Improving Object Detection via Local-global Contrastive Learning

Danai Triantafyllidou [1*], Sarah Parisot[1]
Ales Leonardis[2], Steven McDonagh[3]

Huawei Noah's Ark Lab[1] University of Birmingham[2] University of Edinburgh[3]

danaitri22@gmail.com, s.mcdonagh@ed.ac.uk

BMVC 2024

Domain Adaptive Object Detection

Problem:

Detection models fail to generalise to new domains with visually distinct images

Solution:

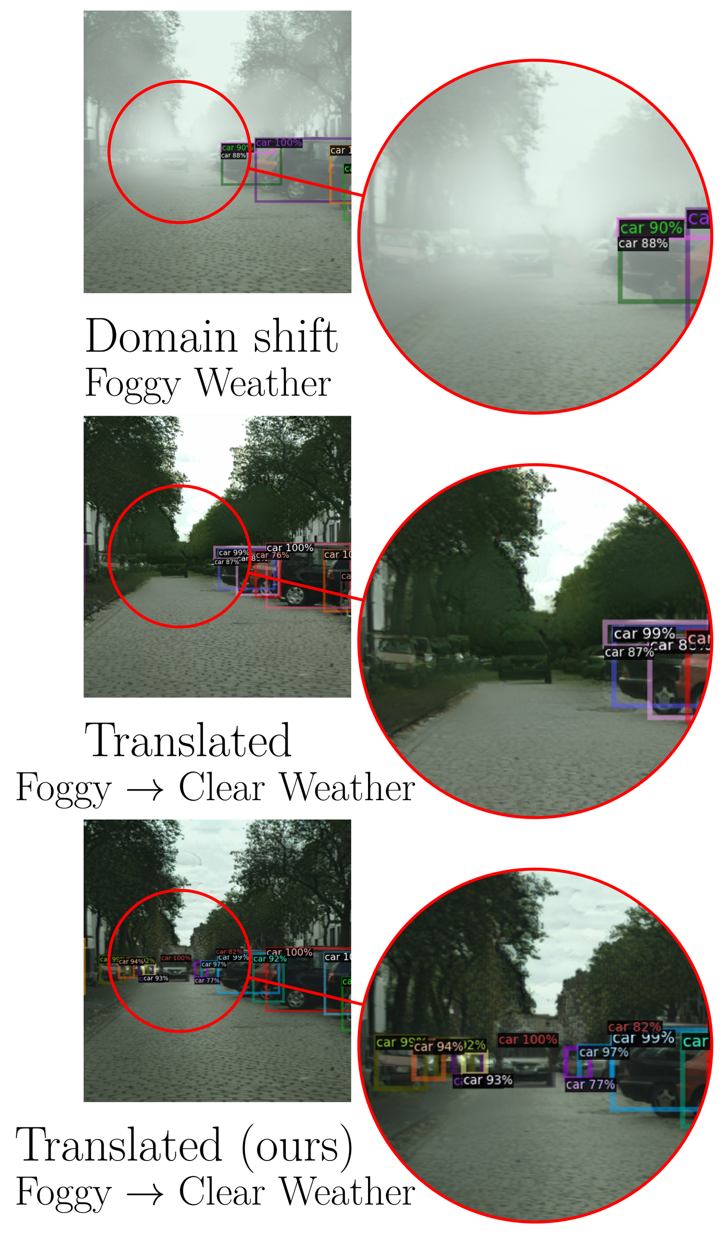
Image-to-image translation → mitigate domain shift at the input level

Challenge:

Global style-translation treats all image regions uniformly, leading to:

- ☐ loss of local structures & object details
- ☐ semantically inconsistent textures

Detection performance suffers as a result



- Prior works leverage object annotations to process object regions separately
- Annotations are expensive and often infeasible to obtain

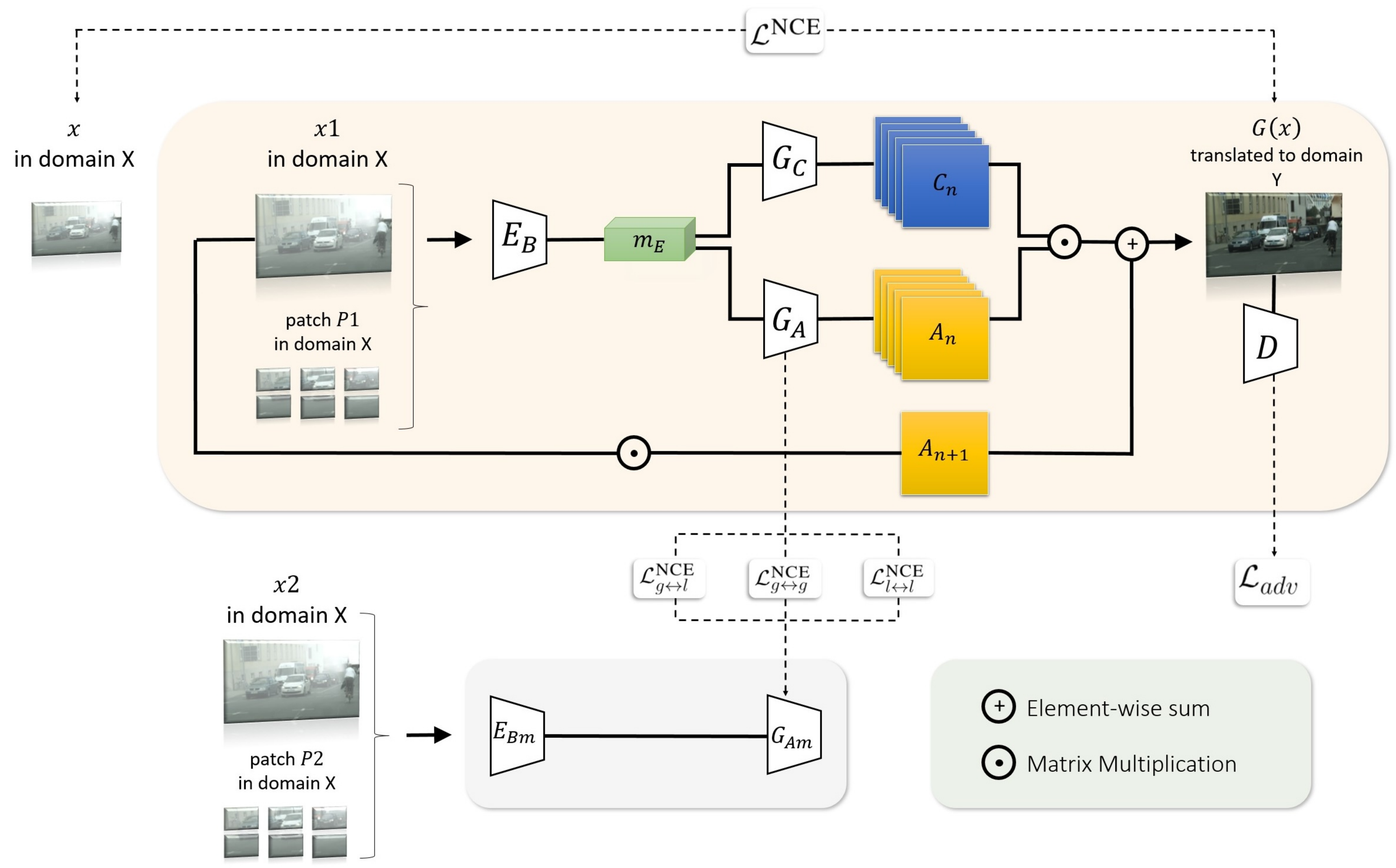
Hypothesis

- Spatial attention can enhance translation quality in local regions
- Content delineation can be facilitated through *local-global* contrastive learning

Contributions

1. Novel I2I translation framework for cross-domain object detection
2. An inductive prior that optimises object appearance through spatial attention maps
3. Leverage local-global contrastive learning to learn discriminative represenations
4. State-of-the-art performance on three visual domain adaptation scenarios; assuming a pre-trained *frozen* detector model

Method



- Detector source domain $\{\mathbf{y}_i\}_{i=1}^N$, target domain $\{\mathbf{x}_i\}_{i=1}^N$
- Learn a mapping $f:\mathcal{X}\rightarrow\mathcal{Y}$ to alleviate visual domain shift and improve detection performance

Spatial Attention

- Encoder-decoder model implicitly separates semantic content into foreground and background regions through spatial attention maps
- Decompose decoder as G_C and G_A, producing a set of n content maps $\{C_t \,|\, t\in[0,n-1]\}$ and a set of $n+1$ attention maps $\{A_t \,|\, t\in[0,n]\}$
- Recover the translated output as $G(\mathbf{x}) = \Sigma_{t=1}^{n}\underbrace{(C^t\odot A^t)}_{\text{foreground}} + \underbrace{(x\odot A^{n+1})}_{\text{background}}$

Optimization

$$\mathcal{L}_{\text{TOTAL}} = \underbrace{\mathcal{L}_{adv}}_{\substack{\text{appearance}\\ \text{transfer}}} + \underbrace{\mathcal{L}^{NCE}}_{\substack{\text{structure}\\ \text{preservation}}} + \underbrace{\mathcal{L}^{G_A}}_{\substack{\text{local-global}\\ \text{attention}\\ \text{guidance}}}$$

$$\mathcal{L}^{NCE} = -\log\frac{\exp(q\cdot k\,/\,\tau)}{\exp(q\cdot k\,/\,\tau)+\Sigma_{k_-}\exp(q\cdot k\,/\,\tau)}$$

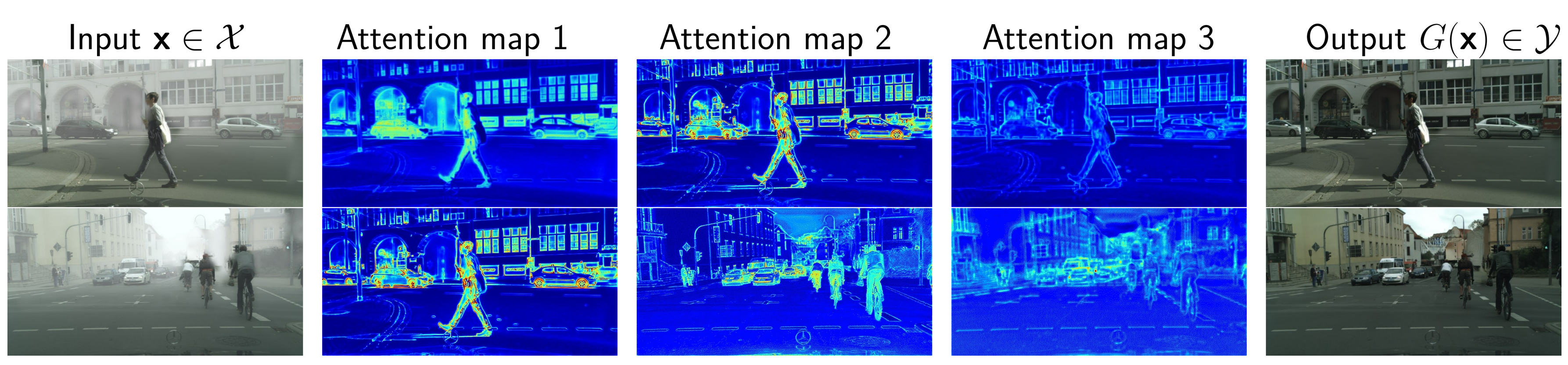
- $\mathcal{L}_{adv}$ adversarial term – translated images match appearance of domain $\mathcal{Y}$
- $\mathcal{L}^{NCE}$ patchwise infoNCE loss maximizing mutual information between input and translated patches – drives structural preservation

Local-global contrastive learning

- Guide the attention generator G_A by contrasting local-global representations; *alleviating the need for object annotations*
- Multi-level supervision directly optimising G_A features

$$\mathcal{L}_{G_A} = \sum_{i=1}^{L} w_i\mathcal{L}^{NCE}_{g\leftrightarrow g} + \sum_{i=1}^{L} w_i\mathcal{L}^{NCE}_{g\leftrightarrow l} + \sum_{i=1}^{L} w_i\mathcal{L}^{NCE}_{l\leftrightarrow l}$$

- ☐ $g\leftrightarrow g$ loss term between *global* representations of $\mathbf{x}$
- ☐ $g\leftrightarrow l$, $l\leftrightarrow l$ terms considering *local-to-global* and *local-to-local* representations of $\mathbf{x}$
- ☐ for network layers L; layer contribution weights w_i

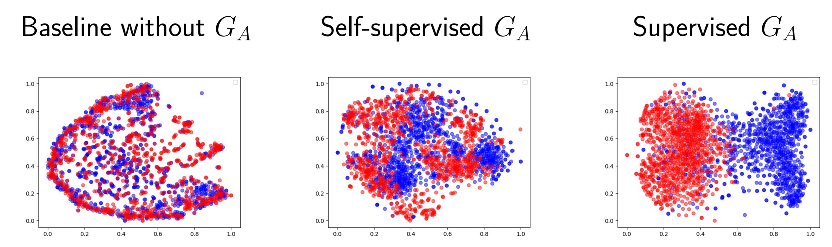


Local-global self-supervision accentuates semantic object regions and improves translation in areas critical for object detection

Ablative study

Det. backbone	G_A	$\mathcal{L}_{G_A}$	Supervision	Attention	mAP @ 0.5
			-		42.7
	✓		-	✓	44.4
Res-50	✓	✓	local–global $\mathcal{L}_{G_A}$	✓	45.3

Method components improve detection performance



t-SNE visuliazation of G_A features; we randomly sample features corresponding to object regions (red) and background regions (blue)

Quantitative results

Foggy cityscapes → Cityscapes [10]

Method	person	rider	car	truck	bus	train	motor	bike	mAP ↑
FGRR [5]	34.4	47.6	51.3	30.0	46.8	42.3	35.1	38.9	40.8
DAF+NLTE [36]	37.0	46.9	54.8	32.1	49.9	43.5	29.9	39.6	41.8
TIA [77]	34.8	46.3	49.7	31.1	52.1	48.6	37.7	38.1	42.3
SCAN [30]	41.7	43.9	57.3	28.7	48.6	48.7	31.0	37.3	42.1
SIGMA [31]	46.9	48.4	63.7	27.1	50.7	35.9	34.7	41.4	43.5
SDA [47]	38.8	45.9	57.2	29.9	50.2	51.9	31.9	40.9	43.3
MGA [79]	43.9	49.6	60.6	29.6	50.7	39.0	38.3	42.8	44.3
DA-DETR [74]	49.9	50.0	63.1	24.0	45.8	37.5	31.6	46.3	43.5
memCLR [60]	37.7	42.8	52.4	24.5	40.6	31.7	29.4	42.2	37.7
MIC [17]	52.4	47.5	67.0	40.6	50.9	55.3	33.7	33.9	**47.6**
CDAT [4]	42.3	51.7	64.0	26.0	42.7	37.1	42.5	44.0	43.8
Ours - supervised ($\mathcal{L}_{G_{A_{sup}}}$)	44.4	49.5	61.4	32.6	50.8	52.2	38.3	44.0	<u>46.7</u>
CUT*† [43]	39.6	45.3	59.4	27.9	47.4	45.4	35.3	39.2	42.4
FeSeSim*† [78]	40.9	47.2	58.4	28.4	48.6	49.8	34.3	42.7	43.8
Qs-Att.*† [19]	42.2	49.0	60.3	23.5	50.5	52.0	36.6	41.4	44.4
NEGCUT*† [63]	42.2	48.2	58.8	27.9	47.8	50.2	34.9	43.7	44.2
Hneg_SCR*† [25]	42.8	46.9	59.7	32.3	48.4	48.9	36.8	43.4	44.9
Santa*† [63]	42.3	47.9	59.4	34.4	49.3	49.1	36.4	42.3	<u>45.1</u>
Source	35.5	38.7	41.5	18.4	32.8	12.5	22.3	33.6	29.4
Target Oracle	47.5	51.7	66.9	39.4	56.8	49.0	43.2	47.3	50.2
Ours - local-global † ($\mathcal{L}_{G_A}$)	43.2	50.1	61.7	33.3	48.6	47.8	35.2	42.6	**45.3**

Methods without access to object annotations during training denoted †. See paper for corresponding references and further details

Adaptation scenarios

Adverse → Clear weather

Foggy Cityscapes → Cityscapes [10]

Synthetic-to-real

Sim10k [23] → Cityscapes [10]

Real-world cross-camera

KITTI [12] → Cityscapes [10]

Links

Contact:

danaitri22@gmail.com
s.mcdonagh@ed.ac.uk

Paper: **Project Page:**

* Currently with Kittl: https://www.kittl.com/