Semi-Supervised Domain Generalization for Object Detection via Language-Guided Feature Alignment (Supplementary Material)

Sina Malakouti
sem238@pitt.edu
Adriana Kovashka
kovashka@cs.pitt.edu

University of Pittsburgh,
Pittsburgh, PA, USA

1 Baselines

We evaluate our proposed model against source-only fully-supervised object detection models (FSOD), domain adaptation (DA), and domain generalization (DG) baselines on two benchmarks: *real-to-artistic* and *adverse weather*. This project leverages a ResNet50 backbone pre-trained by RegionCLIP; therefore, for a fair comparison, we only compare it with related works utilizing the ResNet50 backbone. Notably, larger ResNet models like ResNet101 are not yet provided for RegionCLIP [12]. To our knowledge, DIDN [5] is the only multi-domain DG work in object detection. However, DIDN only conducted experiments on adverse weather, and its implementation code is not publicly available, preventing fair comparison on *real-to-artistic*. Thus, on the DG task in real-to-artistic transfer, we compare to source-only FSOD baselines (i.e., Faster-RCNN and RegionCLIP), domain adaptation methods, namely Adaptive-MT [2], IRG [2], and our defined DG baselines, namely *Direct Visual Alignment (DVA)* and *Caption Pseudo-Labeling (Caption-PL)*.

Extending our method to DA, we compare it across the *real-to-artistic* domain with FSOD baselines, DA methods, and our DG benchmarks. We further compare our model on *adverse weather* DG task with source-only FSOD baselines and DIDN [5]. Following [5], and since most of the previous works have used these benchmarks in their experiments, we have extensively compared our model on the *City* → *Foggy* adaptation task, against both source-only FSOD and DA baselines. SW-DA [6], D&Match [3], MTOR [1], AFAN [8], GPA [11], SFA [9], DSS [10], TTD+FPN [2], and IRG [2] are the DA state-of-the-arts that are compared with on this task.

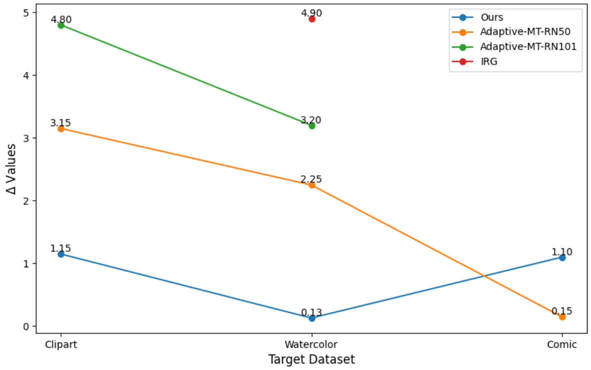


Figure 1: **Comparison of** Δ. Δ represents the drop in the performance on the target domain when the target is used as the source domain (i.e., DA) and when it is unknown throughout the training (i.e., DG). Hence, a lower number means that the model is more stable. Adaptive-MT-RN101 and IRG values are borrowed from [4] and [7], respectively. Since values for all settings were not reported, Δ is reported for Clipart and Watercolor for Adaptive-MT-RN101 and Watercolor for IRG. RN101 is the ResNet101 backbone, while RN50 is the ResNet50 backbone.

2 Qualitative results

2.1 Stability comparison

Fig. 1 demonstrates the stability of our method compared to the domain adaptation methods on the real-to-artistic benchmark. We define $\Delta = DA_{mAP} - DG_{mAP}$. For example, for Clipart, we subtract the average performance on Clipart when Clipart is used as a target domain in the DG task (i.e., Table 1 in the main text) from the performance on Clipart in a DA task (i.e., Table 4.a in the main text). Our model mAP drop on DG sets is relatively much lower than the DA methods, especially on Clipart and Watercolor. For example, on Clipart, our model's performance drops by 1.15%, while Adaptive-MT-RN50's performance drops by 3.15% and Adaptive-MT-RN101's performance drops by 4.8%. Comic Δ for our model is higher than Adaptive-MT-RN50; this is due to the fact that Adaptive-MT-RN50 performs extremely poorly on Comic in both DA and DG settings. For instance, in the DA task, our model achieves 46.3% while Adaptive-MT-RN50 achieves 23.4% (Table 4.a), and the average performance in the DG settings is 45.2% and 23.25% for ours and Adaptive-MT-RN50, respectively (Table 1 in the main text).

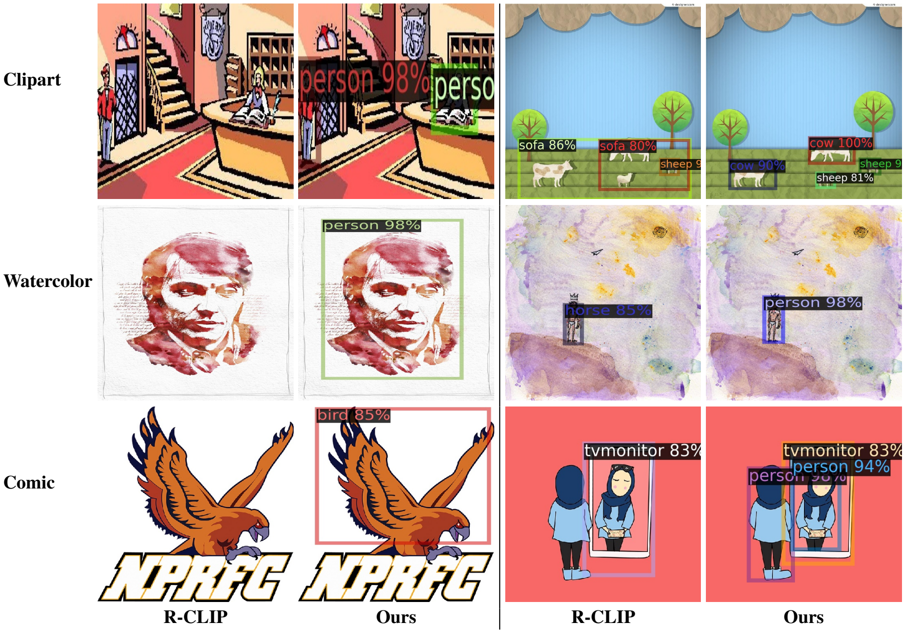
2.2 Caption comparison

Table 1 includes some examples of the generated caption on the validation set of VOC, Clipart, Watercolor, and Comic. Specifically, we use ClipCap to generate captions based on RegionCLIP baseline trained on VOC in a supervised manner and when our method is trained on *VOC, Clipart → Watercolor, Comic*. Even though our method does not produce perfect descriptions, especially in artistic domains, it constantly produces better and more meaningful descriptions than the RegionCLIP baseline, as captions may become meaningless on even natural images from the VOC dataset when FSOD finetuning. For instance,

Table 1: **Caption Comparison**. Comparing generated caption based on RegionCLIP baseline and CDDMSL visual features when training using VOC and Clipart as source domains

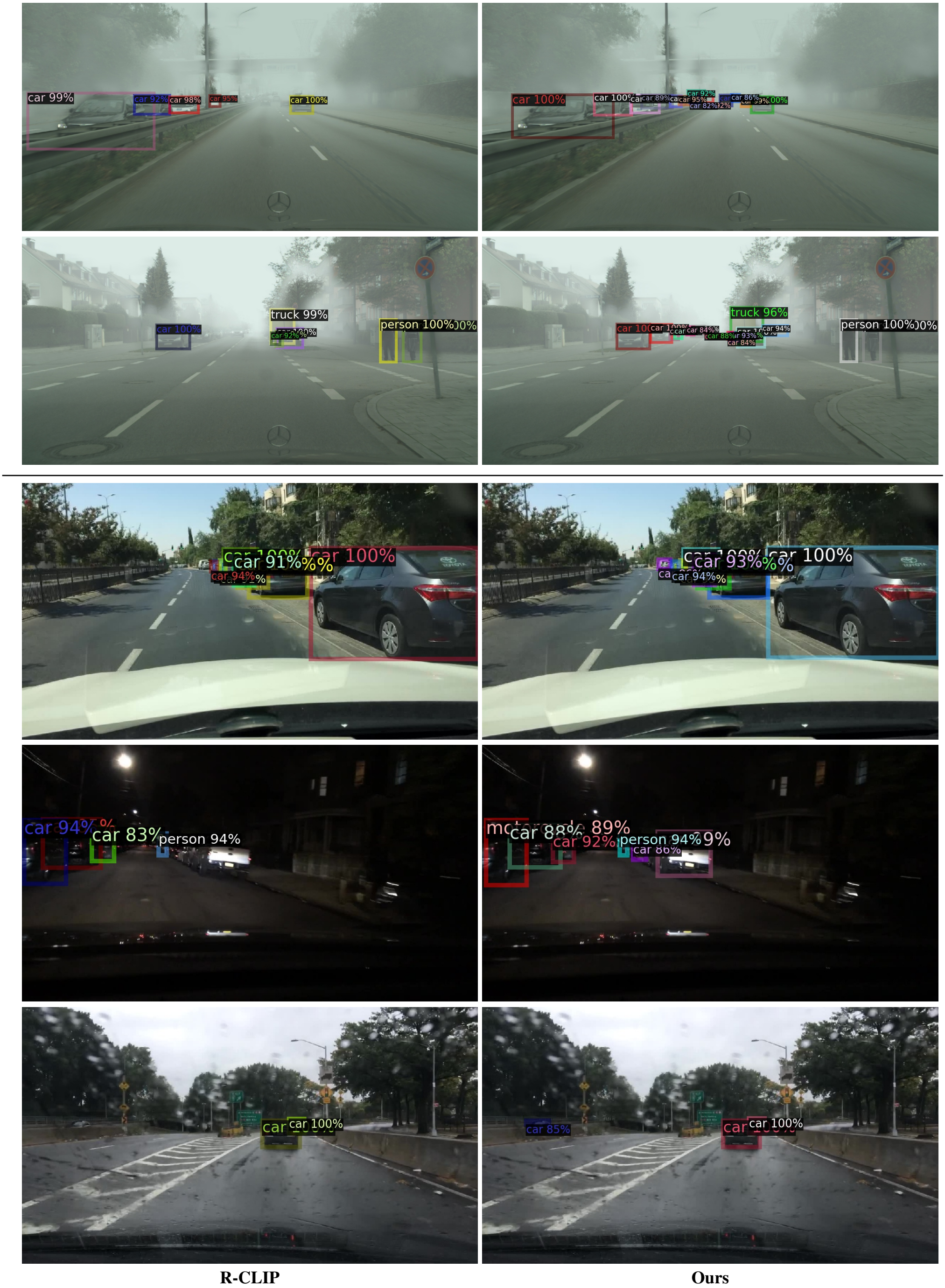
	VOC	Clipart	Watercolor	Comic
RegionCLIP	Two people are standing in a room with a table.	A row of soccer jerseys on a rooftop	A picture of a wall of knit hats	A man is standing in a coffee shop with a group of people watching him.
Ours	A person riding a brown horse on a paved field.	A green and white truck driving down a beach.	A brown and white drawing of a bird.	A man riding a motorcycle on top of a building

Table 2: Visualization inference of our model on real-to-artistic. Each pair of columns corresponds to a different sample with RegionCLIP (left column in each pair) trained on labeled data and Ours (right column in each pair) trained on VOC (labeled) and Clipart (unlabeled) for real-to-artistic generalization.



in Table 1, the RegionCLIP baseline generated an unrelated caption for the image from the VOC dataset that is missing the essential information such as "person" and "horse." This shows the importance of our proposed knowledge distillation-based regularization technique to ensure that captions are meaningful and related to the corresponding image.

Table 3: Visualization inference of our model on adverse-weather domains. RegionCLIP (left column) is trained on labeled data. Ours (right column) is trained on City and Foggy for adverse-weather generalization. The first two rows are results from Foggy, and the last three rows are predictions on the Bdd test set.



References

[1] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11457–11466, 2019.

[2] Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. Cross domain object detection by target-perceived dual branch distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9570–9580, 2022.

[3] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12456–12465, 2019.

[4] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7581–7590, 2022.

[5] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *In IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8771–8780, 2021.

[6] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6956–6965, 2019.

[7] Vibashan VS, Poojan Oza, and Vishal M. Patel. Instance relation graph guided source-free domain adaptive object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3520–3530, June 2023.

[8] 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.

[9] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. Exploring sequence feature alignment for domain adaptive detection transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1730–1738, 2021.

[10] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, YangYang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9603–9612, 2021.

[11] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12355–12364, 2020.

[12] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16793–16803, 2022.