Vision Transformer based Spatially Conditioned Graphs for Long Tail Visual Relationship Recognition CVPR 2023 LTVRR Challenge

Chenyu Wang[1*] Shuo Wang[2*] Shenghua Gao[1†]
[1]ShanghaiTech University [2]Shanghai Jiao Tong University
{wangchy8,gaoshh}@shanghaitech.edu.cn
gavin.wang@sjtu.edu.cn

Abstract

The long-tail visual relationship recognition (LTVRR) task aims at understanding the pairwise visual relationships that follow a long-tail distribution between interacting objects in an image. In this paper, we proposed to learn more effective feature representation for the LTVRR task by applying the powerful vision transformer backbone, relationship recognition, and refinement modules. The vision transformer backbone can improve the feature representation of the subject and object node encoding. We also adopt spatially conditioned graphs to implicitly learn the relation node encoding of the given subject and object pair. Lastly, we use RelTransformer to refine the subject, object, and pairwise relationship feature representation. With our approach, we achieved 23.4% on the overall relation accuracy on the VG8K-LT test set, which improved by 4.4%, and achieved 25.1% on the overall accuracy on the GQA-LT test set, compared to the winner of the 2021 LTVRR Challenge. Our code is available at https://github.com/GWwangshuo/VTSCG_LTVRR.git.

1. Introduction

The task of visual relationship recognition (VRR) requires understanding relationships between pairwise interacting objects in a visual scene. VRR benefits various other vision tasks such as scene graph generation (e.g., [14]), human object interaction (e.g., [18, 28]), and image captioning (e.g., [27]), due to the enriched scene understanding. Previous works of VRR usually assume that the training data are abundant, in which each class typically has a few hundred to thousands of examples. However, visual relationships usually follow a long-tail distribution due to their compositional nature. Long-tail Visual relationship recognition (LTVRR) is a more realistic task and has become more challenging when the vocabulary becomes large.

There are some works for the LTVRR task. Graph-based methods [7, 17] tackle with the LTVRR tasks under a graph scenario, and they use graph attention networks to iteratively pass messages from direct or indirect nodes to the relation. However, this design implicitly constrains the relation to the focus on its surrounding nodes rather than the distant nodes. Recent methods [1, 6] of the LTVRR tasks usually consist of three steps: 1) extracting embeddings from the visual features of cropped image regions of subject, object, and pairwise relationship; 2) augmenting the long-tail relation representation learning; 3) predicting the categories of subject, object, and corresponding pairwise relationship. Regarding feature extraction, most previous methods rely on VGG [21] or ResNet [10] as backbone networks. Due to the limited feature representation of these backbone networks, previous methods can only extract coarse rather than fine visual embeddings, which are important for subject, object, and pairwise relationship recognition.

To alleviate the aforementioned problems, we apply the vision transformer to improve the feature extraction. We use spatially conditioned graphs to learn the pairwise relationships between given subjects and objects, adopting reltransformer to refine the feature embeddings of subjects, objects, and pairwise relationships. The Vision Transformer (ViT) [19] has recently emerged as a powerful alternative and yields state-of-the-art performance in various computer vision tasks. Spatially conditioned graphs are a graph neural network. In this network, the relationships are represented as edge encoding and are obtained by using spatial subject-object and global context information. By message passing, feature representations of the subject, object, and pairwise relationship can be updated. After obtaining the subjects, objects, and their pairwise relationships feature embeddings, these feature embeddings are sent to a relation-

[*] Equal contribution.
[†] Corresponding author.

triplet refinement module to refine them further. The refined feature embeddings are utilized to predict the subject, object, and pairwise relationship label. Extensive experiments demonstrate the effectiveness of the adopted method on two recently large-scale long-tail VRR benchmarks, GQA-LT and VG8K-LT. We also conducted several ablative experiments and showed the usefulness of the vision transformer backbone, spatially conditioned graphs, and reltransformer.

2. Related Work

2.1. Long-Tail Visual Relationship Recognition

Most real-world data distributions are imbalanced and skewed to the few head classes, which typically obtain a low tail accuracy due to the extreme imbalance when doing recognition tasks. Aiming to reduce the accuracy gap between the head and tail classes, previous works revolve around the core idea of alleviating the imbalance of the data, proposing re-sampling [3, 8, 20] and re-weighting [5, 12]. However, those biasing towards tail classes methods come at the cost of sacrificing the performance of the head classes [13], along with difficulty in model convergence. Another solution is to use a two-stage network architecture [13, 30], which decouples the representation and the classifier learning. Drawing on the methods of two network branches proposed in the [30], several works [4, 25] introduce the multi-expert framework to improve the performance. As for information augmentation, [7] adjusts the original Mixup [29] to make it more suitable for long-tailed distributions.

2.2. Human-Object Interaction Detection

Human-object interaction (HOI) detection aims to localize and classify relationships between humans and objects. Existing HOI methods can be divided into two categories: two-stage methods and one-stage methods. One-stage HOI detection framework is first introduced in PPDM [16], where interactions are directly detected as keypoints. Recent one-stage methods [22] formulate the HOI detection task as a set prediction problem and utilize the transformer to learn a number of queries as the feature representations of humans, objects, and their corresponding interactions. The two-stage HOI detection methods first use an off-the-shelf detector to detect humans and objects and then classify the interaction label for pairwise human-object. Several studies [18, 28] attempted to encode contextual information using a message-passing mechanism in a graph structure. Recent works [18, 26] proposed to adopt vision transformer and text-to-image diffusion models to HOI detection tasks.

3. Method

3.1. Problem Definition

In the VRR task, the goal is to predict r between the given n_s and n_o, where n_s, n_o, r denotes a subject, an object, and their relationship, respectively. Previous VRR methods [6] have the below setting.

$$y^r = f(b^s, b^o, b^r, I) \tag{1}$$

where b^s, b^o, and b^r are the subject, object and relationship bounding boxes. b^r is obtained by the minimum enclosing region of b^s and b^o. y^r is the relationship label. I denotes the raw RGB image features. f is the inference model.

An overview of our framework is shown in Fig. 1. We first use ViT to extract the feature representation of subjects and objects. Then, a spatially conditioned graphs network is utilized to classify the pairwise relationships. The subject node, object node, and relation encoding are next sent to reltransformer module to obtain refined subject, object, and relationship features. Finally, classifiers are utilized to predict the refined subject, object, and relationship features to the target label of the subject, object, and relationship label.

3.2. Feature Extraction with ViT

Instead of using the commonly used backbone network VGG, ResNet, for feature extraction, we attempted ViT due to its powerful feature representation. In the case of the VGG and ResNet backbone, visual features are extracted from the feature map of the backbone networks via ROI-Pooling or ROI-Align [9]. However, ROI-Pooling and ROI-Align are not appropriate for the feature extraction of the ViT backbone due to the different shapes of the output feature map compared to the VGG and ResNet.

Masking with Overlapped Area. To improve the feature extraction, we adopted the recently proposed masking with the overlapped area (MOA) module [18], which has been demonstrated to be more effective for feature extraction of ViTs compared with the RoI-Pool and RoI-Align in HOI detection task. The MOA addresses the quantization problem by utilizing the overlapped area between each patch and the given region to the attention mask in the attention function.

3.3. Relation Learning

Spatially Conditioned Graphs. After extracting the subject and object visual features, we need to recognize the corresponding pairwise relationship. To achieve this, we adopted a graph neural network named spatially conditioned graphs (SCG). The SCG network is first introduced in [28] and is designed to jointly reason about the appearance and spatial information of an image.

Negative Samples Reduction. We found many subject-object pairs in both GQA-LT and VG8K-LT benchmarks, directly constructing these pairwise nodes to graph structure using SCG may bring lots of negative subject-object pairs with no relationships. To deal with this, we remove duplicate subject and object bounding boxes and limit the maximum input number of subjects and objects of SCG to 15

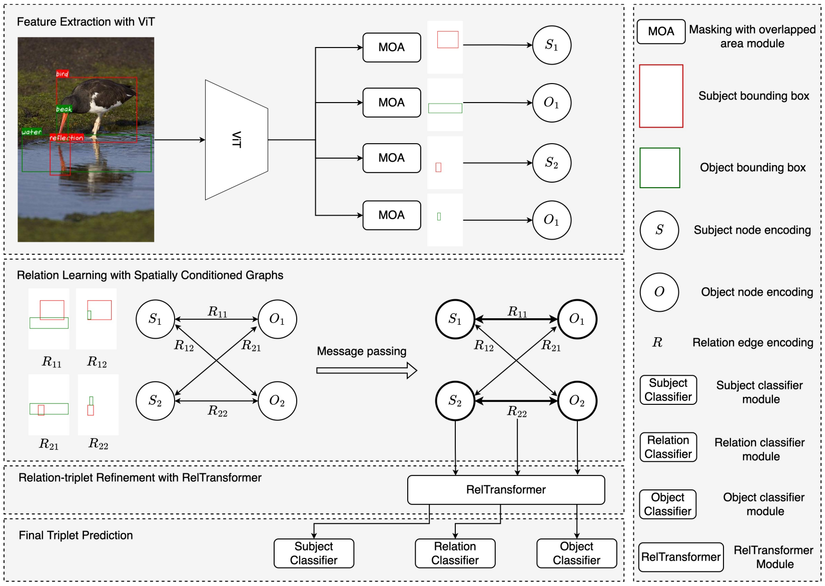
Feature Extraction with ViT
bird
beak
water
reflection
ViT
MOA
MOA
MOA
MOA
S_1
O_1
S_2
O_1
Relation Learning with Spatially Conditioned Graphs
R_{11}
R_{12}
R_{21}
R_{22}
S_1
O_1
S_2
O_2
Message passing
Relation-triplet Refinement with RelTransformer
RelTransformer
Final Triplet Prediction
Subject Classifier
Relation Classifier
Object Classifier
MOA Masking with overlapped area module
Subject bounding box
Object bounding box
S Subject node encoding
O Object node encoding
R Relation edge encoding
Subject Classifier Subject classifier module
Relation Classifier Relation classifier module
Object Classifier Object classifier module
RelTransformer RelTransformer Module

Figure 1. The overall pipeline of our method. We extract features for subjects and objects using a ViT backbone and MOA module. Then, a graph neural network is utilized to learn the relation representation of subject-object pairs. The subject node, object node, and relation edge encoding are updated through the message-passing process. After the message passing, the combination of subject node, object node, and relation edge encoding is sent to reltransformer for feature refinement. Classifiers are utilized to classify the refined subject, object, and relationship embeddings to the subject, object, and relationship label.

for an image. Moreover, we calculate the minimum interaction over the union between the subject-object pairs with relationships and all pairwise subject-object, keeping the minimum 50 subject-object pairs that are not in the subject-object pairs with relationships as the negative pairs. By using SCG and negative sample reduction, GPU memory can be effectively saved, and pairwise relationships feature representations can be well and fast learned.

3.4. Relation-triplet Refinement

RelTransformer. Chen et al. [6] proposed an effective and convenient module, reltransformer, to improve the relation-triplet feature representation. It consists of two variants of Transformer [24] encoders, a global-context encoder, and a relational encoder, linked by meshed attention. The former learns a context representation of the scene, while the latter focus on the relation representation guided by the corresponding subject and object features. To alleviate the model biasing towards the higher-frequent relations caused by the imbalance distribution, a novel memory augmentation method is introduced to allow the information to be shared across the whole dataset. We adopt the reltransformer architecture in this work to refine further the feature representation of subject, object, and pairwise relationships. Finally, the features are finally sent to the subject, object, and relation classifiers, a one-layer MLP, to predict the target subject, object, and relationship label, respectively. More details can be found in [6].

4. Experiments

4.1. Datasets

We show experiments on two large-scale long-tailed benchmarks, i.e., **GQA-LT** [11] and **VG8K-LT** [15], and they are built upon GQA [11] and Visual Genome(v1.4) [15] respectively.

GQA-LT. The dataset contains 1703 object and 310 relationship categories, with 72,580 training, 2,573validation and 7,722 testing images in total. It has a heavy long-tail distribution with examples ranging from 1 to 1,692,068.

VG8K-LT. This dataset comprises 97,623 training, 1,999 validation, and 4,860 testing images. It encompasses a total of 5,330 objects and 2,000 types of relationships, and the example numbers per object class range from 14 to 196,944, and that per relation class range from 18 to 618,687.

Both datasets split into three parts, **Many, Medium, Few**, and the selection ratio of each split follows is based on the frequency of each class: **Many** (top 5%), **Medium** (middle 15%), **Few** (remaining 80%).

4.2. Experimental Setup

Baselines. We compare our model with several state-of-the-art models, including LSVRU [1] and RelTransformer [6]. Besides that, several strategies are always applied to tackle with the imbalanced distribution: 1) re-balance loss functions: weighted cross entropy (WCE), equalization loss (EQL) [23], and ViLHub loss [1]. 2) relation augmentation

Architecture	Learning Methods	VG8K-LT many 100	 medium 300	 few 1,600	 all 2,000	GQA-LT many 16	 medium 46	 few 248	 all 310
LSVRU	VilHub [2]	27.5	17.4	14.6	15.7	63.6	17.6	7.2	11.7
LSVRU	VilHub + RelMix [2]	24.5	16.5	14.4	15.4	63.4	14.9	8.0	11.9
LSVRU	EQL [23]	22.6	15.6	12.6	13.6	62.3	15.8	6.6	10.8
LSVRU	WCE	35.5	24.7	15.2	17.2	53.4	35.1	15.7	20.5
RelTransformer	WCE	36.6	27.4	16.3	19.0	**63.6**	**59.1**	**43.1**	**46.5**
Ours	WCE	**37.1**	**30.6**	**21.1**	**23.4**	57.5	41.6	20.0	25.1

Table 1. Average per-class accuracy in relation prediction on VG8K-LT and GQA-LT datasets. The best performance for each column is underlined.

strategy: RelMix [1].

Evaluation Metrics. We evaluate the model by calculating the top-1 accuracy overall for all classes. Considering the long-tail distribution, we illustrate the accuracy of each split to intuitively reflect the impact of the number of examples on the results.

Implemental Details. All the experiments are conducted with 8 NVIDIA 3090 GPUs, and the batch size is 8. We train our network with AdamW, 8 epochs on both GQA-LT and VG8K-LT, setting the initial backbone's learning rate to 10-5 and others to 10-4, and weight decay to 10-4. We do not directly apply CLIP image pre-processing, due to the center-crop process which is not suitable for object detection. We use random horizontal flip augmentation during training, resizing the input images to 672. The SCG and classifiers are trained with a default dropout of 0.1.

Architecture	Learning Methods	VG8K-LT many 267	 medium 799	 few 4,264	 all 5,330	GQA-LT many 86	 medium 255	 few 1,362	 all 1,703
LSVRU	VilHub [2]	**61.6**	20.3	10.1	14.2	68.6	44.0	10.3	18.3
LSVRU	VilHub + RelMix [2]	59.5	15.1	10.4	13.6	**68.8**	42.1	10.1	18.1
LSVRU	EQL [23]	56.9	12.1	10.0	12.7	68.9	43.7	10.0	18.0
LSVRU	WCE	52.8	27.2	10.8	14.5	53.4	42.0	14.0	20.2
RelTransformer	WCE	50.1	31.3	13.7	18.0	50.3	46.2	28.7	32.4
Ours	WCE	41.5	**38.9**	**21.5**	**25.1**	47.7	**51.6**	**37.6**	**40.2**

Table 2. Average per-class accuracy for subject/object prediction on VG8K-LT and GQA-LT datasets.

4.3. Quantitative Results

Tab.1 and Tab.2 illustrate the results on VG8K-LT and GQA-LT datasets. For VG8K-LT, compared to the other methods, we achieve the best results in all categories. We can see a consistent improvement over the whole band for relation classification, surpassing 3.2% and 4.8% for the medium and few category respectively. It is worth noting that our model achieves an accuracy of 25.1% for sub/obj classification, which outperforms other state-of-the-art models by a large margin. For GQA-LT, we outperform for the overall sub/obj categories and improve 8.9% on the tails. While our model is not as good as the RelTransformer [6] for relation category. We are still exploring this phenomenon since we have not achieved the reported performance yet due to the limited GPU memory.

4.4. Ablation Studies

To validate the design choices of the whole model, we conduct extensive ablation studies from two aspects on VG8K-LT, since it is more complex and covers more categories. Results are in Tab.3 and Tab.4.

Representation quality. We ablate our model with different backbones and the version without refining the relation triplet, the performance is shown in Tab.3. The results indicate that both the ViT backbone and Reltransformer module can benefit all categories, and when they are combined together, gaining the highest performance improvement.

Classifier quality. In Tab.4, we evaluate the performance of our model by each time removing one of the three modules that affect the quality of the classifier. We can find the negative sample reduction strategy has the most significant impact on improving the model's performance, especially for medium and few relations, resulting in an improvement of 6.9% and 6.7% respectively.

Backbone	SCG	RelTrans	relation many	 medium	 few	 all
ResNet-50	✓	✗	27.6	24.0	16.9	18.5
ResNet-50	✓	✓	29.0	27.5	17.8	19.8
ViT-B/16	✓	✗	33.2	27.2	18.5	20.5
ViT-B/16	✓	✓	**37.1**	**30.6**	**21.1**	**23.4**

Table 3. Ablation study on VG8K-LT dataset about representation quality. **RelTrans** represents the RelTransformer module.

NSR	Decoupled	Loss Function	relation many	 medium	 few	 all
✗	✓	WCE	32.6	23.7	14.4	16.7
✓	✓	CE	30.2	22.2	16.4	17.9
✓	✗	WCE	36.1	30.2	20.4	22.7
✓	✓	WCE	**37.1**	**30.6**	**21.1**	**23.4**

Table 4. Ablation study on VG8K-LT dataset about classifier quality. NSR:Negative Samples Reduction. Decoupled: Decoupling classifier. WCE: Weighted cross entropy.

5. Conclusion

We proposed to learn more effective feature representation for the LTVRR task by using the powerful backbone and relationship recognition and refinement modules. Previous methods rely on those backbones and can only learn coarse visual representation, leading to the performance of the relationship prediction dropping significantly. We adopt the ViT backbone to overcome this limitation to extract finer features. Moreover, we leverage the SCG to learn the pairwise relationships between the subject and object. Combined with RelTransformer, our model can predict more precisely the relationships between the subject and object. Extensive experimental results demonstrate that our model outperforms, especially in VG8K-LT, which is more challenging and consists of more classes.

References

[1] Sherif Abdelkarim, Aniket Agarwal, Panos Achlioptas, Jun Chen, Jiaji Huang, Boyang Li, Kenneth Church, and Mohamed Elhoseiny. Exploring long tail visual relationship recognition with large vocabulary. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15921–15930, 2021. 1, 3, 4

[2] Sherif Abdelkarim, Aniket Agarwal, Panos Achlioptas, Jun Chen, Jiaji Huang, Boyang Li, Kenneth Church, and Mohamed Elhoseiny. Exploring long tail visual relationship recognition with large vocabulary. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15921–15930, 2021. 4

[3] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 2

[4] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 112–121, 2021. 2

[5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 2

[6] Jun Chen, Aniket Agarwal, Sherif Abdelkarim, Deyao Zhu, and Mohamed Elhoseiny. Reltransformer: A transformer-based long-tail visual relationship recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19507–19517, 2022. 1, 2, 3, 4

[7] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. In *European Conference on Computer Vision*, pages 95–110. Springer, 2020. 1, 2

[8] Chris Drumnond. Class imbalance and cost sensitivity: Why undersampling beats oversampling. In *ICML-KDD 2003 Workshop: Learning from Imbalanced Datasets*, 2003. 2

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[11] Drew A Hudson and Christopher D Manning. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*, 3(8):1, 2019. 3

[12] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7610–7619, 2020. 2

[13] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *Eighth International Conference on Learning Representations (ICLR)*, 2020. 2

[14] Siddhesh Khandelwal and Leonid Sigal. Iterative scene graph generation. *arXiv preprint arXiv:2207.13440*, 2022. 1

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 3

[16] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020. 2

[17] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019. 1

[18] Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17152–17162, 2023. 1, 2

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[20] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pages 467–482. Springer, 2016. 2

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[22] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021. 2

[23] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11662–11671, 2020. 3, 4

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[25] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021. 2

[26] Jie Yang, Bingliang Li, Fengyu Yang, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Boosting human-object interaction detection with text-to-image diffusion model. *arXiv preprint arXiv:2305.12252*, 2023. 2

[27] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694, 2019. 1

[28] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13319–13327, 2021. 1, 2

[29] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2

[30] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020. 2