Cloud-Device Collaborative Learning for Multimodal Large Language Models

Supplementary Material

A. Overview

To enhance the completeness of our experiments, additional analyses on the generalization capabilities of the MLLM under various domain shifts are included in the supplementary material. These analyses are conducted from two perspectives: image-domain and text-domain, specifically illustrating the enhancement in the multimodal understanding capabilities of our approach when dealing with domain-shift scenarios. The following aspects are included in our supplementary material.

- Supplementary Experimental Analysis
  - Generalization capabilities in text-domain shifts
  - Generalization capabilities in image-domain shifts
  - Generalization capabilities in multimodal-domain shifts
  - Comparison complementary
  - Hyperparameter discussion
- Additional Visualization Results
  - Visualization results in text-domain shifts
  - Visualization results in image-domain shifts
  - Visualization results in multimodal-domain shifts
- Expanded Related Work
  - Expanded related work in MLLMs
  - Expanded related work in **C**loud-Device Collaborative Learning
  - Expanded related work in **C**ontinual Domain Adaptation.
- Demo Video and Dataset

B. Supplementary Experimental Analysis

B.1. Generalization capabilities in text-domain shifts

To validate the enhancement of the generalization performance of MLLMs deployed on the device-side by our method, we employed the VQAv2-to-AOKVQA to simulate text-domain distribution shifts in the real world. Initially, both the pocket-size MLLM on device-side and the teacher MLLM in the cloud were fine-tuned using the VQAv2 dataset. Subsequently, we utilized the AOKVQA dataset for testing, aiming to mimic scenarios in the open world where the image-domain remains largely unchanged while the text-domain input varies. The results, as presented in the main text, distinctly show that our approach demonstrates a notable advantage in enhancing the continual generalization capability of multimodal large models amidst changing text-domain inputs, compared to other recent domain-adaptation and knowledge distillation methods.

B.2. Generalization capabilities in image-domain shifts

To assess the enhancement of generalization performance in device-side deployed MLLMs by our method, we utilized the VQAv2-IDS to simulate image-domain distribution shifts in the real world. Initially, both the device-side pocket-size MLLM and the cloud-based teacher MLLM were fine-tuned using the VQAv2 training dataset. We then constructed the VQAv2-IDS dataset to emulate scenarios in the open world where the text-domain remains constant while the image-domain undergoes significant distribution shifts. Specifically, we modified 2,000 of the most difficult instances in VQAv2 test images by randomly introducing natural elements like rain, snow, and fog, and adjusting lighting conditions, creating the VQAv2-image-domain-shift (VQAv2-IDS) test image dataset. Subsequently, without altering the text-domain inputs of the VQAv2 test data, we replaced the image-domain with VQAv2-IDS for comparison against recent state-of-the-art (SOTA) domain-adaptation and knowledge distillation methods, as shown in Table 1. It is evident that the source-only MLLM deployed on device, constrained by its parameter size, exhibits weaker generalization capabilities for inputs with significant image domain shifts. In contrast, our method rapidly improves the performance of device-deployed MLLMs, achieving sustained generalization.

B.3. Generalization capabilities in multimodal-domain shifts

To validate the enhanced generalization performance of our method for MLLMs deployed on the device-side, we utilized the VQAv2-to-AOKVQA-IDS dataset to simulate multimodal-domain distribution shifts in the real world. Initially, both the pocket-size MLLM at the device-side and the cloud-based teacher MLLM were fine-tuned using the VQAv2 training dataset. In the testing phase, we employed images with image-domain shifts and used AOKVQA's input text as the input for the other modality. This approach helped establish a multimodal domain gap with the training data, simulating multimodal domain shifts in the open world. The VQAv2-to-AOKVQA-IDS dataset encompasses images with significant image-domain shifts and texts with notable text-domain shifts. Based on this, we conducted a series of comparative experiments, the results of which are shown in Table 2. Our method effectively improves the generalization performance of device-side deployed MLLMs in handling multimodal-domain shifts. For scenarios difficult to comprehend by source-only models,

Table 1. **Persistent generalization capability on VQAv2-IDS (image-domain shift).** During the training phase, we fine-tuned the MLLM using the VQAv2 dataset. For testing, we employed a newly constructed dataset, VQAv2-IDS, which introduces random variations of rain, snow, fog, and lighting adjustments to the images while retaining the original VQAv2 question information. The VQAv2-IDS dataset represents image-domain alterations designed to simulate various environmental changes within the image domain in an open-world setting. DA is VQA accuracy (%) calculated following [1] under direct answers. Gain (%) refers to the accuracy improvement compared with the source-only method.

Time	$t \longrightarrow$				
Round	1_{st}	2_{nd}	3_{rd}	Mean_{DA}	Gain_{DA}
Source-only [2]	35.47	35.47	35.47	35.47	/
TENT-continual [3]	36.04	36.20	35.86	36.03	+0.56
CoTTA [4]	35.52	35.94	35.43	35.63	+0.16
PKD [5]	38.06	38.09	38.05	38.07	+2.6
ChannelWiseDivergence [6]	38.40	38.07	38.60	38.36	+2.89
Ours (CD-CCA)	**41.41**	**41.49**	**41.64**	**41.51**	**+6.04**

Table 2. **Persistent generalization capability on VQAV2-to-AOKVQA-IDS (multimodal-domain shift).** During the training phase, we fine-tuned the MLLM using the VQAv2 dataset. For testing, we employed a newly constructed dataset, AOKVQA-IDS, which introduces random variations of rain, snow, fog, and lighting adjustments to the images while retaining the AOKVQA's question information. The VQAv2-to-AOKVQA-IDS dataset represents multimodal-domain (image & text) alterations designed to simulate various environmental changes within both the image and text domain in an open-world setting. MC and DA are VQA accuracy (%) calculated following [1] under different conditions (multiple choices and direct answers). Gain (%) refers to the accuracy improvement compared with the source-only method.

Time	$t \longrightarrow$									
Round	1_{st}		2_{nd}		3_{rd}		All			
Condition	MC	DA	MC	DA	MC	DA	Mean_{MC}	Mean_{DA}	Gain_{MC}	Gain_{DA}
Source-only [2]	46.55	43.60	46.55	43.60	46.55	43.60	46.55	43.60	/	/
TENT-continual [3]	46.72	43.75	45.50	43.75	47.24	43.06	46.48	43.52	-0.07	-0.08
CoTTA [4]	46.98	43.62	47.42	43.53	46.81	44.03	47.07	43.72	+0.52	+0.12
PKD [5]	48.64	46.38	48.38	**46.79**	49.25	46.69	48.76	46.62	+2.21	+3.02
ChannelWiseDivergence [6]	48.55	46.35	49.17	46.60	49.43	46.13	49.05	46.35	+2.50	+2.75
Ours (CD-CCA)	**50.22**	**46.88**	**51.27**	46.60	**51.52**	**46.94**	**51.00**	**46.81**	**+4.45**	**+3.21**

significant performance improvements were achieved after a single round of Cloud-Device Collaborative Learning.

With this, we have completed experimental analyses across three different types of domain shifts. Additionally, our experiments on the COCO-to-nocaps dataset, presented in the main text, further validate our CD-CCA framework's sustained generalization capabilities for MLLMs from the perspective of category-domain shift.

B.4. Comparison complementary

We have conducted more comparison experiments with prompt tuning methods [7–9] to further demonstrate the efficacy of our method (using the same experimental settings as in Table 2). Besides, we also implemented distillation methods of [10–14] in our system, As shown in Table 3, the results indicate that our method surpasses the aforementioned studies in terms of persistent generalization capability enhancement for device-side MLLMs.

Table 3. **Comparison Complementary (left) & Hyperparameter Discussion**. We conducted further comparative experiments on the continuous optimization of performance for multi-modal large models deployed on the device side, comparing with our state-of-the-art prompt tuning method and knowledge distillation method. The results are shown in the comparison experimental results (left), along with a discussion on hyperparameters in the system loss function (right) through ablation experiments, as presented in the right table.

Method	Mean (BLeU)
PromptSRC [7]	33.93
Black-VIP [8]	36.48
MaPLe [9]	35.92
DearKD [10]	37.92
Co-advise [11]	37.89
ALPKD [12]	38.01
ReviewKD [13]	37.65
SimKD [14]	37.52
Ours	**38.35**

λ_{query}	λ_{repr}	λ_{CE}	Mean (BLeU)
5	1	1	36.60
1	5	1	36.83
1	1	5	37.45
2	1	1	37.70
1	2	1	37.51
1	1	2	37.95
1	1	0.5	37.33
1	1	0.1	36.25
1	1	1	**38.35**

B.5. Hyperparameter discussion

Hyperparameters in Eq.6 are designed to balance various losses for effective parameter optimization. To provide a more detailed showcase of our method and further demonstrate its effectiveness, we conducted the following ablation

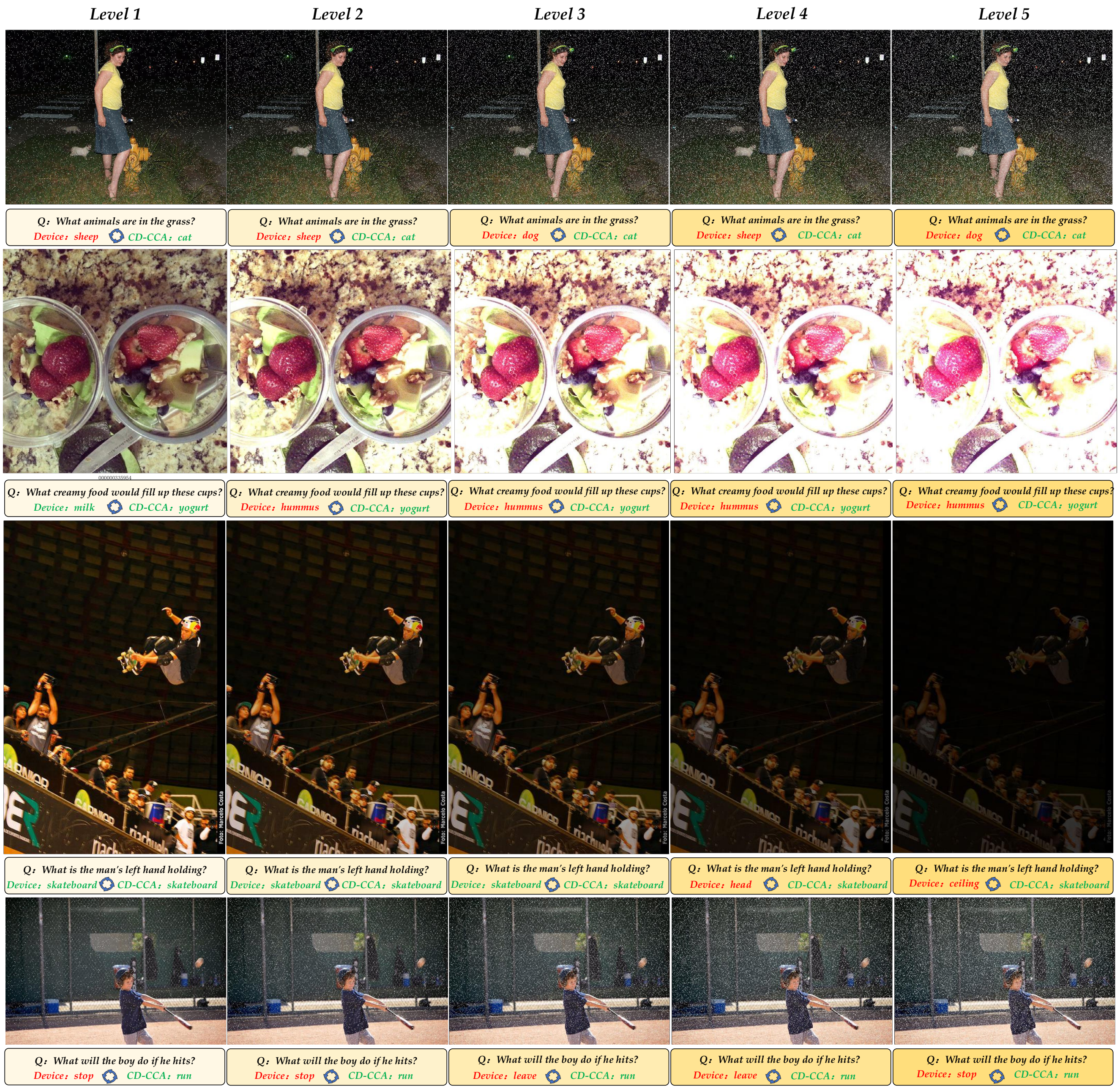
Level 1
Level 2
Level 3
Level 4
Level 5
Q：What animals are in the grass?
Device：sheep CD-CCA：cat
Q：What animals are in the grass?
Device：sheep CD-CCA：cat
Q：What animals are in the grass?
Device：dog CD-CCA：cat
Q：What animals are in the grass?
Device：sheep CD-CCA：cat
Q：What animals are in the grass?
Device：dog CD-CCA：cat
Q：What creamy food would fill up these cups?
Device：milk CD-CCA：yogurt
Q：What creamy food would fill up these cups?
Device：hummus CD-CCA：yogurt
Q：What creamy food would fill up these cups?
Device：hummus CD-CCA：yogurt
Q：What creamy food would fill up these cups?
Device：hummus CD-CCA：yogurt
Q：What creamy food would fill up these cups?
Device：hummus CD-CCA：yogurt
Foto: Marcelo Costa
Q：What is the man's left hand holding?
Device：skateboard CD-CCA：skateboard
Q：What is the man's left hand holding?
Device：skateboard CD-CCA：skateboard
Q：What is the man's left hand holding?
Device：skateboard CD-CCA：skateboard
Q：What is the man's left hand holding?
Device：head CD-CCA：skateboard
Q：What is the man's left hand holding?
Device：ceiling CD-CCA：skateboard
Q：What will the boy do if he hits?
Device：stop CD-CCA：run
Q：What will the boy do if he hits?
Device：stop CD-CCA：run
Q：What will the boy do if he hits?
Device：leave CD-CCA：run
Q：What will the boy do if he hits?
Device：leave CD-CCA：run
Q：What will the boy do if he hits?
Device：stop CD-CCA：run

Figure 1. **Visualization of Experimental Results under Different Domain Shifts.** We artificially introduced uncertainty elements (rain, snow, fog, brightness, etc.) into the multimodal inputs to simulate the continuously changing natural environments. The intensity of added uncertainty gradually increases from level 1 to level 5. The resulting figures illustrate the performance of the device-side deployed MLLM in visual question answering tasks, as well as the improved outcomes following the CD-CCA optimization of the device-side MLLM.

experiments. Experimental results are shown in Table 3.

C. Additional Visualization results

To visually demonstrate the enhanced continual generalization ability of device-side pocket-size MLLMs in handling domain shifts, facilitated by our proposed CD-CCA, we present the following experimental results in a visual format. As shown in Figure 1, we depict the comprehension capabilities of the device-side MLLM under various environments, both before and after the application of CD-CCA. It is clearly observable that, as domain shifts intensify, the generalization ability of the source-only MLLM deployed on the device-side progressively decreases. However, following efficient collaborative learning through CD-

CCA, a notable improvement in its generalization capability is achieved.

D. Expanded Related Work

MLLMs. Current MLLMs extend beyond linguistic processing by expanding the scale of data and model architecture, enabling real-world perception and addressing limitations in tasks like image captioning [15] and visual question answering [16]. Due to constraints in model size and training costs, some scholars attempted to predominantly utilize frozen LLM backbones, focusing exclusively on training visual components, or adopting more streamlined and efficient training strategies such as parameter-efficient fine-tuning[17] rather than training from scratch. Considering the limitations imposed by computational power and network bandwidth for model deployment on devices, merely reducing the number of model-trainable parameters is insufficient. Therefore, we propose the CD-CCA framework as a solution.

Cloud-Device Collaborative Learning. Merely offloading the computational workload to the cloud without considering the collaboration between the cloud and the device, although alleviating the computational limitations of the device, has minimal impact on enhancing the device's ability to handle complex model processing tasks. Some work has explored the transmission of tokens during the computation process on devices through an Uncertainty Guided Sampling [18] approach, aiming to enhance bandwidth utilization. However, this method exhibits strong randomness in the token selection, and the tokens selected may lack sufficient semantic information. Our UTS strategy allows us to reduce bandwidth while maximizing the semantic information richness of the selected image tokens. In the cloud, knowledge distillation can be leveraged to transfer knowledge from large models to smaller models, aiming to minimize the parameter volume transmitted from the cloud to the device.

Knowledge Distillation (KD) is a method of model compression and transfer learning [19]. Over the years, many KD methods have been proposed that perform distillation over intermediate features [20, 21], relation representation [22, 23], attention [24, 25] for various vision tasks. However, for MLLMs, there is currently no specific knowledge distillation method available to compress them effectively. In this paper, to better serve this system, we propose an adapter-based knowledge distillation(AKD) to get the manifolds embedded in multi-modal space.

Continual Domain Adaptation. When devices are deployed in the real world, the continuous variation of data in real-world scenarios poses significant requirements for the generalization capability of models. To achieve better generalization performance on target data without access to source data, TENT [26] optimizes the pre-trained model's Batch Normalization layers through entropy minimization, while SHOT [27] utilizes both entropy minimization and a diversity regularizer for information maximization. References [28] and [29] enhance model performance in target domains without source data by generating target-style data. Our work proposes a Cloud-Device Collaborative Continual Adaptation framework enables models to adapt to dynamically changing data distributions, significantly enhancing the generalization capability of device models.

E. Demo Video and Dataset

We provide a video demo (the attached MP4 file), which contains the motivation and intuitive introduction of our proposed CD-CCA paradigm, the workflow of the overall framework, and the visualization results. Furthermore, the domain-shift VQAv2-to-AOKVQA-IDS dataset has been made available on Google Drive for researchers to access and utilize.

References

[1] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 2

[2] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 2

[3] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 2

[4] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2022. 2

[5] Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *Advances in Neural Information Processing Systems*, 35: 15394–15406, 2022. 2

[6] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021. 2

[7] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 2

[8] Changdae Oh, Hyeji Hwang, Hee-young Lee, YongTaek Lim, Geunyoung Jung, Jiyoung Jung, Hosik Choi, and Kyungwoo Song. Blackvip: Black-box visual prompting for robust transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24224–24235, 2023. 2

[9] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2

[10] Xianing Chen, Qiong Cao, Yujie Zhong, Jing Zhang, Shenghua Gao, and Dacheng Tao. Dearkd: Data-efficient early knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12052–12062, 2022. 2

[11] Sucheng Ren, Zhengqi Gao, Tianyu Hua, Zihui Xue, Yonglong Tian, Shengfeng He, and Hang Zhao. Co-advise: Cross inductive bias distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 16773–16782, 2022. 2

[12] Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, and Qun Liu. Alp-kd: Attention-based layer projection for knowledge distillation. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 35, pages 13657–13665, 2021. 2

[13] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2021. 2

[14] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11933–11942, 2022. 2

[15] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 4

[16] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 4

[17] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp, 2019. 4

[18] Yulu Gan, Mingjie Pan, Rongyu Zhang, Zijian Ling, Lingran Zhao, Jiaming Liu, and Shanghang Zhang. Cloud-device collaborative adaptation to continual changing environments in the real-world, 2022. 4

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NeurIPS Workshop*, 2014. 4

[20] Tao Huang, Yuan Zhang, Shan You, Fei Wang, Chen Qian, Jian Cao, and Chang Xu. Masked distillation with receptive tokens. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=mWRngkvIki3. 4

[21] Tao Huang, Yuan Zhang, Mingkai Zheng, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge diffusion for distillation. *NeurIPS*, 30, 2023. 4

[22] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *ICLR*, 2020. 4

[23] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *CVPR*, pages 12319–12328, 2022. 4

[24] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *CVPR*, pages 4643–4652, 2022. 4

[25] Yuan Zhang, Weihua Chen, Yichen Lu, Tao Huang, Xiuyu Sun, and Jian Cao. Avatar knowledge distillation: Self-ensemble teacher paradigm with uncertainty. *arXiv preprint arXiv:2305.02722*, 2023. 4

[26] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization, 2021. 4

[27] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation, 2021. 4

[28] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9641–9650, 2020. 4

[29] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation, 2021. 4