CLIP the Gap: A Single Domain Generalization Approach for Object Detection

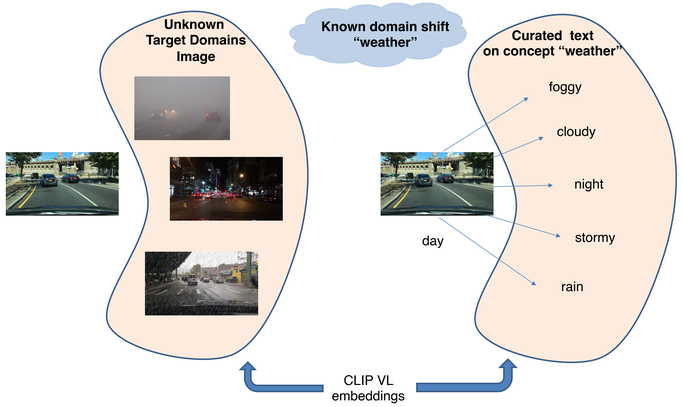
Vidit, Martin Engilberge, Mathieu Salzmann
CVLab, EPFL

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

- SDG aims for generalization to new domains
- One source domain is available
- Domain shifts can be broadly defined via text prompts
- Vision-Language models like CLIP [1] can be helpful

Text Prompts are used to estimate target domain features

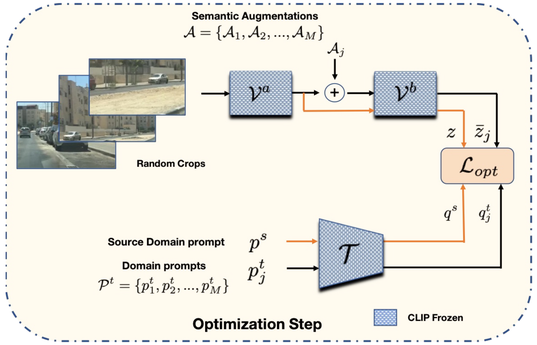


Contribution

- Leveraging vision - text aligned embeddings
- Use textual domain prompts to generate semantic augmentations
- Train-time only augmentation

Method

Step 1.



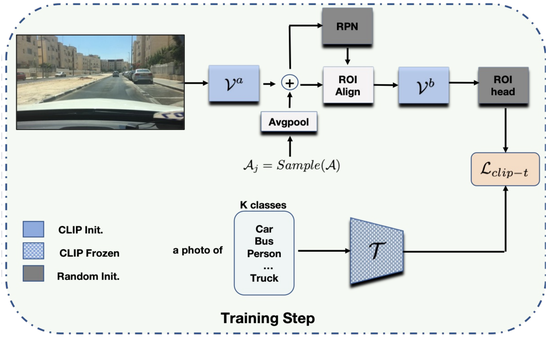
$$z^*_j = z + \frac{q^t_j - q^s}{\|q^t_j - q^s\|_2}$$

$$\mathcal{L}_{opt} = \sum_{\mathcal{I}_{crop}} \sum_{j} \mathcal{D}(z^*_j, \bar{z}_j) + \|\bar{z}_j - z\|_1$$

$\mathcal{D}(z^*_j, \bar{z}_j)$ cosine distance

- We curate a set of textual prompts $\mathcal{P}^t$ which can represent new domains
- For ***weather*** dataset, we can create prompts $\mathcal{P}^t$ as
  `an image taken on a {weather} {time of the day.}`
- Source domain prompt p^s for the same is
  `an image taken during the day.`

Step 2.

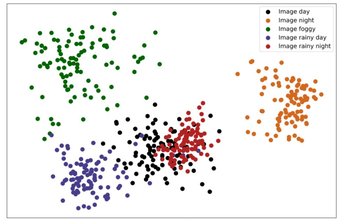


$$\mathcal{L}_{clip-t} = \sum_{r} \mathcal{L}_{CE}\left(\frac{e^{\gamma \cdot sim(\mathcal{F}_r, \mathcal{Q}_k)}}{\sum_{k=0}^{K} e^{\gamma \cdot sim(\mathcal{F}_r, \mathcal{Q}_k)}}\right)$$

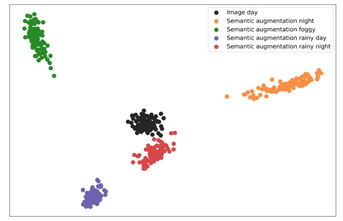
- Train by sampling augmentations
- Inference without augmentation

Results

Original image features



Features obtained by optimization



Adverse Weather Dataset [2]
Source Domain: Day

Method	mAP				
	Day Clear	Night Clear	Dusk Rainy	Night Rainy	Day Foggy
S-DGOD [2]	**56.1**	36.6	28.2	16.6	33.5
Ours	51.3	**36.9**	**32.3**	**18.7**	**38.5**

Cross domain dataset [3]
Source Domain: Pascal VOC[4]

Method	mAP	
	Comic	Watercolor
CLIP-init	26.2	41.9
Ours w/o seg-aug	32.8	42.7
Ours	**33.5**	**43.4**

Ablations

Model Component				mAP				
				Source		Target		
CLIP init	$\mathcal{L}_{clip-t}$	Attn. Pool	Sem. Aug	Day Clear	Night Clear	Dusk Rainy	Night Rainy	Day Foggy
				48.1	34.4	26.0	12.4	32.0
✓				51.2	**37.0**	31.0	15.7	37.5
✓	✓			50.7	36.0	31.3	16.3	36.9
✓	✓	✓		51.0	35.9	31.3	16.7	37.7
✓	✓	✓	✓	51.3	36.9	**32.3**	**18.7**	**38.5**

Aug. Type	mAP				
	Day Clear	Night Clear	Dusk Rainy	Night Rainy	Day Foggy
no-aug.	51.0	35.9	31.3	16.7	37.7
random	51.2	36.0	30.4	15.3	37.3
clip-random	51.5	36.4	30.2	15.9	37.9
Ours w/ seg.aug	51.3	**36.9**	**32.3**	**18.7**	**38.5**

Conclusion

- Textual description of underlying domain shift can be helpful
- Vision-Text embeddings helps in augmenting missing target image features

References:

[1] Alec Radford et al. Learning transferable visual models from natural language supervision, ICML'21
[2] Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self distillation, CVPR'22
[3] Naoto Inoue,et al. Cross-domain weakly-supervised object detection through progressive domain adaptation, CVPR'22
[4] M. Everingham, et al. The PASCAL Visual Object Classes Challenge 2012 (VOC2012), ECCV'12