation Graph for Visual Question Answering," in *ICME*, 2022.

[77] P. Xiong, Q. You *et al.*, "SA-VQA: Structured alignment of visual and semantic representations for visual question answering," *arXiv:2201.10654*, 2022.

[78] X. Chen, X. Wang *et al.*, "PaLI: A jointly-scaled multilingual language-image model," in *ICLR*, 2023.

[79] Y. Wang, M. Yasunaga *et al.*, "VQA-GNN: Reasoning with Multimodal Knowledge via Graph Neural Networks for Visual Question Answering," in *ICCV*, 2023.