ViM: Vision Middleware for Unified Downstream Transferring
Supplementary Material

Yutong Feng[1], Biao Gong[1], Jianwen Jiang[1], Yiliang Lv[1], Yujun Shen[2], Deli Zhao[1], Jingren Zhou[1]
[1]Alibaba Group [2]Ant Group
{fengyutong.fyt, a.biao.gong, jianwen.alan, shenyujun0302, zhaodeli}@gmail.com
{yiliang.lyl, jingren.zhou}@alibaba-inc.com

1. Details of ViM Training

In this section, we present the detailed training configurations and results of all the midstream tasks.

1.1. Midstream Training Configurations

Image Classification. A fully-connected layer is appended after the backbone for classification learning. For most datasets, the model is trained in 100 epochs (30 epochs for ImageNet-21K [9]) with batch size of 256, optimized by the AdamW [27] optimizer with the initial learning rate of $1e-3$ and weight decay of $1e-4$. The learning rate is decayed following the cosine scheduler with the minimum of $1e-5$. For single-label classification task, we use the cross-entropy loss and evaluate with top-1 accuracy. For multi-label classification task, we use the binary cross-entropy loss and evaluate with mAP.

Object Detection. The ViTDet [23] framework with FastRCNN detector is adopted for learning object detection, where a feature pyramid with size $1/32$, $1/16$, $1/8$ and $1/4$ of the original image size is generated by convolution layers on the last-layer feature map of the backbone. The images are resized into size of 512×512 with large-scale jittering in $[0.1, 2.0]$. The models are trained for 100 epochs (10 epochs for Objects365 [32]) and batch size of 64, optimized by the AdamW optimizer with the initial learning rate of $1e-4$, which times 0.1 in the 89 and 96 epoch.

Instance Segmentation. We follow the configuration of object detection, except for the task-head is replaced by MaskRCNN to detect instances with dense masking.

Semantic Segmentation. We adopt a task-head of the UperNet [43], the input of which are from the transposed 2D-convolution layers or max pooling layers on the 3, 5, 7 and 11 ViT layer's feature map, following BEiT [2]. The images are resized into size of 512×512. We train 160K iterations with batch size of 16. The optimizer is AdamW with initial learning rate $3e-5$ and weight decay 0.05, and the learning rate is linearly decayed to 0.

Keypoints Detection. We follow the ViTDet [23] framework, replacing the detector by KeypointRCNN, which is a simple adoption of MaskRCNN via viewing keypoints as one-hot masks. The images are resized into size of 512×512. We train 90K iterations and keep the remaining configurations same as the object detection.

Depth Estimation. To predict the depth map, we feed the last layer's output into a decoder module. The decoder contains a sequence of 3 deconvolution layers (kernel size as 2×2 and hidden dimensions as $512/256/128$), 2 convolution layers (kernel size as 3×3 and hidden dimension as 128, and a up-sampling layer (ratio as 2.0). We train 25 epochs with batch size of 24. The optimizer is AdamW with intial learning rate $3e-5$ and weight decay 0.05. The learning rate follows the cosine scheduler.

Visual Question Answering. We adopt a baseline solution with two-tower framework. Except the existing visual backbone, we use an additional pre-trained BERT-base [10] text encoder to extract embeddings of the input question. The text embedding is then concatenated with the visual backbone generated embedding, and fed into a MLP classifier to find the correct answer of current question. We train the model in 60K iterations with batch size of 480. The optimizer is AdamW with initial learning rate of $5e-5$, which is linearly decayed.

Vision-and-Language Tasks. For the remain vision-and-language tasks, we adopt the PEVL [45] framework that unifies these tasks into a masked language modeling (MLM) format, including the visual question answering (on GQA [19]), referring expression comprehension, phrase grounding, visual relationship detection (VRD) and visual commonsense reasoning (VCR). During the training process, the task input is converted into masked sentence, fed into the pre-trained text encoder and predict to fill-in the masking token together with the image encoder.

Self-Supervised Learning. We adopt tasks introduced by 2 self-supervised learning (SSL), *i.e.*, contrastive learning with MoCo-v2 [6] and masked image modeling with MAE [17]. (i) For contrastive learning, images are strongly augmented into two views, and the model is trained to

Table 1. **Full list of midstream training datasets and results.** Some of the midstream results are not presented for varying reasons, *e.g.*, missing labels, empty categories, no evaluation metric.

Task Type (#mid-tasks)	Datasets	Type	Size	Results ([metric]: [value])
Global Recognition				
Image Classification (21)	ImageNet-21K-P [9, 31]	common	12M	top-1: 42.26
	ImageNet-1K [9]	common	1.33M	top-1: 82.04
	iNaturalist-2018 [18]	natural	0.46M	top-1: 67.31
	iNaturalist-2021 [18]	natural	2.79M	top-1: 73.34
	iWildCam-2022 [3]	natural	0.20M	top-1: 59.42
	Herbarium-2021 [8]	plant	2.26M	top-1: 63.67
	Danish Fungi 2020 [29]	fungus	0.30M	top-1: 70.96
	Tsinghua Dogs [51]	dog	0.07M	top-1: 84.23
	NABirds [37]	bird	0.02M	top-1: 82.49
	Places365 [26]	scene	1.84M	top-1: 56.29
	GLD-v2 [42]	landmark	1.58M	top-1: 70.58
	BigEarthNet-S2 [36]	satellite	0.59M	mAP: 80.73
	MLRSNet [30]	satellite	0.11M	mAP: 88.04
	iMaterialist-2018 [15]	fashion	1.01M	-
	iMet-2019 [1]	art	0.11M	-
	CelebA [25]	face	0.20M	mAP: 81.02
	CompCars [44]	car	0.63M	top-1: 98.17
	Logo-2K+ [39]	logo	0.17M	top-1: 88.36
	SOP [35]	product	0.12M	top-1: 68.86
	FoodX-251 [20]	food	0.13M	top-1: 77.19
	Food-101 [4]	food	0.10M	top-1: 92.55
Local Recognition				
Object Detection (7)	Objects365 [32]	common	1.8M	AP_{box}: 16.15, $AP_{box}50$: 27.13
	COCO [24]	common	123K	AP_{box}: 38.74, $AP_{box}50$: 61.34
	LVIS [16]	common	100K	AP_{box}: 24.71, $AP_{box}50$: 42.37
	DHD-traffic [28]	traffic	50K	AP_{box}: 49.21, $AP_{box}50$: 75.08
	DHD-campus [28]	campus	45K	AP_{box}: 49.08, $AP_{box}50$: 74.94
	LogoDet-3K [38]	logo	159K	AP_{box}: 63.53, $AP_{box}50$: 88.67
	CrowdHuman [33]	person	19K	AP_{box}: 30.57, $AP_{box}50$: 63.98
Instance Segmentation (2)	COCO [24]	common	123K	AP_{seg}: 34.41, $AP_{seg}50$: 57.42
	LVIS [16]	common	100K	AP_{seg}: 23.89, $AP_{seg}50$: 39.54
Semantic Segmentation (4)	ADE20K [50]	common	20K	mIoU: 44.47
	COCO-Stuff-164K [5]	common	164K	mIoU: 45.78
	COCO-Stuff-10K [5]	common	10K	mIoU: 43.31
	iSAID [41]	satellite	46K	mIoU: 45.78
Keypoints Detection (1)	COCO-keypoints [24]	person	57K	AP_{kpt}: 54.37, $AP_{kpt}50$: 80.29
Depth Estimation (2)	NYU Depth V2 [34]	indoor	25K	δ_1: 86.86, RMSE: 0.41
	KITTI [12, 13]	traffic	23K	δ_1: 95.11, RMSE: 2.52
Vision and Language				
Visual Question-Answering (2)	VQA-v2 [14]	common	265K	val acc: 35.76
	GQA [19]	common	110K	val acc: 66.14
Referring Expression Comprehension (3)	RefCOCO [47]	common	20K	val acc: 79.92, testA: 84.30, testB: 71.62
	RefCOCOg [47]	common	20K	val acc: 60.79
	RefCOCO+ [47]	common	20K	val acc: 55.84
Phrase Grounding (1)	Flickr30K [46]	common	32K	val acc: 55.84
Visual Relationship Detection (1)	Visual Genome [22]	common	101K	R@20: 54.67, R@50: 60.19, R@100: 61.85
Visual Commonsense Reasoning (1)	VCR [48]	common	100K	-
Self-Supervised Learning				
Contrastive Learning (1)	ImageNet-1K [9]	common	1.33M	-
Masked Image Modeling (1)	ImageNet-1K [9]	common	1.33M	-

Table 2. Detailed results of downstream classification.

Midstream	Downstream	VTAB-1k			FGVC				
		Natural	Specialized	Structural	CUB-200	NABirds	Flowers	Dogs	Cars
-	Fully	74.55	85.06	54.51	89.18	90.93	98.86	87.06	94.32
ImageNet-21K ft.	ConvPass	**82.38**	86.97	55.84	**90.21**	**91.27**	**99.66**	**90.42**	91.18
Objects365 ft.	ConvPass	67.09	83.25	52.70	83.02	85.17	96.78	81.93	90.05
COCO-Stuff164K ft.	ConvPass	62.85	80.67	50.51	77.63	78.02	95.46	77.26	87.09
-	Linear	71.47	81.48	31.32	79.10	75.64	94.60	77.23	81.84
-	VPT	78.10	82.47	53.25	81.64	77.17	96.75	81.05	89.69
-	Adapter	77.99	84.69	56.61	86.78	88.63	98.70	84.83	91.54
-	ConvPass	77.56	84.66	57.13	86.56	87.84	98.73	85.24	92.72
+ViM	**ViM-agg (rep.)**	79.14	86.21	**58.98**	86.49	88.51	98.78	84.49	92.66
+ViM	**ViM-agg (ens.)**	79.87	**87.18**	58.89	88.07	90.56	99.11	86.78	**94.16**

pair these two views among many negative images. We train with defaulted configuration of MoCo-v2, including the queue size of $65,536$, momentum 0.999, temperature 0.07 and MLP head. (ii) For masked image modeling, the visual backbone is trained as encoder on randomly masked image patches, and another decoder module is introduced to recover the image. We use the default configuration of MAE decoder with 8 layers and dimension of 512. Considering there are 2D-convolution inside the ViM module, which is not suitable for forwarding on sampled image patches, we append additional parameters as the mask tokens to fulfill the masked patches.

1.2. Midstream Training Results

We then present the training results, together with the full list of midstream training datasets in Table 1. It is noteworthy that we DO NOT require to achieve *competitive performance with the SoTA* methods for the following reasons: (i) The final goal of ViM is to benefit unified downstream transferring instead of midstream tasks, thus the midstream training results are only referenced to understand how the ViM module learn about each task. (ii) Only the parameters of ViM module is trained in the midstream, without fine-tuning the backbone model. (iii) Considering specific conditions for each task, we might not use the most advanced training configurations. For instance, we train with resolution of 512×512 for objection detection since the plain ViT backbone requires large computation cost.

2. Details of Downstream Transferring

In this section, we present the detailed configurations of downstream transferring and more transferring results .

2.1. Downstream Training Configurations

Classification on VTAB-1K. For all the 19 datasets in the VTAB-1k [49] benchmark, we firstly train 100 epochs on the train and validation sets with $1,000$ samples, then evaluate on the test sets. We train with batch size 64, intial learning rate $1e-3$ and weight decay $1e-4$. The optimizer is AdamW, with 10 epochs of learning rate warm-up and cosine decaying scheduler.

Classification on FGVC. For the 5 datasets in the FGVC benchmark, we train 50 epochs with batch size 64, optimized by AdamW with initial learning rate $1e-3$ and weight decay 0.01, with 10 epochs of warm-up and cosine decaying scheduler.

Object Detection. For training object detection on PASCAL [11] and Cityscapes [7], we adopt the ViTDet [23] similar to midstream training configurations. We train with image resolution of 1024×1024. To reduce the computation cost of plain ViT on larger images, we follow ViTDet to apply window-based attention in layers except the layer 2, 5, 8 and 11, with window size of 14. We also follow ViTDet and BEiT [2] to append relative position bias to better training. We train 24K iterations for both datasets, with learning rate decaying at iteration 16K and 21K.

Semantic Segmentation. We introduce a semantic FPN [21] module to learn the segmentation task. The input images are resizes into resolution of 512×512. We train 20K/10K iterations for PASCAL [11]/LoveDA [40] with batch size 16, using AdamW optimizer with initial learning rate $3e-5$ and weight decay 0.05.

Depth Estimation. We follow the same configuration as the midstream training of depth estimation.

2.2. Detailed Results of Downstream Classification

For the downstream classification, we evaluate with two benchmarks in the main experiment. Here we present the detailed classification results for each dataset in the benchmarks. The results are shown in Table 2.

References

[1] iMet Collection 2019 - FGVC6. 2

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *Int. Conf. Learn. Represent.*, 2021. 1, 3

[3] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021. 2

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 2

[5] Holger Caesar, Jasper R. R. Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1209–1218, 2018. 2

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3213–3223, 2016. 3

[8] Riccardo de Lutio, John Y Park, Kimberly A Watson, Stefano D'Aronco, Jan D Wegner, Jan J Wieringa, Melissa Tulig, Richard L Pyle, Timothy J Gallaher, Gillian Brown, et al. The herbarium 2021 half–earth challenge dataset and machine learning competition. *Frontiers in Plant Science*, page 3320, 2022. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009. 1, 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics, 2019. 1

[11] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson W. H. Lau. Location-aware single image reflection removal. In *Int. Conf. Comput. Vis.*, pages 4997–5006, 2021. 3

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Adv. Neural Inform. Process. Syst.*, 2014. 2

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, pages 1231–1237, 2013. 2

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6325–6334, 2017. 2

[15] Sheng Guo, Weilin Huang, Xiao Zhang, Prasanna Srikhanta, Yin Cui, Yuan Li, Hartwig Adam, Matthew R Scott, and Serge Belongie. The imaterialist fashion attribute dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[16] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5356–5364, 2019. 2

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16000–16009, 2022. 1

[18] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8769–8778, 2018. 2

[19] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6700–6709, 2019. 1, 2

[20] Parneet Kaur, Karan Sikka, Weijun Wang, Serge Belongie, and Ajay Divakaran. Foodx-251: a dataset for fine-grained food classification. *arXiv preprint arXiv:1907.06167*, 2019. 2

[21] Alexander Kirillov, Ross B. Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6399–6408, 2019. 3

[22] 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. *Int. J. Comput. Vis.*, pages 32–73, 2017. 2

[23] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 1, 3

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755, 2014. 2

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2

[26] Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós, and Álvaro García-Martín. Semantic-aware scene recognition. *Pattern Recognit.*, page 107256, 2020. 2

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[28] Yanwei Pang, Jiale Cao, Yazhao Li, Jin Xie, Hanqing Sun, and Jinfeng Gong. Tju-dhd: A diverse high-resolution dataset for object detection. *IEEE Trans. Image Process.*, 2020. 2

[29] Lukáš Picek, Milan Šulc, Jiří Matas, Thomas S. Jeppesen, Jacob Heilmann-Clausen, Thomas Læssøe, and Tobias Frøslev. Danish fungi 2020 - not just another image recognition dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1525–1535, January 2022. 2

[30] Xiaoman Qi, Panpan Zhu, Yuebin Wang, Liqiang Zhang, Junhuan Peng, Mengfan Wu, Jialong Chen, Xudong Zhao,

Ning Zang, and P Takis Mathiopoulos. Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:337–350, 2020. 2

[31] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 2

[32] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Int. Conf. Comput. Vis.*, pages 8430–8439, 2019. 1, 2

[33] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 2

[34] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *Eur. Conf. Comput. Vis.*, pages 746–760, 2012. 2

[35] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[36] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904, 2019. 2

[37] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 595–604, 2015. 2

[38] Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, and Shuqiang Jiang. Logodet-3k: A large-scale image dataset for logo detection. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, pages 1–19, 2022. 2

[39] Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, Haishuai Wang, and Shuqiang Jiang. Logo-2k+: A large-scale logo dataset for scalable logo classification. In *Assoc. Adv. Artif. Intell.*, pages 6194–6201, 2020. 2

[40] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, 2021. 3

[41] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. 2

[42] T. Weyand, A. Araujo, B. Cao, and J. Sim. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[43] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Eur. Conf. Comput. Vis.*, pages 432–448, 2018. 1

[44] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3973–3981, 2015. 2

[45] Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Pevl: Position-enhanced pre-training and prompt tuning for vision-language models. In *Annual Conference on Empirical Methods in Natural Language Processing*, 2022. 1

[46] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, pages 67–78, 2014. 2

[47] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Eur. Conf. Comput. Vis.*, pages 69–85, 2016. 2

[48] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6720–6731, 2019. 2

[49] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 3

[50] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5122–5130, 2017. 2

[51] Ding-Nan Zou, Song-Hai Zhang, Tai-Jiang Mu, and Min Zhang. A new dataset of dog breed images and a benchmark for finegrained classification. *Computational Visual Media*, pages 477–487, 2020. 2