Semi-Supervised Domain Generalization for Object Detection via Language-Guided Feature Alignment

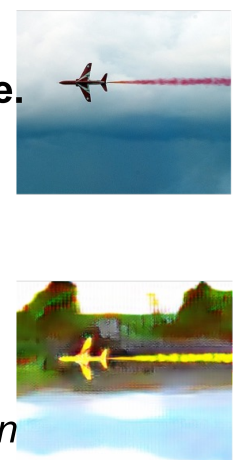
Sina Malakouti and Adriana Kovashka

Department of Computer Science – University of Pittsburgh

Motivation

- Existing domain adaptation (DA) and generalization (DG) methods in object detection enforce **feature alignment in the visual space.**
- But they face challenges, such as **object appearance variability** and **scene complexity**, making distinguishing objects difficult and preventing accurate detection.
- Image descriptions offer **rich semantic** data for object localization and detection.
- **Enforcing consistency in captions** across domains will **enable** model to learn robust representation for **recognizing objects** and **their relations** across domains.

Key idea: *Enforcing **generated image description/captions** to **be consistent across domains** to learn domain robust representation*



Ours A red and white **airplane** flying in the sky.

R-CLIP A red and white **airplane** flying in the sky.

R-CLIP A large red and yellow **highway** with a yellow and yellow **flag**.

Ours A **plane** flying over water.

Overview

- We build a model for **Semi-Supervised Domain Generalization** in Object Detection.
- Capitalizing on the **generalizability** of **vision-language pre-training**, we utilize the RegionCLIP (Zhong et al. 2022) backbone.
- We employ a novel multi-scale contrastive-based consistency objective over generated descriptive features of an image and its stylized version.
- Our approach significantly outperforms baselines in both DA and DG tasks over two benchmarks.

Ours **Two women** smile as they pose for a picture.

R-CLIP **Two women** smile as they pose for a picture.

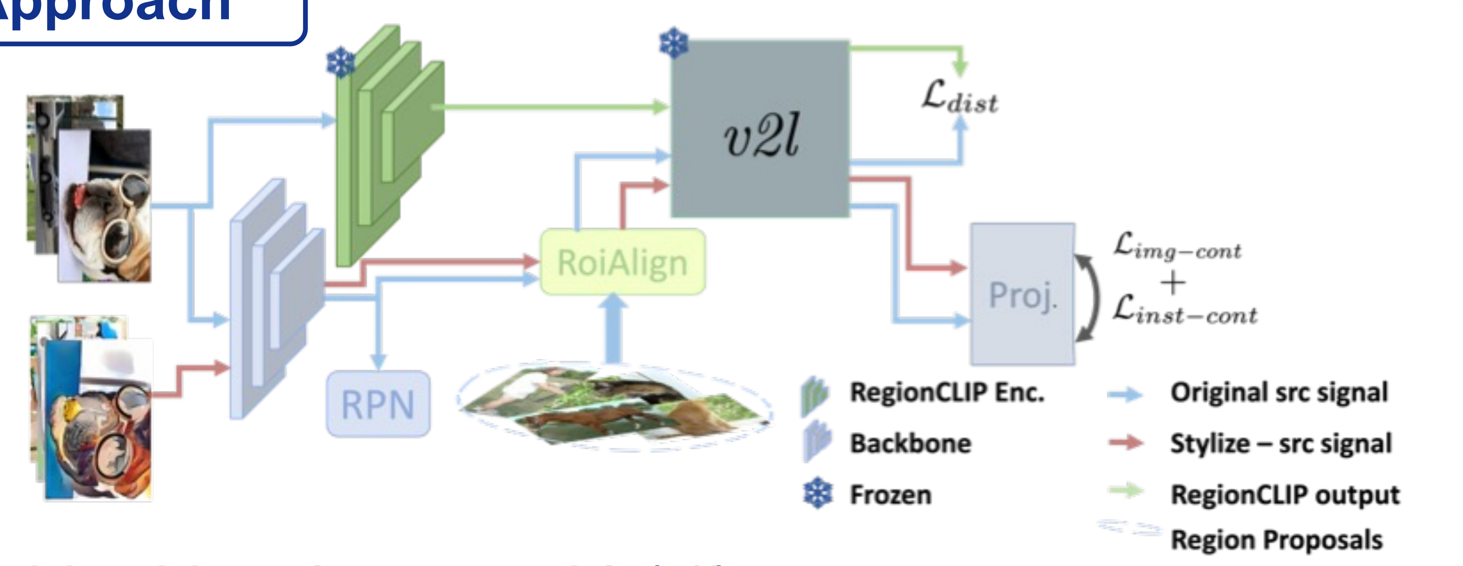
R-CLIP **Two** young **women** with colorful faces and **face masks** standing in front of a **statue**.

Ours **A man and woman** are smiling and posing for a picture.

Related Works

- Our method only **requires one labeled domain**. In contrast, prior DG work (Lin et al. 2021) rely on multiple fully-annotated source domains.
- Prior DA and DG methods in object detection employ different techniques such as pseudo-labeling, data augmentation, mean-teacher framework, etc. by enforcing their objective in the visual domain.

Approach



Pre-training vision-to-language module (*v2l*)

- A transformer-based $v2l$ layer is pre-trained according to ClipCap (Mokady et al. 2021) to project visual features to language space.

Instance-level & Image-level Descriptive Consistency Learning

- Image-level loss is computed by replacing instance features with image-level images

$$\mathcal{L}_{inst-cont} = \frac{1}{N}\sum_i -\log\left(\frac{\exp(s_{i,i}/\tau)}{exp(s_{i,i}/\tau) + \sum_k \exp(s_{i,k}/\tau)}\right); \ k \neq i, \ h_i = g(z_i^{\ell}) \qquad s_{i,j} = s(h_i, \tilde{h}_j) = \frac{h_i^{\top} \cdot \tilde{h}_j}{||h_i|| \cdot ||\tilde{h}_j||}$$

Regularization via Knowledge Distillation (KD)

- A KD regularization is employed to ensure maintaining meaningful representation.

$$\mathcal{L}_{dist} = \frac{1}{N_{\mathcal{L}}}\sum_{i=1}^{N_{\mathcal{L}}} \mathbf{d}(v2l(z_i), v2l(z_i^R)); \quad z_i^R = F_{R-CLIP}(x_i)$$

Object Detector Training

$$\mathcal{L}_{tot} = \mathcal{L}_{det} + \mathcal{L}_{inst-contr} + \mathcal{L}_{img-contr} + \omega \cdot \mathcal{L}_{dist}$$

Experimental Setup

Real-to-Artistic

Domain Generalization

- Source
  - **Labeled:** Pascal-VOC (Everingham et al. 2012,2007)
  - **Unlabeled**: Clipart1k, Watercolor2k, or Comic2k (Inoue et al. 2018)
- Target(s)
  - Clipart1k, Watercolro2k, or Comic2k

Domain Adaptation

- Source
  - Pascal-VOC
- Target(s)
  - Clipart1k, Watercolor2k, and Comic2k

Adverse-Weather

Domain Generalization

- Source
  - **Labeled:** Cityscapes (Cordts et al. 2016)
  - **Unlabeled**: Foggy-Cityscapes (Cordts et al. 2016)
- Target(s)
  - Bdd100k (Yu et al. 2020)

Domain Adaptation

- Source
  - Cityscapes
- Target(s)
  - Foggy-Cityscapes

Baselines

Direct Visual Alignment (DVA)

- Applying Contrastive loss in the visual space

Caption-PL

- Caption Pseudo Labeling

Conclusion

- We developed an approach for Semi-Supervised Domain Generalization in Object Detection.
- We stylized labeled source domain in the unlabeled domain using a style transfer model.
- We leveraged vision-language pretraining by utilizing RegionCLIP.
- We developed a multi-scale contrastive-based approach to ensure consistency of descriptive features in the language latent space.

Quantitative Results

Real-to-Artistic Generalization

Method	VOC&Clip →Water,Com		VOC&Water→Clip,Com		VOC&Com→Clip,Water		Max ↑
	Watercolor	Comic	Clipart	Comic	Clipart	Watercolor	
Faster-RCNN	41.2	17.9	24.1	17.9	24.1	41.2	-
RegionCLIP	44.7	34.2	33.9	34.2	33.9	44.7	16.3/16.3/9.8
Adaptive MT (CVPR'22)	40.6 (-4.1)	22.2 (-12.0)	29.0 (-4.9)	24.3 (-9.9)	25.7 (-8.2)	42.3 (-2.4)	4.3/6.4/1.6
IRG (CVPR'23)	48.1 (+3.4)	25.9 (-8.3)	-	-	-	-	8.0/-/-
DVA	45.6 (+0.9)	38.1 (+3.9)	32.6 (-1.3)	34.2 (+0.0)	35.9 (+2.0)	45.9 (+1.2)	20.2/16.3/11.8
Caption-PL	45.0 (+0.3)	36.4 (+2.2)	30.1 (-3.8)	30.3 (-3.9)	34.7 (+0.8)	42.1 (-2.6)	18.5/12.4/10.6
Ours	**49.8** (+5.1)	**45.9** (+11.7)	**38.7** (+4.8)	**43.5** (+9.3)	**39.8** (+5.9)	**49.4** (+4.7)	**28.0/25.6/15.7**

- Our outperforms baselines on all settings and improves the baseline by up-to 11.7%.
- Our method outperforms DVA and Caption-PL, which shows the effectiveness of **enforcing the consistency objective** in **through the language space** and **the latent space**, respectively.

Real-to-Artistic Adaptation

- Our proposed approach outperforms state-of-the-arts DA methods.
- It also significantly improves the baselines (source-only, DA, and DG).

Method	Target Domain		
	Clipart	Watercolor	Comic
Faster-RCNN	24.1	41.2	17.9
RegionCLIP (CVPR'22)	33.3	44.7	34.2
Adaptive MT (CVPR'22)	30.5	43.7	23.4
IRG (CVPR'23)	31.5	**53.0**	-
DVA	36.6	43.9	35.9
Caption-PL	35.2	44.2	34.2
Ours	**40.4**	49.7	**46.3**

Adverse Weather Generalization

Method	prsn	rider	car	truck	bus	motor	bike	mAP
Faster-RCNN	27.9	27.5	43.1	16.6	15.1	5.6	21.0	19.6
RegionCLIP (CVPR'22)	40.6 (+12.7)	31.3 (+3.8)	47.9 (+4.8)	16.8 (+0.2)	12.0 (-3.1)	11.2 (+5.6)	23.2 (+2.2)	26.1 (+6.5)
DIDN (ICCV'21)	34.5 (+6.6)	30.4 (+2.9)	44.2 (+1.1)	**21.2** (+4.6)	**19.0** (+3.9)	9.2 (+3.6)	22.8 (+1.8)	22.7 (+3.1)
Ours	**41.4** (+13.5)	**31.7** (+4.2)	**49.8** (+6.7)	18.1 (+1.5)	11.4 (-3.7)	**12.4** (+6.8)	**25.6** (+4.6)	**27.1** (+7.5)

- We also show the effectiveness of our method on Cityscapes , Foggy-Cityscapes -> Bdd100k and improve DIDN by 7.5%.

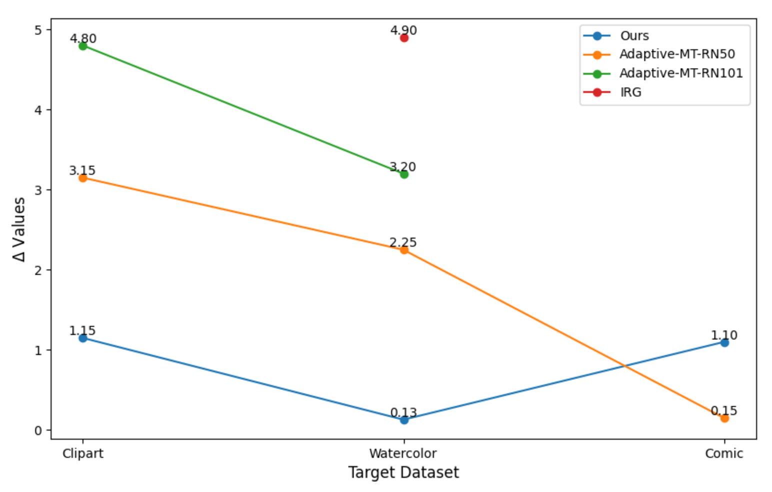
Adverse Weather Adaptation

	Method	prsn	rider	car	truck	bus	train	motor	bike	mAP
-	Faster-RCNN	36.9	36.1	44.5.6	21.7	32.3	9.2	21.5	32.4	28.3 (-20.8)
	RegionCLIP (CVPR'22)	46.5	51.8	57.6	27.3	45.1	19.7	34.8	50.2	**41.6** (-7.5)
DA	SW-DA (CVPR'19)	31.8	44.3	48.9	21.0	43.8	28.0	28.9	35.8	35.3
	D&Match (CVPR'19)	31.8	40.5	51.0	20.9	41.8	34.3	26.6	32.4	34.9
	SC-DA (CVPR'19)	33.8	42.1	52.1	26.8	42.5	26.5	29.2	34.4	35.9
	MTOR (CVPR'19)	30.6	41.4	44.0	21.9	38.6	40.6	28.3	35.6	35.1
	AFAN (TIP'21)	42.5	44.6	57.0	26.4	48.0	28.3	33.2	37.1	39.6
	GPA (CVPR'20)	32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5
	ViSGA (ICCV'21)	38.8	45.9	57.2	29.9	50.2	**51.9**	31.9	40.9	43.3
	SFA (acmmm'21)	46.5	48.6	62.6	25.1	46.2	29.4	28.3	44.0	41.3
	DSS (CVPR'21)	**50.9**	**57.6**	61.1	**35.4**	50.9	36.6	38.4	51.1	47.8
	TTD+FPN (CVPR'22)	50.7	53.7	**68.2**	35.1	53.0	45.1	38.9	49.1	**49.2**
	IRG (CVPR'23)	37.4	45.2	51.9	24.4	39.6	25.2	31.5	41.6	37.1
DG	DIDN (ICCV'21)	38.3	44.4	51.8	28.7	53.3	34.7	32.4	40.4	40.5 (-8.6)
	Ours	50.5	55.1	66.9	35.0	**56.2**	33.5	**41.0**	**54.3**	**49.1**

- We extensively compare against DA methods on Cityscapes -> Foggy-Cityscapes.
- We observe that while our method is not designed to adapt to a specific domain it still outperform most of the DA methods by a large margin.

Ablation & Visualization

Stability Comparison

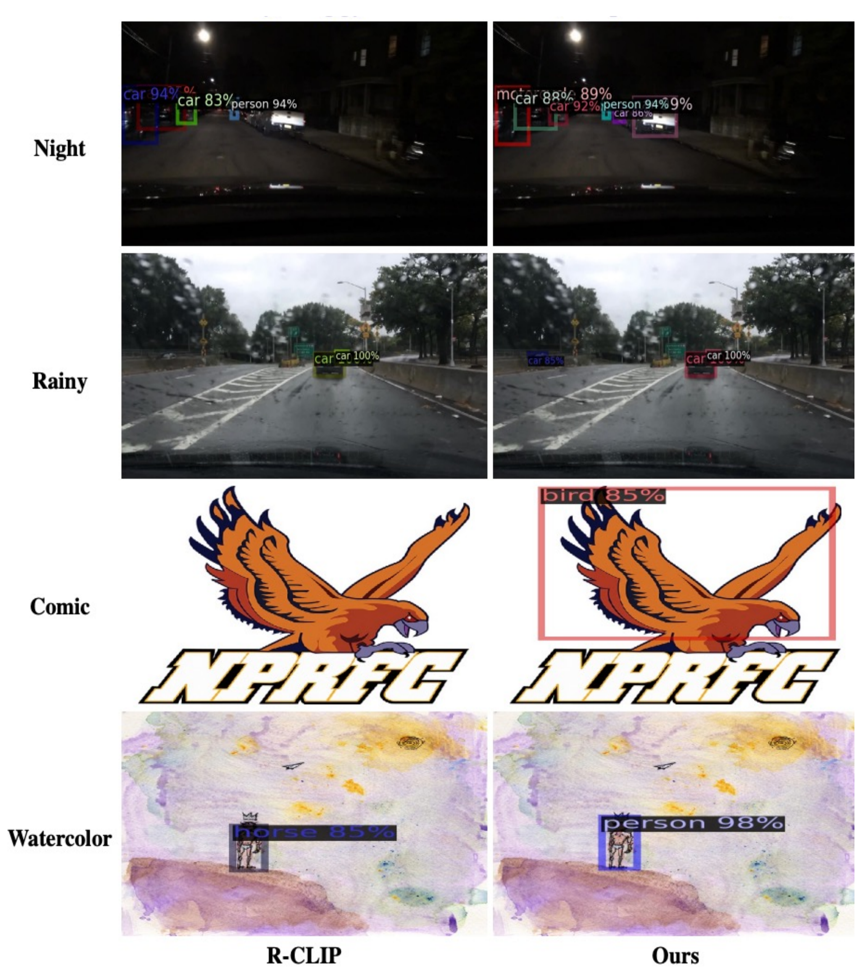


$$\Delta = DA_{mAP}(target) - DG_{mAP}(target)$$

Effectiveness of each component

R-CLIP init.	$\mathcal{L}_{img-cont}$	$\mathcal{L}_{inst-cont}$	$\mathcal{L}_{dist}$	DA Clipart	DG Watercolor	DG Comic
				24.1	41.2	17.9
✓				32.3	44.7	34.2
✓	✓			32.3	41.7	35.1
✓		✓		34.6	45.0	35.4
✓	✓	✓		35.1	44.2	35.7
✓	✓	✓	✓	**40.4**	**49.8**	**45.9**

Visualization



- Vision-Language pre-training **is more robust** compared to ImageNet pre-training.
- Instance-level and Image-level consistency **together achieve the best performance**.
- KD regularization ensures **semantically meaningful features,** resulting in **best performance**.

Contact & Acknowledgement

- For more information, please contact: sem238@pitt.edu, or friend me via LinkedIn.
- https://sinamalakouti.github.io/

- We are grateful for NSF Grant No. 2006885.