HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230124

Representation Learning for Semantic Scene Understanding

Azade FARSHAD [a,b]
[a] *Technical University of Munich*
[b] *Munich Center for Machine Learning (MCML)*

Abstract. Recent advances in semantic scene understanding have underscored its growing significance in the field of computer vision. Enhanced representations can be achieved by incorporating semantic information derived from textual data and applying it to generative models for scene modeling. Nevertheless, the features extracted from text prompts may not seamlessly model a scene.

Scene graphs offer a robust solution to address this challenge, serving as a powerful representation for semantic image generation and manipulation. In this study, we delve into the utilization of scene graphs for this purpose and propose novel methodologies to augment both the representation and learning processes involved in image generation and manipulation.

For image generation, we examine meta-learning for producing images in unprecedented scenes and refine the generated images using an autoregressive scene graph generation model. In terms of image manipulation, we put forth a novel self-supervised method that eliminates the need for paired before-and-after data. Additionally, we boost image manipulation performance by disentangling latent and graph representations in a self-supervised manner.

By evaluating the efficacy of our proposed approaches on a diverse range of publicly available benchmarks, we demonstrate their superiority, ultimately achieving state-of-the-art performance in the domain of semantic image generation and manipulation.

Keywords. Scene Graphs, Representation Learning, Image Generation and Manipulation

1. Introduction

The field of computer vision has witnessed remarkable progress in recent years, with a growing emphasis on the interpretation [1] and understanding of complex scenes. This has led to the increasing importance of integrating semantic information from textual sources into generative models, which can facilitate richer and more detailed scene representations. Despite the promise of these advancements, a crucial problem that remains is the effective modeling of scenes based on features extracted from text prompts, as these features may not always capture the intricacies and relationships inherent in a given scene.

Previous work in semantic scene understanding has primarily focused on leveraging textual information to inform generative models. However, these approaches have exhibited shortcomings in terms of accurately modeling and capturing complex semantic relationships between various scene components. To address this limitation and advance

the state of the art in semantic image generation and manipulation, our study explores the potential of scene graphs as a powerful and versatile solution.

Scene graphs provide a comprehensive representation that encapsulates both the semantic relationships and the structural composition of scenes, offering a more effective means of addressing the challenges associated with the integration of textual information into generative models. Our research is centered around devising innovative methodologies that can enhance both the representation and learning processes involved in image generation and manipulation tasks by harnessing the capabilities of scene graphs.

In summary, this research project includes the following key contributions:

- We propose using meta-learning [2,3,4,5] for generating images in novel scenes.
- We refine the image generation process through the implementation of an autoregressive scene graph generation model.
- We address the challenges associated with image manipulation by introducing a pioneering self-supervised method, eliminating the need for paired before-and-after data.
- We further enhance the performance of image manipulation tasks by proposing a self-supervised disentangling latent and graph representations in a self-supervised manner.

Through extensive experimentation and evaluation on a diverse range of publicly available benchmarks, our research demonstrates the effectiveness of the proposed methodologies in outperforming existing approaches. By achieving state-of-the-art performance in the domain of semantic image generation and manipulation, this study not only contributes valuable insights and advancements to the broader field of computer vision but also lays the groundwork for future research and practical applications in semantic scene understanding.

2. Related Works

Scene Graphs Scene graphs provide a directed graph representation that characterizes an image [6], with objects represented as nodes and their relationships depicted as edges. A wide range of research has delved into generating scene graphs from images [7,8,9,10,11,12,13] and more recently, from point clouds [14,15]. The primary objective of this task is to discern the objects present in a scene and their associated visual relationships. To achieve this goal, various strategies have been investigated, such as iterative message-passing [16], graph decomposition [17], and attention mechanisms [18,19]. Scene graphs have proven to be a potent alternative for conditional scene generation [20,21,22] and manipulation [23], which we will further examine in the subsequent sections.

Image Generation Recent advancements in image generation have predominantly stemmed from Generative Adversarial Networks [24] and diffusion models [25,26]. The research community has delved into conditional variants [27], which facilitate image generation based on various input modalities. For instance, Pix2Pix [28] serves as a model for translating between different image domains, while CycleGAN [29] addresses this task without requiring paired images for training. On the other hand, studies focused on unconditional generation [30,31] are typically domain-specific, such as facial images.

A series of approaches [32,33,34] propose semantic image generation, whereby input semantic maps produce corresponding images. Alternative methods involve image generation from layout [35,36], using bounding boxes and class labels for each scene instance. More closely related to our work are techniques that generate images conditioned on scene graphs [20,37,38], with the layout serving as an intermediate step to translate the graph structure into image space. Johnson et al. [20] pioneered this approach with Sg2im, a supervised method utilizing a combined object-level and image-level GAN loss. Subsequent research has enhanced performance in this challenging task by incorporating per-object neural image features to boost diversity [37] and leveraging contextual information to refine the layout (CoLoR) [38].

Image Manipulation Image generation typically incorporates a user interface to specify the subject of change [39]. Early works in scene-level image editing employed handcrafted techniques, which involved replacing parts of an image with sample patches from a database [40]. One such manipulation method is image inpainting [41], where a user specifies a mask for removal and automatic filling of an image area [42]. This can be further enhanced with semantics [43] or edges [44,45] to guide the missing region. Hong et al. [46] used a learned model on a semantic layout representation, allowing users to modify images by adding, moving, or removing bounding boxes. SESAME [47] enables users to draw a mask with semantic labels on an image to indicate the category of changed pixels. Similarly, EditGAN [48] allows users to alter object appearance by modifying a detailed object part segmentation map [49,50]. SIMSG [23] employs scene graphs as the interface, where users can manipulate images by altering the nodes or edges of a graph. Recently, Su et al. [51] introduced an enhancement to this model by utilizing masks instead of bounding boxes for object placement. Contrasting with these approaches, our objective is to model an object representation that disentangles appearance and pose.

3. Method

In this section, we describe our methodology for enhancing the representation and learning processes involved in image generation and manipulation tasks using scene graphs. Our approach consists of several key components: (1) meta-learning for generating images in novel scenes (MIGS), (2) autoregressive scene graph generation (SGGen), (3) self-supervised semantic image manipulation (SIMSG), and (4) disentangling latent and graph representations (DisPositioNet).

3.1. Definitions

A scene graph $\mathscr{G}$ is a directed graph representation of an image I, where the nodes V represent objects and the edges E represent the semantic relationships between the objects. Formally, a scene graph can be defined as $\mathscr{G} = (V,E)$, where $V = v_1, v_2, \ldots, v_n$ is a set of object nodes and $E = e_{ij}$ is a set of directed edges representing relationships between objects v_i and v_j. Each object node v_i is associated with an object category label $c_i \in \mathscr{C}$, where $\mathscr{C}$ denotes the set of all object categories. Similarly, each edge e_{ij} is associated with a relationship label $r_{ij} \in \mathscr{R}$, where $\mathscr{R}$ denotes the set of all relationship types.

3.2. Scene Graph to Image

The goal of the scene graph to image model (SG2Im) is to learn a model that generates the image I conditioned on the scene graph G. We employ a graph neural network (GNN) to extract features H_o of each object in the scene from the scene graph. The GNN parameters are represented by θ_{GNN}. Then a layout L is constructed after predicting the bounding box and pseudo-segmentation map from the objects. Finally, the decoder network is conditioned on the layout and the image is generated.

Object Feature Extraction The features for each object in the scene H_o are the result of message passing inside the GNN after numerous iterations. The message passing enforces the object embeddings to be updated based on their neighbouring nodes as well as their connected relationships.

Layout Generation Once we have the object embeddings H_o from the graph, we construct a layout L that defines the spatial arrangement of objects in the image. First, we predict the bounding box coordinates x_o for each object from their corresponding embeddings using a multi-layer perceptron (MLP) called boxNet, denoted by θ_{box}. Then, a pseudo-segmentation map is predicted by another MLP that defines the object shape. The layout L is constructed by arranging the object embeddings in the spatial locations of their corresponding bounding box locations. Therefore, the layout becomes similar to a high-dimensional segmentation map with the object embeddings forming its depth information.

Image Synthesis With the constructed layout L, we synthesize the final image I' using a conditional GAN. The generator G of the conditional GAN takes the layout L as input and produces an image I':

$$I' = D(\mathscr{G}; \theta_G), \tag{1}$$

where θ_G denotes the parameters of the generator. The discriminator D of the conditional GAN aims to distinguish between real and generated image pairs, with parameters θ_D. The generator and discriminator are trained in an adversarial manner to optimize the following objective:

$$\mathscr{L}_{GAN} = \min_{\theta_G} \max_{\theta_D} \mathbb{E}_{I \sim p_{\text{data}}}[\log D(I; \theta_D)] + \mathbb{E}_{I' \sim p_{\text{G}}}[\log(1 - D(G(\mathscr{G}; \theta_G); \theta_D))], \tag{2}$$

where I is sampled from the true data distribution p_{data}, and $G(\mathscr{G}; \theta_G)$ represents the generated image conditioned on the scene graph $\mathscr{G}$. In addition to the GAN objective, the model is trained with auxiliary classifier loss, and a bounding box prediction loss:

$$\mathscr{L}_{SG2Im} = \mathscr{L}_{GAN} + \|x' - x\| + \mathscr{L}_{BCE}(c', c) \tag{3}$$

, where for each node in the graph, c and c' are the ground truth and predicted object classes and x, x' are the ground truth and predicted bounding box coordinates, respectively.

By optimizing the above objectives, we learn to generate images from scene graphs that are semantically consistent and visually realistic.

3.3. Meta-Learning for Image Generation

We propose a meta-learning framework (MIGS) to generate images in novel scenes. Given the scene graph $\mathcal{G}$, our goal is to generate the corresponding image I. The generator model G is trained on a set of training scene graphs $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_m$ and corresponding images $I_1, I_2, \ldots, I_m$.

During meta-learning, we employ episodic training, where each episode consists of a support set S and a query set Q. The support set contains a subset of scene graphs and their corresponding images, while the query set contains a separate subset of scene graphs for which we aim to generate images. In each episode, we train multiple networks $\theta_{G,1}, \theta_{G,2}, \ldots, \theta_{G,n}$ on the given support sets. The we average all the model parameters to obtain the model for the current episode:

$$\theta_G = \frac{\Sigma_{n=0}^{N} \theta_{G,n}}{N} \tag{4}$$

Afterwards, the model parameters are updated similar to SG2Im as in Equation 3.

3.4. Graph representation learning via unconditional scene graph generation

To improve the quality of generated images, we model the generation process using an autoregressive scene graph generation model, SGGen [52]. Given a scene graph $\mathcal{G}_i = (V_i, E_i)$, we aim to generate images by predicting one node at a time in a sequential manner. The SGGen model is a deep auto-regressive generative model that learns the probability distribution over labeled and directed graphs. It generates a scene graph in a sequence of steps, with each step producing an object node followed by a sequence of relationship edges connecting to the previous nodes.

3.5. Semantic image manipulation using scene graphs

To address the challenges associated with image manipulation, we propose a self-supervised method that eliminates the need for paired before-and-after data. We utilize image reconstruction as a proxy task, where the image and scene graph are partially randomly masked with Gaussian noise. Let I represent the input image, and $\mathcal{G}$ and $\mathcal{G}'$ denote the original and manipulated scene graphs, respectively. Our goal is to reconstruct the image using the information in the scene graph. Using this proxy task, we learn a model ϕ, that generates a manipulated image I' at inference time based on the manipulation mode and the target scene graph $\mathcal{G}'$.

We optimize the parameters ϕ by minimizing the reconstruction loss function $\mathcal{L}_{rec}$, in addition to the GAN and auxiliary losses similar to MIGS [53] and SG2Im as shown in Equation 3:

$$\mathcal{L}_{rec} = \|I' - I\| \tag{5}$$

Additionally, the model is trained with perceptual loss (L_{LPIPS}) [54] to improve the realism of the reconstructed images. Then the total loss becomes:

$$\mathscr{L}_{SIMSG} = \mathscr{L}_{rec} + \mathscr{L}_{SG2Im} + \mathscr{L}_{LPIPS} \tag{6}$$

Disentangled Representation Learning To further improve the performance of image manipulation tasks, we propose a method called DisPositoNet [55] to disentangle the representations within the graph neural network and the latent space of the GAN. This ensures that certain features such as pose and appearance be preserved in the image manipulation process, while the other features change.

The disentanglement is performed on graph-level features by adapting the DisenGCN [56] architecture to scene graphs. The original framework was designed to disentangle the node features in the graph. Here, we modify the model to consider the edge level features as well. This results in object embeddings H'_o, which are distangled through a neighbourhood routing mechanism in a self-supervised manner based on the effect of their neighbour nodes.

The latent space disentanglement is done through modeling the latent space with two variational autoencoders (VAE). We encode the disentangled object embeddings H'_o using the pose and appearance encoders E_p, E_a, respectively. The pose representation H_p captures the spatial arrangement of objects in the scene, while the appearance representation H_a encodes the visual features of the image. By disentangling these representations in the latent space of the GAN, we aim to better model the relationships between different elements in the scene.

Since the data does not have any annotations regarding the pose or appearance, we disentangle the features by learning a transformation function ω, using the pose VAE by having the pose decoder $\mathscr{G}_p$ predict the affine transformation parameters. On the other hand, the second decoder $L' = \mathscr{G}_a$ is supposed to predict the scene layout without the pose information. To reconstruct the final image, we generate the original scene layout L by applying the affine transformation function using the predicted parameters to the non-transformed scene layout L':

$$L = \omega(L') \tag{7}$$

In addition to the SIMSG [23] objective functions provided in Equation 6, we add variational loss on the two VAEs in the latent space which try to minimize the KL divergence between the data distribution in the latent space and the normal distribution.

4. Experiments and Results

In this section, we present the results of our experiments and discuss the performance of our proposed methodology for semantic image generation and manipulation. The evaluation was conducted on multiple public benchmarks, which allowed for a comprehensive assessment of our approach in comparison to existing state-of-the-art methods.

4.1. Semantic Image Generation

We show the performance of the SGGen [52] and MIGS [53] on the task of image generation from scene graphs. In addition, we show the graph generation error for SGGen.

Quantitative Results To compare the distribution of the generated scene graphs by SGGen with the ground truth scene graphs, we measure the Maximum Mean Discrepancy (MMD) distance between them. This is shown in Table 1. In addition, we condition SG2Im [20] on the scene graphs generated by SGGen and compute the FID, Inception Score and Precision / Recall between the ground truth and generated images.

We evaluate MIGS on BDD100k [57] and Visual Genome (VG) [58] datasets, and show the quantitative results in Table 2 and Table 3, respectively. We evaluate the model in few-shot (with 5 and 10 samples) setting, as well as full data (160-shot) setting.

Table 1. Quantitative evaluation of the graph samples (left) and image samples (right)

Model	Ordering	Graph		Image			
		MMD node ($\times 10^3$) ↓	MMD graph ($\times 10^3$) ↓	FID ↓	IS ↑	Precision ↑	Recall ↑
GraphRNN [59]	BFS	2.3	1.3	75.8	4.88	0.680	0.660
	Random	0.39	1.2	74.5	4.85	0.679	0.664
SGGen [52]	BFS	2.05	1.82	73.3	5.04	0.679	0.690
	Hierarchical	1.85	0.63	72.2	**5.26**	0.717	**0.714**
	Random	**0.37**	**0.11**	**71.2**	4.95	**0.727**	**0.714**
Ground Truth		0.018	0.023	73.0	5.22	0.693	0.707

Table 2. Quantitative results on BDD100k fine-tuned on 5,10 and 160 shots.

Method	Decoder	FID ↓	KID $\cdot 10^3$ ↓	FID ↓	KID $\cdot 10^3$ ↓	FID ↓	KID $\cdot 10^3$ ↓
		160-shot		10-shot		5-shot	
SG2Im [20]	CRN	194	210	176	186.5	196.8	224.2
MIGS [53]	CRN	158.5	156.4	157	158.4	183.5	187.6
SG2Im [20]	SPADE	66.1	42.2	70.6	48.3	95.2	73.1
MIGS [53]	SPADE	**49.5**	**26.7**	**46.1**	**24**	**53.5**	**30.7**

Table 3. Quantitative results on VG fine-tuned on 5, 10 and 160 shots.

Method	Decoder	FID ↓	KID $\cdot 10^3$ ↓	FID ↓	KID $\cdot 10^3$ ↓	FID ↓	KID $\cdot 10^3$ ↓
		160-shot		10-shot		5-shot	
SG2Im [20] (All epochs)	SPADE	55.20	35.54	81.42	59.39	91.79	68.52
MIGS [53] (1/3 epochs)	SPADE	54.83	34.21	76.56	52.02	84.87	59.38
MIGS [53] (All epochs)	SPADE	**54.24**	**29.00**	**75.96**	**50.69**	**83.54**	**55.28**

Qualitative Results We show the qualitative results of MIGS [53] in Figure 1. The qualitative analysis of the generated images revealed that our method was able to produce visually coherent and semantically meaningful scenes even when trained with a few samples in the new environment.

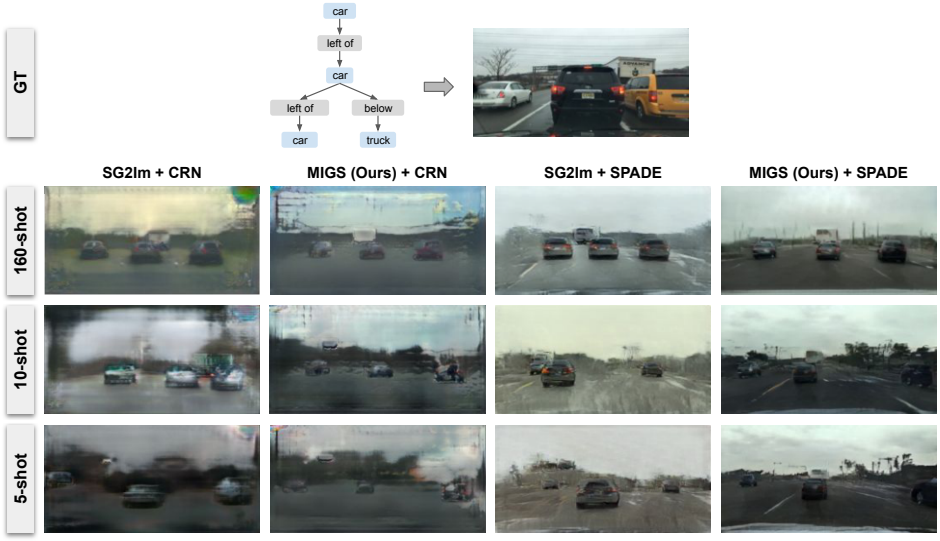

Figure 1. Qualitative results of MIGS [53] compared to SG2Im [20] on BDD100k dataset [60].

4.2. Semantic Image Manipulation

In this section, we provide the results for semantic image manipulation and compare the results of our proposed methods, SIMSG [23] and DisPositioNet [53].

Quantitative Results The quantitative evaluation of image manipulation for real world datasets is not possible, since there are no pairs of before and after manipulation. Therefore, we evaluate the performance of our methodologies for semantic image manipulation through the image reconstruction task and measuring the reconstruction error. In addition, we measure the common image generation metrics, i.e. FID, Inception Score (IS) and KID. The results of the quantitative evaluation and comparison between SIMSG [23], DisPositioNet [53] and the related work are provided in Table 4 and Table 5 on VG and COCO datasets, respectively.

Qualitative Results The qualitative results provided in Figure 2 for various image manipulation modes demonstrates that the disentangled representations within both the GAN and the graph neural network contributed to the generation of images with consistent object appearances, while maintaining accurate spatial relationships between different scene elements.

4.3. Discussion

The results of our experiments demonstrate the effectiveness of our proposed methodology in addressing the challenges associated with semantic image generation and manipulation. The integration of scene graphs, meta-learning, and disentangled representation learning enabled our approach to accurately model complex scenes and generate high-quality images that outperformed existing state-of-the-art methods.

Table 4. Image reconstruction on Visual Genome. We compare the results of our method to previous works using ground truth (GT) and predicted scene graphs. In the experiments denoted by (Generative), the whole input image is masked. N/A: Not Applicable.

Method	Decoder	All pixels					RoI only	
		MAE ↓	SSIM ↑	LPIPS ↓	FID ↓	IS ↑	MAE ↓	SSIM ↑
Generative, GT Graphs								
ISG [37]	Pix2pixHD	46.44	28.10	0.32	58.73	6.64±0.07	N/A	N/A
SIMSG [23]	SPADE	41.88	34.89	0.27	44.27	7.86±0.49	N/A	N/A
DisPositioNet [55]	SPADE	**41.62**	**35.30**	**0.26**	**40.75**	**7.93**±0.36	N/A	N/A
GT Graphs								
Cond-sg2im [20]	CRN	14.25	84.42	0.081	13.40	11.14±0.80	29.05	52.51
SIMSG [23]	SPADE	8.61	87.55	0.050	**7.54**	**12.07**±0.97	**21.62**	**58.51**
DisPositioNet [55]	SPADE	**8.41**	**87.56**	**0.048**	7.66	11.65±0.58	21.76	58.18
Predicted Graphs								
SIMSG [23]	SPADE	13.82	83.98	0.077	16.69	10.61±0.37	28.82	49.34
DisPositioNet [55]	SPADE	**9.39**	**86.91**	**0.052**	**14.42**	**10.69**±0.33	**25.40**	**51.85**

Table 5. Image reconstruction on COCO

Method	All pixels			RoI only	
	MAE ↓	SSIM ↑	LPIPS ↓	MAE ↓	SSIM ↑
Generative					
SIMSG [23]	54.03	24.12	0.490	N/A	N/A
DisPositioNet [55]	**51.07**	**26.53**	**0.418**	N/A	N/A
Non Generative					
SIMSG [23]	9.36	87.00	0.086	27.68	49.93
DisPositioNet [55]	**9.24**	**88.26**	**0.057**	**27.52**	**50.35**

The disentangling of latent and graph representations within both the GAN and the graph neural network proved to be particularly beneficial, as it facilitated better modeling of the relationships between different scene elements, leading to more coherent and visually appealing images. Furthermore, our self-supervised method for image manipulation allowed for greater flexibility in the absence of paired before-and-after data, demonstrating the potential of our approach for real-world applications.

In summary, our experiments confirm the superiority of our proposed methodology in the domain of semantic image generation and manipulation, showcasing its potential for various applications in computer vision and related fields.

5. Conclusion

In conclusion, this study has presented a novel approach to semantic image generation and manipulation by leveraging the power of scene graphs, meta-learning, and disentangled representation learning. Our methodology, which incorporates a self-supervised method for image manipulation and disentangles latent and graph representations within

Figure 2. Comparison of DisPositioNet [55] (denoted as Ours) to the previous work SIMSG [23] on (A) VG [58] and (B) COCO [57] datasets.

both the GAN and the graph neural network, has demonstrated its effectiveness by outperforming existing approaches and achieving state-of-the-art performance on a diverse range of publicly available benchmarks.

The results of our research not only contribute valuable insights and advancements to the broader field of computer vision but also open up new possibilities for future research and practical applications in semantic scene understanding. Our approach can potentially be extended to various domains, such as virtual reality, autonomous vehicles, robotics, and video game development, where accurate and detailed scene representations are essential.

As for future work, several research directions can be explored to further enhance the performance and applicability of our methodology:

- Diffusion Models: Integration of the proposed approaches into SOTA generator models such as diffusion model.
- Dynamic scenes: Extending our methodology to handle dynamic scenes, such as videos or interactive environments, would provide a more robust representation of the temporal aspects of scenes and enable the development of advanced applications in video analysis and virtual reality.
- Unsupervised and semi-supervised learning: Exploring unsupervised and semi-supervised learning techniques for scene graph feature extraction and image generation tasks could reduce the reliance on large amounts of labeled data and improve the overall efficiency of our approach.

By addressing these potential research directions, we envision that our work will continue to contribute to the ongoing advancements in the field of computer vision and semantic scene understanding, ultimately benefiting various real-world applications and opening up new avenues for exploration.

Acknowledgements

I express my deepest gratitude to my doctoral supervisor, Prof. Nassir Navab, for his invaluable guidance throughout my research journey. I am also immensely grateful to my partner and colleague, Yousef, for his unwavering assistance and collaboration, and to Helisa, who has been an exceptional collaborator and friend. I am also immensely grateful to Sabrina and Sarthak for their invaluable assistance in these projects.

Lastly, I acknowledge the generous financial support provided by the Munich Center for Machine Learning (MCML) and Deutsche Forschungsgemeinschaft (DFG), which has been instrumental in enabling the realization of this research.

References

[1] Y. Zhang, A. Khakzar, Y. Li, A. Farshad, S. T. Kim, and N. Navab, "Fine-grained neural network explanation by identifying input features with predictive information," *Advances in Neural Information Processing Systems*, vol. 34, pp. 20040–20051, 2021.

[2] A. Farshad, Y. Yeganeh, and N. Navab, "Learning to learn in medical applications: A journey through optimization," in *Meta-Learning with Medical Imaging and Health Informatics Applications*, pp. 3–25, Elsevier, 2023.

[3] A. Farshad, A. Makarevich, V. Belagiannis, and N. Navab, "Metamedseg: Volumetric meta-learning for few-shot organ segmentation," in *Domain Adaptation and Representation Transfer: 4th MICCAI Workshop, DART 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pp. 45–55, Springer, 2022.

[4] A. G. Roy, S. Siddiqui, S. Pölsterl, A. Farshad, N. Navab, and C. Wachinger, "Few-shot segmentation of 3d medical images," in *Meta-Learning with Medical Imaging and Health Informatics Applications*, pp. 161–183, Elsevier, 2023.

[5] Y. Yeganeh, A. Farshad, J. Boschmann, R. Gaus, M. Frantzen, and N. Navab, "Fedap: Adaptive personalization in federated learning for non-iid data," in *Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health: Third MICCAI Workshop, DeCaF 2022, and Second MICCAI Workshop, FAIR 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18 and 22, 2022, Proceedings*, pp. 17–27, Springer, 2022.

[6] J. Johnson, R. Krishna, M. Stark, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *CVPR*, 2015.

[7] R. Herzig, M. Raboh, G. Chechik, J. Berant, and A. Globerson, "Mapping images to scene graphs with permutation-invariant structured prediction," in *NeurIPS*, 2018.

[8] M. Qi, W. Li, Z. Yang, Y. Wang, and J. Luo, "Attentive relational networks for mapping images to scene graphs," *arXiv preprint arXiv:1811.10696*, 2018.

[9] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *CVPR*, pp. 5831–5840, 2018.

[10] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *ICCV*, pp. 1261–1270, 2017.

[11] A. Newell and J. Deng, "Pixels to graphs by associative embedding," in *NeurIPS*, pp. 2171–2180, 2017.

[12] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3716–3725, 2020.

[13] M. Suhail, A. Mittal, B. Siddiquie, C. Broaddus, J. Eledath, G. Medioni, and L. Sigal, "Energy-based learning for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13936–13945, 2021.

[14] J. Wald, H. Dhamo, N. Navab, and F. Tombari, "Learning 3d semantic scene graphs from 3d indoor reconstructions," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] S.-C. Wu, J. Wald, K. Tateno, N. Navab, and F. Tombari, "Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7515–7525, 2021.

[16] D. Xu, Y. Zhu, C. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *CVPR*, 2017.
[17] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, and X. Wang, "Factorizable net: an efficient subgraph-based framework for scene graph generation," in *ECCV*, pp. 335–351, 2018.
[18] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *ECCV*, pp. 670–685, 2018.
[19] Y. Yeganeh, A. Farshad, N. Navab, and S. Albarqouni, "Inverse distance aggregation for federated learning with non-iid data," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2*, pp. 150–159, Springer, 2020.
[20] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *CVPR*, 2018.
[21] A. Luo, Z. Zhang, J. Wu, and J. B. Tenenbaum, "End-to-end optimization of scene layout," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3754–3763, 2020.
[22] H. Dhamo, F. Manhardt, N. Navab, and F. Tombari, "Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs," in *IEEE International Conference on Computer Vision (ICCV)*, 2021.
[23] H. Dhamo, A. Farshad, I. Laina, N. Navab, G. D. Hager, F. Tombari, and C. Rupprecht, "Semantic image manipulation using scene graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5213–5222, 2020.
[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.
[25] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*, pp. 8162–8171, PMLR, 2021.
[26] A. Farshad, Y. Yeganeh, Y. Chi, C. Shen, B. Ommer, and N. Navab, "Scenegenie: Scene graph guided diffusion models for image synthesis," *arXiv preprint arXiv:2304.14573*, 2023.
[27] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
[28] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.
[29] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.
[30] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.
[31] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *arXiv preprint arXiv:2106.12423*, 2021.
[32] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *ICCV*, pp. 1511–1520, 2017.
[33] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018.
[34] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *CVPR*, 2019.
[35] B. Zhao, L. Meng, W. Yin, and L. Sigal, "Image generation from layout," in *CVPR*, 2019.
[36] W. Sun and T. Wu, "Image synthesis from reconfigurable layout and style," in *ICCV*, October 2019.
[37] O. Ashual and L. Wolf, "Specifying object attributes and relations in interactive scene generation," in *ICCV*, pp. 4561–4569, 2019.
[38] M. Ivgi, Y. Benny, A. Ben-David, J. Berant, and L. Wolf, "Scene graph to image generation with contextualized object layout refinement," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2428–2432, IEEE, 2021.
[39] L. Li, K. Fan, and C. Yuan, "Cross-modal representation learning and relation reasoning for bidirectional adaptive manipulation," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22* (L. D. Raedt, ed.), pp. 3222–3228, International Joint Conferences on Artificial

Intelligence Organization, 7 2022. Main Track.

[40] S.-M. Hu, F.-L. Zhang, M. Wang, R. R. Martin, and J. Wang, "Patchnet: A patch-based image representation for interactive library-driven image editing," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, pp. 1–12, 2013.

[41] Y. Yeganeh, A. Farshad, and N. Navab, "Shape-aware masking for inpainting in medical imaging," *arXiv preprint arXiv:2207.05787*, 2022.

[42] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016.

[43] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *CVPR*, pp. 5485–5493, 2017.

[44] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[45] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[46] S. Hong, X. Yan, T. E. Huang, and H. Lee, "Learning hierarchical semantic image manipulation through structured representations," in *Advances in Neural Information Processing Systems*, pp. 2713–2723, 2018.

[47] E. Ntavelis, A. Romero, I. Kastanis, L. Van Gool, and R. Timofte, "SESAME: Semantic Editing of Scenes by Adding, Manipulating or Erasing Objects," in *Computer Vision – ECCV 2020* (A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds.), (Cham), pp. 394–411, Springer International Publishing, 2020.

[48] H. Ling, K. Kreis, D. Li, S. W. Kim, A. Torralba, and S. Fidler, "Editgan: High-precision semantic image editing," *arXiv preprint arXiv:2111.03186*, 2021.

[49] A. Farshad, Y. Yeganeh, P. Gehlbach, and N. Navab, "Y-net: A spatiospectral dual-encoder network for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II*, pp. 582–592, Springer, 2022.

[50] Y. Yeganeh, A. Farshad, G. Guevercin, A. Abu-zer, R. Xiao, Y. Tang, E. Adeli, and N. Navab, "Scope: Structural continuity preservation for medical image segmentation," *arXiv preprint arXiv:2304.14572*, 2023.

[51] S. Su, L. Gao, J. Zhu, J. Shao, and J. Song, "Fully functional image manipulation using scene graphs in a bounding-box free way," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1784–1792, 2021.

[52] S. Garg, H. Dhamo, A. Farshad, S. Musatian, N. Navab, and F. Tombari, "Unconditional scene graph generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16362–16371, 2021.

[53] A. Farshad, S. Musatian, H. Dhamo, and N. Navab, "Migs: Meta image generation from scene graphs," in *BMVC*, 2021.

[54] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[55] A. Farshad, Y. Yeganeh, H. Dhamo, F. Tombari, and N. Navab, "Dispositionet: Disentangled pose and identity in semantic image manipulation," in *BMVC*, 2022.

[56] J. Ma, P. Cui, K. Kuang, X. Wang, and W. Zhu, "Disentangled graph convolutional networks," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 4212–4221, PMLR, 09–15 Jun 2019.

[57] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[58] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, no. 1, pp. 32–73, 2017.

[59] J. You, R. Ying, X. Ren, W. L. Hamilton, and J. Leskovec, "Graphrnn: Generating realistic graphs with deep auto-regressive models," *ICML*, 2018.

[60] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.