Language Adaptive Weight Generation for Multi-task Visual Grounding Supplementary Material

A. Experiments on Swin Transformer

Apart from using ViT [3] as the visual backbone, we also conduct experiments on the Swin Transformer [8]. Following the settings reported by QRNet [14] and LAVT [13], we evaluate the proposed VG-LAW framework on Swin-S and Swin-B for REC and RES tasks, respectively. The main results are summarized to Tab. 1.

Methods	Visual Backbone	Multi-task	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val	RefCOCOg test	ReferItGame test
REC:											
RefTR [7]	RN101	✓	82.23	85.59	76.57	71.58	75.96	62.16	69.41	69.40	71.42
QRNet [14]	Swin-S	✗	84.01	85.85	82.34	72.94	76.17	63.81	71.89	73.03	74.61
VG-LAW	Swin-S	✗	84.82	87.22	81.94	74.36	78.49	65.24	75.61	76.28	74.83
VG-LAW	Swin-S	✓	85.77	87.48	82.36	75.65	80.12	65.48	76.33	76.81	76.08
RES:											
RefTR [7]	RN101	✓	70.56	73.49	66.57	61.08	64.69	52.73	58.73	58.51	-
LAVT [13]	Swin-B	✗	74.46	76.89	70.94	65.81	70.97	59.23	63.62	63.66	-
VG-LAW	Swin-B	✗	75.09	77.02	72.46	66.56	70.67	59.09	64.43	65.39	-
VG-LAW	Swin-B	✓	75.37	77.31	72.64	66.81	70.92	59.41	65.46	65.68	-

Table 1. Comparison with state-of-the-art methods on RefCOCO [15], RefCOCO+ [15], RefCOCOg [10] and ReferItGame [5] for REC and RES tasks. RN101, Swin-S, and Swin-B are shorthand for the ResNet101, Swin-Transformer Small, and Swin-Transformer Base, respectively. We highlight the best and second best performance in the red and blue colors.

It can be observed that: (1) for the REC task, VG-LAW achieves the best performance on all four datasets when using the multi-task configuration, and the second-best performance except for the testB split on RefCOCO. Compared to the state-of-the-art REC method QRNet [14], which follows the TransVG [2] by using the transformer-based cross-modal interaction module and introduces extra multiscale fusion structures, ours VG-LAW is more compact and lightweight by just using a Swin backbone filled with expression-adaptive weights and a neat multi-task head. (2) For the RES task, VG-LAW achieves the best and second-best performance on RefCOCO and RefCOCOg datasets, and comparable performance on the RefCOCO+ dataset. Compared to the state-of-the-art RES method LAVT [13], which introduces the PWAM module based on the scaled dot-product attention and FPN-like decoder head, the VG-LAW is still compact and lightweight. The most obvious difference between VG-LAW and PWAM is that VG-LAW simply modifies the weights which are then used to extract expression-aware visual features, whereas PWAM incorporates linguistic features directly into the computation of the dot-product attention.

B. Experiments on Large-scale Pre-training Datasets

To compare with the methods [1, 4, 9, 12] trained on large-scale datasets, we also build a large-scale pre-training dataset by collecting images and annotations from the train split of RefCOCO/+/g, ReferItGame, Flickr30k Entities [11], and VG regions [6]. This dataset contains 174K images with nearly 6.1M referring expressions. We pre-train the models for 40 epochs with a batch size of 512, which are then fine-tuned on each specific dataset for 20 epochs with a batch size of 256. The pre-training results are summarized to Tab. 2.

C. Comparison of FLOPs and Inference Time

We also evaluate the FLOPs using fvcore and inference time on one 1080Ti GPU. The results are summarized to Tab. 3.

Method	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val-u	RefCOCOg test-u
ViLBERT [9]	-	-	-	72.34	78.52	62.61	-	-
VL-BERT [12]	-	-	-	72.59	78.57	62.30	-	-
UNITER [1]	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
MDETR [4]	87.51	90.40	82.67	81.13	85.52	72.96	83.35	83.31
VG-LAW	89.27	91.63	86.46	81.56	85.77	74.10	83.56	84.37

Table 2. Comparison with large-scale pre-training SOTA methods.

Method	Multi-task	Visual Backbone	FLOPs(G)	Runtime(ms)
QRNet [14]	✗	Swin-S	81.9	64.7
LAVT [13]	✗	Swin-B	193	67.9
VG-LAW	✗	ViT-B	74.3	49.4
VG-LAW	✓	ViT-B	77.3	50.2

Table 3. Comparison of FLOPs and inference time.

References

[1] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120, 2020. 1, 2

[2] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *ICCV*, pages 1769–1779, 2021. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[4] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021. 1, 2

[5] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 1

[6] 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. *IJCV*, 2017. 1

[7] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *NeurIPS*, 34:19652–19664, 2021. 1

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 1

[9] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 32, 2019. 1, 2

[10] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, pages 792–807, 2016. 1

[11] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 1

[12] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 1, 2

[13] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18155–18165, 2022. 1, 2

[14] Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *CVPR*, pages 15502–15512, 2022. 1, 2

[15] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 1