Semantic Guided Feature Disentanglement and Reconstruction for Domain Adaptive Object Detection

Supplementary Material

6. Prompt training

Drawing inspiration from DA-Pro [26], our framework co-optimizes two complementary prompt types: cross-domain prompts that capture transferable features, and domain-aware prompts that retain domain-distinctive characteristics. The cross-domain prompts establish detection head uniformity between source and target environments through parameter synchronization, whereas the domain-aware prompts identify unique patterns by amplifying classification certainty within individual domains.

Our adaptive detection module computes the likelihood $P(y|r_j, d, \Omega)$ that region proposal r_j is assigned to category y in domain d:

$$P(y|r_j, d, \Omega) = \frac{\exp(\cos(\mathcal{V}(r_j), \mathcal{T}(t_y^d))/\tau)}{\sum_{\omega\in\Omega}\sum_{c=1}^{C}\exp(\cos(\mathcal{V}(r_j), \mathcal{T}(t_c^d))/\tau)}, \tag{14}$$

Here, $\cos(\cdot)$ denotes cosine similarity, while $\Omega \in \{\{s\}, \{t\}, \{s,t\}\}$ determines the prompt sets: source-only $\{t_c^s\}_{c=1}^C$, target-only $\{t_c^t\}_{c=1}^C$, or their combination $\{t_c^s\}_{c=1}^C \cup \{t_c^t\}_{c=1}^C$. For a domain-$d$ image x_i with detected regions $\{r_j^d\}_{j=1}^M$ and corresponding annotations $\{l_j^d\}_{j=1}^M$, the scross-entropy loss on the source domain is:

$$\mathcal{L}_s^\Omega = \mathbb{E}_{\mathcal{X}^s}\left[-\frac{1}{M}\sum_{j=1}^{M}\log P(l_j^s|r_j^s, s, \Omega)\right]. \tag{15}$$

To encourage domain-agnostic representations, we enforce consistent correct predictions from both classifiers:

$$\mathcal{L}_d^{\text{inv}} = \mathcal{L}_{s,\{s\}} + \mathcal{L}_{s,\{t\}}. \tag{16}$$

For domain-specific characteristics, we aim for the dedicated classifier to exhibit stronger confidence through joint prompt optimization:

$$\mathcal{L}_d^{\text{spec}} = \mathcal{L}_{s,\{s,t\}}. \tag{17}$$

The comprehensive learning objective for source data integrates both aspects:

$$\mathcal{L}_s = \mathcal{L}_s^{\text{inv}} + \mathcal{L}_s^{\text{spec}}. \tag{18}$$

For the unannotated target domain, we exploit CLIP's zero-shot recognition capacity to create pseudo-annotations. Specifically, we employ the template ”A photo of [class]” to designate the highest-probability category:

$$l_j^{tgt} = \operatorname*{argmax}_y P(y|r_j^{tgt}), \tag{19}$$

with probabilities derived from Eq.14. To mitigate noise from uncertain pseudo-labels, we filter predictions below confidence threshold τ. The target domain optimization objective then becomes:

$$\mathcal{L}_{t,\Omega} = \mathbb{E}_{\mathcal{X}^t}\left[-\frac{1}{M}\log P(l_j^t|r_j^t, t, \Omega)\right], \tag{20}$$

$$\mathcal{L}_t = \mathcal{L}_{t,\{t\}} + \mathcal{L}_{t,\{s,t\}}. \tag{21}$$

The unified optimization framework combines both domains:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_s + \mathcal{L}_t. \tag{22}$$

Table 6. mAP(%) comparison with existing methods on Pascal→ Watercolor and Pascal→ Comic tasks.

Methods	P → W	P → Comic
DBGL [3](ResNet-101)	53.8	29.7
FGRR [4] (ResNet-101)	55.7	32.7
SIGMA++[33] (ResNet-101)	57.1	37.1
DA-Pro [26] (ResNet-50)	58.1	44.6
Ours	58.9	46.3

7. More experimental results

Additional Experimental Results. To further assess the effectiveness of SG-FDR under more challenging and diverse domain shift conditions, we conduct evaluations on two harder benchmarks: Pascal→Watercolor and Pascal→Comic. Pascal VOC [12] is a large-scale real-world dataset with 20 object categories, comprising subsets from 2007 and 2012. Watercolor and Comic [19] are art-style datasets, each containing 1,000 training and 1,000 test images, and sharing 6 categories with Pascal VOC. Together, these benchmarks allow evaluation under more pronounced domain shifts and in multi-class scenarios. As shown in Table 6, our method, SG-FDR, using a weaker ResNet-50 backbone, consistently outperforms the other approaches across the two benchmarks, further validating the efficacy of SG-FDR.

Analysis of pseudo labels. To enhance domain-adaptive prompts, we leverage annotated source data and distill CLIP's powerful zero-shot semantic knowledge into the target-domain detection head. A naïve strategy is to employ soft labels derived from the fixed prompt “A photo of [class]”. However, forcing the model's predictions to align with such probability distributions causes the learnable

Table 7. The comparison (%) of different options of pesudo labels on Cross-Weather adaptation scenario Cityscapes→Foggy Cityscapes.

Options	Cityscapes→Foggy	Cityscapes KITTI→Cityscapes	Sim10K→Cityscapes
probabilities	54.3	60.8	62.1
pseudo label (DA-Pro)	55.9	61.4	62.9
pseudo label (Ours)	58.9	67.3	68.2

Table 8. Analysis of different initialization of $\mathbf{P}^c$.

Method	C →F	K →C	S →C
Baseline	48.6	59.1	58.9
+random	53.4	62.7	59.6
+CLIP-token-based	55.3	64.8	65.9
+two-stage pretrain	58.9	67.3	68.2

prompt to collapse toward the handcrafted prompt, thereby limiting discriminative learning. In contrast, threshold-based pseudo labels (i.e., hard labels) avoid encoding handcrafted inter-class relations and instead require the learned prompt to remain close to the correct category while being separated from incorrect ones. This promotes a more discriminative and domain-adaptive prompt space. Our experiments on the three benchmarks (C→ F, K→C, S→C) verify this effect: replacing hard pseudo labels with probability supervision leads to a performance drop, as shown in Table 7.

While pseudo-label noise is a common challenge across domain adaptation methods, our framework exhibits markedly stronger robustness under the same pseudo-label settings. The Semantic-Guided Disentanglement Module mitigates noisy domain-specific biases through adaptive semantic prompts, while the Adaptive Feature Reconstruction Module selectively extracts reliable transferable cues. Together, these modules consistently yield superior performance compared with existing pseudo-label-based approaches, highlighting the effectiveness of our design. Future work will further investigate confidence-aware weighting strategies and iterative pseudo-label refinement.

Analysis of prompt initialization. Table 8 shows that prompt initialization has a clear impact on cross-domain performance. Random initialization already improves over the baseline, indicating that learnable prompts provide useful adaptation capacity. Initializing $\mathbf{P}^c$ with CLIP token (CLIP hand-crafted prompt) embeddings yields further gains across all tasks, demonstrating the benefit of injecting pretrained semantic priors. The two-stage pretraining strategy refers to training on the source first, and then initializing as the ^{c} achieves the best results, significantly outperforming other variants, showing that warming up shared prompts with source supervision leads to more stable optimization and stronger cross-domain generalization. Overall, semantically enriched initialization consistently enhances prompt

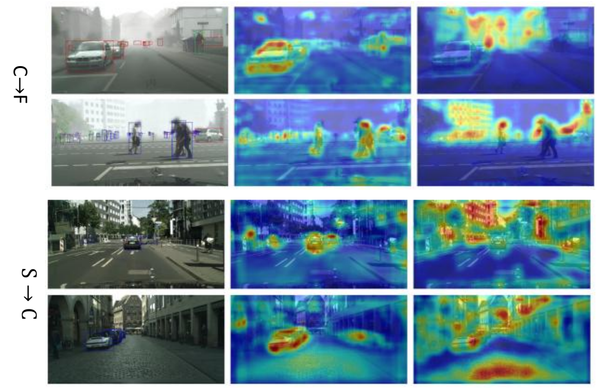

Figure 5. Qualitative analysis on the disentanglement effectiveness of our method under C → F and S → C.

discriminability and transferability.

Disentanglement Evidence Analysis. We conduct a qualitative evaluation to assess the effectiveness of our proposed disentanglement-based SG-FDR framework. Specifically, we visualize both domain-shared and domain-private features at the global level. As illustrated in Figure 5, the domain-shared features captured by our method focus predominantly on instance objects that are essential for object detection. Meanwhile, the domain-private features effectively encapsulate domain-unique characteristics, such as weather-related attributes. These observations indicate that our approach successfully separates domain-invariant and domain-specific information.

To quantitatively measure the disentanglement capability, we compute feature distribution discrepancies under the Cityscapes → Foggy Cityscapes adaptation scenario. We randomly sample 100 foreground object features per category from each domain and employ two metrics: Proxy A-distance (PAD) [1] and Earth Mover's Distance (EMD) [40]. Results in Table 9 confirm that our method achieves lower domain divergence in both global and local stages. With more discriminative domain-invariant features and reduced cross-domain discrepancy, our approach effectively enhances the baseline through feature disentanglement.

Table 9. Quantitative evaluation on the domain distance under the Cityscapes → Foggy Cityscapes setting.

Method	PAD ↓	EMD ↓
DA-Pro [26]	1.65	5.86
FGPro [49]	1.33	2.83
Ours	0.53	1.62

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006. 2

[2] Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang. Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23839–23848, 2023. 1, 6

[3] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 2683–2692. IEEE, 2021. 1

[4] Chaoqi Chen, Jiongcheng Li, Hong-Yu Zhou, Xiaoguang Han, Yue Huang, Xinghao Ding, and Yizhou Yu. Relation matters: Foreground-aware graph-based relational reasoning for domain adaptive object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3677–3694, 2023. 1

[5] Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, et al. Learning domain adaptive object detection with probabilistic teacher. *arXiv preprint arXiv:2206.06293*, 2022. 2

[6] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 1, 2, 6

[7] Zichong Chen, Jian Cheng, Ziying Xia, Yongxiang Hu, Xiaochen Li, Zhicheng Dong, and Nyima Tashi. Focusing on feature-level domain alignment with text semantic for weakly-supervised domain adaptive object detection. *Neurocomputing*, 622:129435, 2025. 6

[8] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5

[9] Yiming Cui, Liang Li, Jiehua Zhang, Chenggang Yan, Hongkui Wang, Shuai Wang, Heng Jin, and Li Wu. Stochastic context consistency reasoning for domain adaptive object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1331–1340, 2024. 6

[10] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. Masked face recognition challenge: The insightface track report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1437–1444, 2021. 1

[11] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4091–4101, 2021. 2

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, pages 303–338, 2010. 6, 1

[13] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. Poda: Prompt-driven zero-shot domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18623–18633, 2023. 3

[14] Yongchao Feng, Shiwei Li, Yingjie Gao, Ziyue Huang, Yanan Zhang, Qingjie Liu, and Yunhong Wang. Dsd-da: Distillation-based source debiasing for domain adaptive object detection. *arXiv preprint arXiv:2311.10437*, 2023. 6

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 6

[16] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3

[17] 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. 7

[18] Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu, and Ling Shao. Category contrast for unsupervised domain adaptation in visual tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1203–1214, 2022. 2

[19] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5001–5009. Computer Vision Foundation / IEEE Computer Society, 2018. 1

[20] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 6

[21] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016. 6

[22] Junjie Ke, Lihuo He, Bo Han, Jie Li, Di Wang, and Xinbo Gao. Vldadaptor: Domain adaptive object detection with

vision-language model distillation. *IEEE Transactions on Multimedia*, 2024. 2

[23] Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T Tan. Cat: Exploiting inter-class dynamics for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16541–16550, 2024. 6

[24] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022. 1

[25] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6092–6101, 2019. 1

[26] Haochen Li, Rui Zhang, Hantao Yao, Xinkai Song, Yifan Hao, Yongwei Zhao, Ling Li, and Yunji Chen. Learning domain-aware detection head with prompt tuning. *Advances in Neural Information Processing Systems*, 36:4248–4262, 2023. 2, 4, 6, 7, 8, 1, 3

[27] Haochen Li, Rui Zhang, Hantao Yao, Xin Zhang, Yifan Hao, Xinkai Song, and Ling Li. React: Remainder adaptive compensation for domain adaptive object detection. *IEEE Transactions on Image Processing*, 33:3735–3748, 2024. 6

[28] Haochen Li, Rui Zhang, Hantao Yao, Xin Zhang, Yifan Hao, Xinkai Song, Xiaqing Li, Yongwei Zhao, Yunji Chen, and Ling Li. Da-ada: Learning domain-aware adapter for domain adaptive object detection. *Advances in Neural Information Processing Systems*, 37:103574–103598, 2024. 2, 6, 8

[29] Haochen Li, Rui Zhang, Hantao Yao, Xin Zhang, Yifan Hao, Xinkai Song, Shaohui Peng, Yongwei Zhao, Chen Zhao, Yanjun Wu, et al. Seen-da: Semantic entropy guided domain-aware attention for domain adaptive object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25465–25475, 2025. 6, 8

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 3

[31] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan. Scan: Cross domain object detection with semantic conditioned adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1421–1428, 2022. 6

[32] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5291–5300, 2022. 1

[33] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9022–9040, 2023. 1

[34] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7581–7590, 2022. 6, 8

[35] Mingfu Liang, Jong-Chyi Su, Samuel Schulter, Sparsh Garg, Shiyu Zhao, Ying Wu, and Manmohan Chandraker. Aide: An automatic data engine for object detection in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14695–14706, 2024. 1

[36] Yabo Liu, Jinghua Wang, Chao Huang, Yaowei Wang, and Yong Xu. Cigar: Cross-modality graph reasoning for domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23776–23786, 2023. 1, 6

[37] Xinzhu Ma, Wanli Ouyang, Andrea Simonelli, and Elisa Ricci. 3d object detection from images for autonomous driving: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3537–3556, 2023. 1

[38] Poojan Oza, Vishwanath A Sindagi, Vishal M Patel, et al. Unsupervised domain adaptation of object detectors: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4018–4040, 2023. 1

[39] 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, 3

[40] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000. 2

[41] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6956–6965, 2019. 2

[42] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 5

[43] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3219–3229, 2023. 3

[44] Vibashan Vs, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4516–4526, 2021. 6

[45] Fei Wang, Luhui Zhao, Shijie Hong, Zhe Wang, Chen Liu, Changxin Gao, Jinsheng Li, Xin Li, and Dapeng Luo. Dual-domain teacher for unsupervised domain adaptation detection. *IEEE Transactions on Multimedia*, 2025. 1

[46] Hongsong Wang, Shengcai Liao, and Ling Shao. Afan: Augmented feature alignment network for cross-domain object detection. *IEEE Transactions on Image Processing*, 30: 4046–4056, 2021. 1

[47] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, YangYang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9603–9612, 2021. 6, 8

[48] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8032–8041, 2023. 1

[49] Zhitao Wen, Jinhai Liu, Huaguang Zhang, and Fengyuan Zuo. Exploring fine-grained visual-text feature alignment with prompt tuning for domain-adaptive object detection. *IEEE Transactions on Cybernetics*, 2025. 2, 6, 3

[50] Weixi Weng and Chun Yuan. Mean teacher detr with masked feature alignment: A robust domain adaptive detection transformer framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5912–5920, 2024. 6

[51] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9342–9351, 2021. 6

[52] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11724–11733, 2020. 1

[53] Liang Zhao and Limin Wang. Task-specific inconsistency alignment for domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14217–14226, 2022. 6

[54] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022. 3, 6

[55] Wenzhang Zhou, Heng Fan, Tiejian Luo, and Libo Zhang. Unsupervised domain adaptive detection with network stability analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6986–6995, 2023. 6