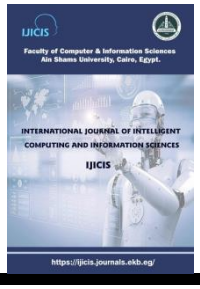
IJICIS, Vol.24, No.3, 1-10
DOI:10.21608/ijicis.2024.301597.1342

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



TRANSFORMER-BASED BACKBONES FOR SCENE GRAPH GENERATION: A COMPARATIVE ANALYSIS

Mohammad Essam*

Scientific Computing,
Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
Mohamed97@cis.asu.edu.eg

Dina Khattab

Scientific Computing,
Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
dina.khattab@cis.asu.edu.eg

Howida A. Shedeed

Scientific Computing,
Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
dr_howida@cis.asu.edu.eg

Mohamed F. Tolba

Scientific Computing,
Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
fahmytolba@cis.asu.edu.eg

Received 2024-07-04; Revised 2024-08-01; Accepted 2024-08-08

Abstract: *The Scene Graph is a modern structured representation of an image scene that explicitly describes the scene as a set of objects, attributes, and links between the objects (relationships). With the great advancements in the computer vision field, researchers dedicated their efforts towards more complex reasoning and a high level of understanding of visual scenes. Tasks like Visual Question Answering, image generation, and cross-modal retrieval are examples of Complex vision tasks that require a high level of visual scene understanding. Scene Graph is an effective data structure that highlights complex visual relationships presented in a scene. In this work, we provide a comparative analysis of Scene Graph Generation (SGG) backbone models. The contributed work aims to compare the Convolution Neural Networks (CNN) backbones and the vision transformer-based backbones using the RelTR model. The conducted analysis proved that both SwiftFormer L3 and MiT-B2 transformer backbones increased the model performance over the ResNet50 CNN backbone by 2.1 % and 2.5% Recall@50 respectively when experimented on the same Visual Genome 50 test split. The Visual Genome 50 is a tailored version of The Visual Genome dataset. It contains only the 50 most common relationships and the most frequent 150 object classes.*

Keywords: *Scene Graph, Scene Graph Generation, Transformer-Based Backbone, Visual Relationship Detection, Low Resolution*

1. Introduction

***Corresponding Author**: Mohammad Essam

Scientific Computing Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

Email address: Mohamed97@cis.asu.edu.eg

In recent years, computer vision has evolved as a cutting-edge solution to incorporate visual scene analysis features into machines. Starting from classical computer vision systems, they contained a sequence of processes to solve simple visual tasks like face recognition [1] and image classification [2]. These processes involve image acquisition, data preprocessing, hand-crafted feature extraction, and pattern recognition. Most of the contributions in these classical systems were focused on hand-crafted feature extraction techniques and how to overcome the hardware limitations at that time.

With the rapid advancement of hardware capabilities, computer vision systems gained extensive attention after introducing the modern CNN AlexNet [3]. This innovation paved the way for working on more complex CNN architectures to solve challenging visual tasks like SGG [4], cross-modal retrieval [5], and small object detection [6]. Building on the foundation of the CNN, Faster R-CNN [7] was proposed to solve object detection task with a new concept called backbone. The backbone of Faster R-CNN is a pre-trained CNN which acts as an automatic feature extractor and boosts model performance. Many advanced CNN architectures adopted the same concept later in object detection [8] as well as other tasks such as semantic segmentation [9] and pose estimation [10].

Accompanied by the transformer research wave in natural language processing [11], the computer vision community started to adapt the transformer techniques to several fundamental computer vision problems like image recognition [12], and instance Segmentation [13]. With the progressive contributions of vision transformers [14], [15], most of the modern vision transformers achieved better results than the complex CNNs in almost every computer vision task [16]. These results suggest that replacing CNNs with vision transformers can boost the performance of any module.

Consequently, some research directions tended to design novel data representations [17] that replace the pixel values with better visual information for the computer vision models. One important image data representation is the Scene Graph. The Scene Graph concept was formulated by Johnson et al. [18] to crack the image retrieval task. Scene Graph encodes the whole visual scene as a group of triplets (Subject - Relationship - Object), where each item in the scene can be whether subject or object and connected to other items by relationships. The scene items are represented by graph nodes and the relationships serve as the graph edges that connect these nodes.

As the initial Scene Graph developed was manually labelled [18], a great effort has been dedicated to automating the process of SGG [19], [20], [21], to facilitate the utilization of Scene Graphs in downstream vision tasks. One of the promising SGG architectures contributed was the RelTR [22] model. It adapted the transformer architecture in the task of SGG to boost the performance and achieve State-Of-The-Art (SOTA) results.

The contribution of this paper is to provide a comparative analysis between transformer-based backbones and CNN backbones for the SGG task using the RelTR model. Our analysis shows that both SwiftFormer L3 [23] and MiT-B2 [14] transformer backbones boosted the model performance over the ResNet50 [24] CNN backbone by 2.1 % and 2.5% Recall@50 respectively when tested on the same Visual Genome 50 dataset [25].

The taxonomy of this research is as follows: Section 2 provides a comprehensive survey about the related research done in the SGG task. Section 3 discusses the contributed system architecture along with its

hardware configuration. The experimental results and their outcomes are demonstrated in Section 4. Section 5 outlines the conclusion and the potential future directions for this research.

2. Related Work

The Scene Graph serves as a global scene encoder, where the scene objects along with their relationships represent the nodes and the edges between these nodes respectively. As noticed in Figure. 1, the SGG model takes a colored image and produces a graph where Scene Objects (e.g., Giraffe, Head, Grass) are connected through the relationships (on, has). The following sub-sections will delve into the novel contributions of SGG and Transformer-based backbones concepts [26].

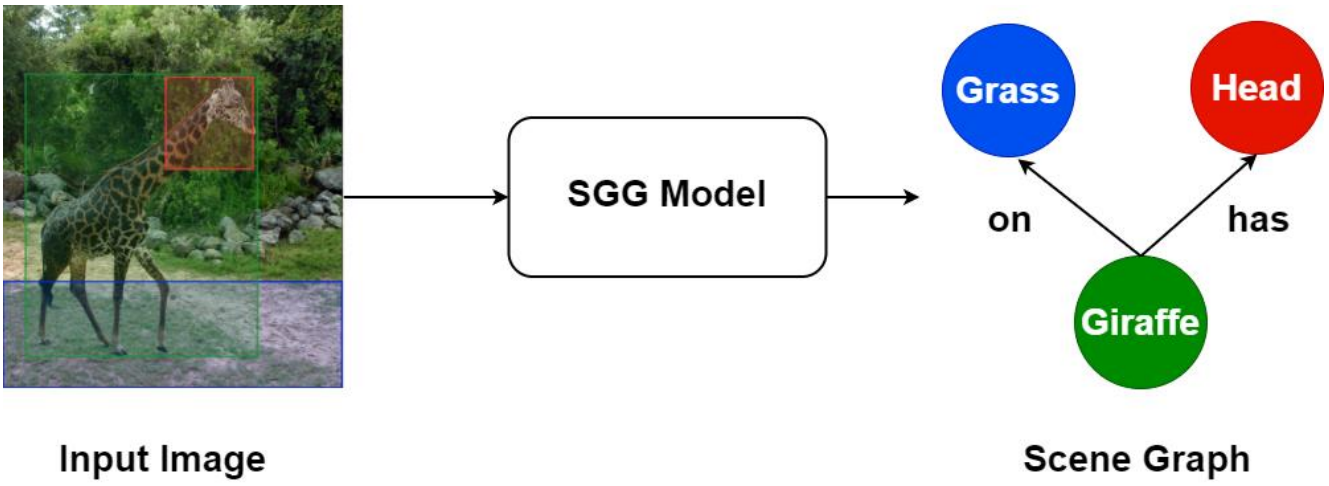


Figure. 1: An Illustration of Scene Graph Generation Task

2.1. Scene Graph Generation

Building on the success of the DETR [27] transformer in object detection task, Cong, et al. [22] introduced the RelTR SGG transformer. The RelTR follows the same building blocks of DETR as the input image is given to a CNN backbone, this backbone is followed by a vision transformer encoder-triple decoder architecture which is the main modification from the DETR architecture. This modification enabled the model to extract triplet features for (subject-relationship-object) and so enhance the performance for the SGG task.

Focusing on the process of human perception, Zhang, et al. [28] modified the Scene Graph to add an importance factor for each relationship (edge) in the graph. This factor upgraded the default message-passing modules to a saliency-based message-passing one, enabling the model to focus more on the most crucial relationships to increase the SGG accuracy. Zheng, et al. [29] proposed an SGG prototype-based network to solve the problem of intra-class variation in relationship classes. The contributed network modelled the relationships with an aligned prototype to generate robust embedding space that is vital for relationship detection. Kundu, et al. [30] invented a generative approach to tackle the SGG problem. They utilized a generative two-stage transformer, where the first stage samples an initial Scene Graph from the detected objects, and then the second stage filters the generated graph with relationship refinement as the top priority.

2.2. Transformer-Based Backbones

Wu, et al. [31] employed the concept of transformer backbone in the biological cell detection task. They introduced a modified Yolov5s model, incorporating a Swin V2 [32] transformer as a feature extraction backbone. This backbone boosted cell detection accuracy by acquiring vital global visual information.
Han, et al. [33] designed a deep-learning network to enhance wireless communication quality. The proposed network is built upon a basic transformer backbone that provides feedback on the communication state information. This technique exhibits an advanced performance over the classic methods. Reza, et al. [34] extended the concept of the transformer-based backbone in the segmentation of temporal actions in videos. They introduced a dual attention mechanism that boosted the performance of the transformer backbone in capturing better hierarchical details. They also included cross-connections between both transformer encoder and decoder to outperform the original transformer performance. Motivated by the results of transformer backbones in 2D computer vision tasks, Yang, et al. [35] utilized the transformer backbone concept in the indoor 3D scene understanding. They pretrained a Swin3D model on a large-scale 3D synthetic dataset to generalize on various segmentation and detection downstream tasks. The Swin3D model showed a superior performance when validated for scalability and generality.

3. Methodology

The following sub-sections introduce a specified description of the main components of our research. Starting from the hardware configuration needed to train the models, we delve into the description of the dataset used for training the contributed models. Finally, we conclude the section with an emphasis on the system architecture highlighting each component and its role in the system.

3.1. Hardware Specifications

The contributed analysis is conducted on one Google Colab GPU instance. This instance contains an Intel Xeon 2.2 GHz CPU, 53 GB of system RAM, and an Nvidia L4 GPU with 24 GB of VRAM. The dedicated GPU is optimized specifically for accelerated AI model training, which enables finishing the training of all experiments in around 200 GPU hours only.

3.2. Dataset

The Visual Genome [25] dataset is a pioneering dataset for the SGG task. With more than 108,000 images, Visual Genome is the default choice to test the generalization of any SGG model as each image has an average of 25 relationships taken from real-life scenarios. In this research, we utilize a customized version of this dataset named Visual Genome 50 [4], which contains only the 50 most appearing relationships besides the top-occurring 150 object classes. The key idea of working on Visual Genome 50 is to prevent the long tail distribution problem [36] in the original Visual Genome. Figure. 2 presents a sample image from Visual Genome 50 accompanied by part of its labels. The dataset is divided into training, validation, and test sets with the following proportions 70%, 5k examples and 30% respectively. The training strategy utilized 70% of the dataset, while the validation set was randomly selected from the training data in each epoch, and the final 30% was kept for testing scenarios.


```
{
"image_id": 1,
"url": "https://cs.stanford.edu/people/rak248/VG_100K_2/1.jpg",
"width": 800,
"height": 600,
"relationships": [
  {
    "relationship_id": 15927,
    "predicate": "ON",
    "subject": {
      "object_id": 5045,
      "x": 119,
      "y": 338,
      "w": 274,
      "h": 192,
      "names": [
        "shade"
      ]
    },
    "object": {
      "object_id": 5046,
      "x": 77,
      "y": 328,
      "w": 714,
      "h": 262,
      "names": [
        "street"
      ],
    }
  }
  ...
}
```


a) b)

Figure. 2: A sample from the Visual Genome 50 dataset, where a) represents the image and b) is the partial graph label of the image

3.3. Model Architecture

The proposed system architecture in this research is mainly inspired by the RelTR SGG model. As shown in Figure. 3a, the original system architecture of the RelTR model is built upon using ResNet50 CNN [37] as the main feature extractor for the whole system. The extracted features are then passed through the RelTR encoder to produce a continuous representation from the input. The encoded representation is finally passed to the RelTR decoder to generate the scene graph for the given input image. As Figure. 3b highlights, that the key change in the proposed model is the utilization of a vision transformer backbone instead of the usual CNN backbone. The use of MiT-B2 [14] and SwiftFormer L3 [23] transformer backbones within this research are based on selecting those having nearly the same number of parameters as the ResNet50 CNN, (see Table 1), to allow for a fair comparison.

Table 1 contributes a comparison between the three backbone models utilized in this research. The table focuses on the differences in the complexity of each model, measured by the number of parameters, reflecting the varying techniques and efficiencies in their designs.

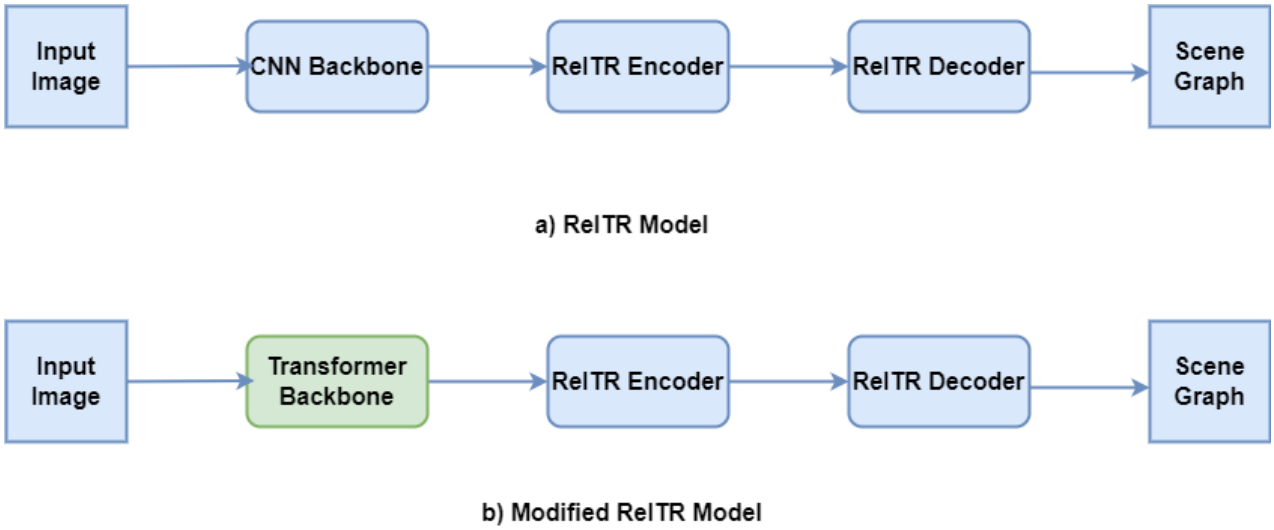


Figure. 3: A Diagram of a) The original RelTR Model, b) The modified RelTR Model

Table 1: A Comparison between ResNet50, SwiftFormer L3, and MiT-B2 Models

Backbone Model	Backbone Type	Number of Parameters (Millions)
ResNet50 [37]	CNN	25.6
SwiftFormer L3 [23]	Transformer	28.5
MiT-B2 [14]	Transformer	24.2

3.3.1. SwiftFormer-L3 Transformer

The SwiftFormer-L3 transformer is the largest variant of the SwiftFormer transformer series that is designed specifically to boost the mobile inference speed. The SwiftFormer-L3 proposed an efficient design, unlike the traditional transformer architectures, by replacing the heavyweight self-attention [11] operation with a more lightweight additive attention operation. This operation enabled a more real-time architecture as well as high accuracy in many downstream computer vision tasks.

3.3.2. MiT-B2 Transformer

The MiT-B2 transformer is one of the SegFormer transformers family that vary from SegFormer-B0 to B5. The SegFormer architecture consists of a multi-scale encoder (the MiT transformer) as well as a simple MLP decoder that combines the features from the different scales. Unlike most vision transformer [38], the MiT-B2 transformer does not use the positional encoding technique, which means there is no interpolation process when using input data of different resolutions. The MiT-B2 is built hierarchically so that the encoder model can extract both local and global features from a given input. This feature boosted the performance of the architecture, especially in segmentation tasks.

3.3.3. RelTR Model

The RelTR SGG model is a lightweight transformer architecture that emerged as a modified version of the DETR transformer to tackle the SGG task. The RelTR transformer consists of a transformer encoder and a triplet transformer decoder, The main role of the encoder is to project the input sequence into a high-level continuous space that keeps all the information from the input sequence. The triplet decoder gets the encoded high-level sequence and the previous output to generate the required scene graph. The main cause of utilizing a triplet decoder is due to the nature of the SGG task which requires predicting the scene graph in the form of triplets (Subject-Relationship-Object).

4. Experiments and Results

In this section, the performed experiments are explained, including the formulation of the evaluation metrics used. The details of training parameters are then highlighted. The performance comparison between the CNN backbone and transformer backbones is provided as the final part of the section.

4.1. Evaluation Metrics

The Scene Graph Detection (SGDet) metric is frequently used in the evaluation of the SGG models, mainly when focusing on measuring how well the model can predict both the objects and the relationships in the scene. The SGDet is built upon the Recall@K formula, which calculates both the portion of correct relationship predictions as well as the involved object detection. The Recall@K in the SGG task is computed with different K values like 20, 50, and 100. We utilize the K parameter as 50 in the performed experiments. To compute Recall@K, the SGG model predicts a set of (Subject, Relationship, Object) Triplets, the top K triplets are then compared to the ground truth graph of the input image. After that, a confusion matrix is calculated to get the False Negative (FN) and the True Positive (TP) rates which are required parameters to estimate the Recall using Eq. (1).

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

4.2. Training Setup

In all our trained models, we follow the same training setup as [22]. We keep the same data splits of the Visual Genome 50 and the fixed image resolution of 384 × 384 for all the performed experiments. Two learning rates of 10^{-4} and 10^{-5} are utilized for the RelTR transformer and the transformer backbones (MiT-B2, SwiftFormer L3) respectively. All training models concluded after 50 epochs as in [23] to provide a fair comparison.

4.3. Experimental Results

In the following experiment, we replaced the CNN-based feature extraction part of the original RelTR model with a transformer backbone utilizing SwiftFormer L3 and MiT-B2 respectively. We compare these results with the original RelTR model on the same resolution 384x384 using Visual Genome 50 test split. As noticed in Table 2, both the proposed modifications outperform the original architecture by 2.1 % and 2.5 % Recall@50 respectively.

Table 2: Experimental Results on the Visual Genome 50 Test Split

Model	Backbone Model	Recall@50
Original RelTR	ResNet50	18.3
Modified RelTR	SwiftFormer L3	20.4
Modified RelTR	MiT-B2	20.8

These results indicate that the vision transformers are better feature extractors than the traditional CNN under the same train and test constraints. Unlike CNNs, which focus on extracting visual features locally, vision transformers employ attention mechanisms to capture the global context within the image. This leads to a better understanding of complex scenes and thus boosts the performance over the CNNs. Comparing the trained models with the benchmark SwinRelTR [39], that achieves higher recall@50 but is considered a more complex model, it highlights that lightweight vision transformer backbones like MiT-B2 and SwiftFormer L3 enhance the recall with reduced computational complexity compared to SwinRelTR.

5. Conclusion

In this research, we discussed the idea of using vision transformers as backbones for the SGG models instead of the classic CNN backbone. We applied two experiments with different transformer architectures (SwiftFormer L3 and MiT-B2) as backbones for the RelTR SGG transformer. Our experiments outperformed the original RelTR by 2.1 % and 2.5 % Recall@50 respectively, thus suggesting that SGG models can benefit from the utilization of vision transformer backbones as the main feature extractors. Regardless of the observed improvements, this study has specific limitations. The analysis is limited to the Visual Genome 50 dataset, which may not generalize well to other datasets with different distributions of object classes and relationships. Furthermore, the comparison is limited to a limited selected transformer backbone models. In future work, more vision transformers can be tested as feature extractors like SwinV2 [40] and FasterViT [41]. Also, more architectural modifications in the RelTR model can be applied to add more complicated graphical modules for the sake of better scene graphs generated. Finally, expanding the contributed comparative analysis to a wider range of datasets to validate the generalizability of the results.

References

[1] G. Srivastava and S. Bag, “Modern-day marketing concepts based on face recognition and neuro-marketing: a review and future research directions,” Benchmarking: An International Journal, vol. 31, no. 2, pp. 410–438, 2024.

[2] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Adv Neural Inf Process Syst, vol. 25, 2012.

[4] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, “Neural Motifs: Scene Graph Parsing with Global Context,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5831–5840.

[5] S. Wang, R. Wang, Z. Yao, S. Shan, and X. 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, pp. 1508–1517.

[6] X. Wu, D. Hong, and J. Chanussot, "UIU-Net: U-Net in U-Net for infrared small object detection," IEEE Transactions on Image Processing, vol. 32, pp. 364–376, 2022.
[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in Advances in neural information processing systems 28, 2015. [Online]. Available: https://github.com/
[8] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.10934
[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
[10] X. Chu, W. Ouyang, X. Wang, and others, "Crf-cnn: Modeling structured information in human pose estimation," Adv Neural Inf Process Syst, vol. 29, 2016.
[11] A. Vaswani et al., "Attention is all you need," Adv Neural Inf Process Syst, vol. 30, 2017.
[12] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning", Accessed: May 30, 2022. [Online]. Available: www.aaai.org
[13] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," IEEE Trans Pattern Anal Mach Intell, vol. 43, no. 5, pp. 1483–1498, 2019.
[14] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in Advances in Neural Information Processing Systems, 2021, pp. 12077–12090.
[15] M. Caron et al., "Emerging Properties in Self-Supervised Vision Transformers." [Online]. Available: https://github.com/facebookresearch/dino
[16] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[17] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for Scene Graph Generation," in Proceedings of the European conference on computer vision, 2018, pp. 670–685.
[18] J. Johnson et al., "Image retrieval using scene graphs," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3668–3678.
[19] J. Chiou, H. Ding, H. Yan, C. Wang, R. Zimmermann, and J. Feng, "Recovering the Unbiased Scene Graphs from the Biased Ones," in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 1581–1590. [Online]. Available: https://github.com/coldmanck/recovering-unbiased-scene-graphs
[20] H. Liu, N. Yan, M. Mortazavi, and B. Bhanu, "Fully convolutional scene graph generation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11546–11556.
[21] X. Lin, C. Ding, J. Zeng, and D. Tao, "GPS-Net: Graph Property Sensing Network for Scene Graph Generation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3746–3753.
[22] Y. Cong, M. Y. Yang, and B. Rosenhahn, "Reltr: Relation transformer for scene graph generation," IEEE Trans Pattern Anal Mach Intell, 2023.
[23] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "SwiftFormer: Efficient additive attention for transformer-based real-time mobile vision applications," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 17425–17436.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-December, pp. 770–778, Dec. 2016, doi: 10.1109/CVPR.2016.90.

[25] R. Krishna et al., "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," Int J Comput Vis, vol. 123, no. 1, pp. 32–73, Jun. 2017, doi: 10.1007/s11263-016-0981-7.

[26] M. Goldblum et al., "Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks," Adv Neural Inf Process Syst, vol. 36, 2024.

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in European conference on computer vision, May 2020, pp. 213–229. [Online]. Available: http://arxiv.org/abs/2005.12872

[28] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C.-W. Chen, "Boosting scene graph generation with visual relation saliency," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 19, no. 1, pp. 1–17, 2023.

[29] C. Zheng, X. Lyu, L. Gao, B. Dai, and J. Song, "Prototype-based embedding network for scene graph generation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22783–22792.

[30] S. Kundu and S. N. Aakur, "Is-ggt: Iterative scene graph generation with generative transformers," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6292–6301.

[31] P. Wu et al., "An improved Yolov5s based on transformer backbone network for detection and classification of bronchoalveolar lavage cells," Comput Struct Biotechnol J, vol. 21, pp. 2985–3001, 2023.

[32] Z. Liu et al., "Swin transformer v2: Scaling up capacity and resolution," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 12009–12019.

[33] X. Han et al., "AI enlightens wireless communication: A transformer backbone for CSI feedback," China Communications, 2024.

[34] S. Reza, B. Sundareshan, M. Moghaddam, and O. Camps, "Enhancing transformer backbone for egocentric video action segmentation," arXiv preprint arXiv:2305.11365, 2023.

[35] Y.-Q. Yang et al., "Swin3d: A pretrained transformer backbone for 3d indoor scene understanding," arXiv preprint arXiv:2304.06906, 2023.

[36] 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, 2020, pp. 3716–3725. [Online]. Available: https://github.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-December, pp. 770–778, Jun. 2016, doi: 10.1109/CVPR.2016.90.

[38] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[39] M. Essam, H. A. Shedeed, M. F. Tolba, and D. Khattab, "SwinRelTR: an efficient single-stage scene graph generation model for low-resolution images," Int. J. of Intelligent Engineering Informatics, vol. 12, no. 2, pp. 169–187, 2024.

[40] Z. Liu et al., "Swin transformer v2: Scaling up capacity and resolution," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 12009–12019.

[41] A. Hatamizadeh et al., "Fastervit: Fast vision transformers with hierarchical attention," arXiv preprint arXiv:2306.06189, 2023.