Improving Object Detection via Local-global Contrastive Learning

Danai Triantafyllidou[1*,†]
danaitri22@gmail.com
Sarah Parisot[1]
sarah.parisot@huawei.com
Ales Leonardis[2,†]
a.leonardis@cs.bham.ac.uk
Steven McDonagh[3,†]
s.mcdonagh@ed.ac.uk

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

Abstract

Visual domain gaps often impact object detection performance. Image-to-image translation can mitigate this effect, where contrastive approaches enable learning of the image-to-image mapping under unsupervised regimes. However, existing methods often fail to handle content-rich scenes with multiple object instances, which manifests in unsatisfactory detection performance. Sensitivity to such instance-level content is typically only gained through object annotations, which can be expensive to obtain. Towards addressing this issue, we present a novel image-to-image translation method that specifically targets cross-domain object detection. We formulate our approach as a contrastive learning framework with an inductive prior that optimises the appearance of object instances through spatial attention masks, implicitly delineating the scene into foreground regions associated with the target object instances and background non-object regions. Instead of relying on object annotations to explicitly account for object instances during translation, our approach learns to represent objects by contrasting local-global information. This affords investigation of an under-explored challenge: obtaining performant detection, under domain shifts, without relying on object annotations nor detector model fine-tuning. We experiment with multiple cross-domain object detection settings across three challenging benchmarks and report state-of-the-art performance.

Project page: https://local-global-detection.github.io

1 Introduction

Deep learning based object detection has become an indispensable part of many computer vision applications such as autonomous navigation. State-of-the-art detection models typically rely on large-scale annotated data in order to learn representative features and yet often fail to generalize well to new target domains that exhibit visual disparity, (such as foggy vs. clear weather scenes), with common benchmarks typically reporting falls in detection accuracy in

† Work partially done at Huawei Noah's Ark Lab

∗ Currently at Kittl Technologies

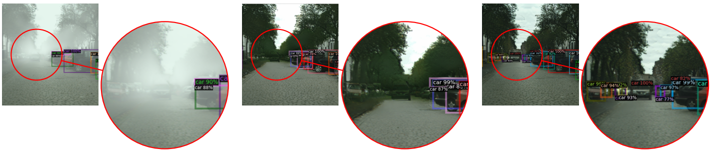


Figure 1: **Left:** visual domains, unseen during object detector training, hurt detection performance. **Middle:** global image-to-image translation (foggy → clear weather) provides some benefit to downstream detection performance, yet homogeneous image translation strategies result in small objects, with low contrast regions, that remain undetectable. **Right:** Our local-global approach is guided to better delineate objects during translation and thus cross-domain detection is improved.

excess of 20% (see Sec. 4.2 for details). Performance drops drastically due to the domain shift problem. Image-to-image (I2I) translation aims to mitigate such domain gaps at the input level and thereby reduce the distribution shift in the visual domain. Such approaches enable an existing (i.e. pre-trained) detector, trained on the source domain, to function well on *source-like* images, translated from the target domain. It has been evidenced that this process is able to improve target domain detection performance [5, 41, 63].

The high costs related to (domain-wise) paired image data collection have steered community interest towards unpaired I2I translation. Pioneering unpaired image translation work has made use of cycle-consistency [71] and shared latent space assumptions [32]. Such methods have become the de facto translation modules in many works. However, they can often lead to severe content distortions and shape deformation as they assume a bijective relationship between source and target domains [36]. Such failures to ensure content preservation may in turn adversely affect performance in downstream object detection tasks [35], where this effect is exemplified in Fig. 1.

Explicitly accounting for object instances has provided an intuitive direction for improving image translation in spatial regions that are critical to down-stream detection [2, 20, 46, 47]. However, these works rely on object annotations in order to treat object and background spatial image regions distinctly in the target domain, which fundamentally limits their applicability. In cases where strong pre-trained detectors exist, yet object labels are inaccessible or are otherwise infeasible to obtain, such strategies become unsuitable. We offer a new perspective and alternatively consider the scenario where labels are unavailable, an under-explored and yet practical problem setting, where we further propose a method to account for this gap in the literature.

Contrastive learning has emerged as a promising strategy for solving I2I translation, through the maximisation of mutual information between corresponding input and output patches [18, 37, 69]. While recent contrastive-based translation results report promising increases for standard image quality metrics [15, 39], these approaches consider image translation as a *global* task; i.e. a translation problem where all image regions are treated uniformly. Framing translation in this manner can lead to unsatisfactory outcomes when considering object-rich images with complex local structures; the visual disparity between objects and background is often large. We hypothesize that implicitly modeling background and foreground object regions can enhance translation quality in local salient areas, significantly improving downstream object detection. Additionally, we propose that the separation of foreground and background can be accomplished through local-global contrastive learning.

Motivated by this intuition, we propose a contrastive learning-based I2I translation framework for cross-domain object detection. We introduce an architectural inductive prior that optimises object instance appearance using spatial attention masks, effectively disentangling scenes into background and foreground regions. We note that while previous studies [66] have employed the “background / foreground” terminology to describe explicit content separation based on object annotations, our approach learns to delineate content under unsupervised conditions[1]. Inspired by the recent success of region-based representation learning [56, 59, 62], we alternatively rely on contrasting local and global views to learn discriminative representations and separate content. Our main contributions can be summarised as:

- We propose a novel attention guided I2I translation framework for cross-domain object detection. Our approach encourages the model to optimise local image region appearance without requiring object annotations and can be used in conjunction with a frozen pre-trained object detector.
- We illustrate how the idea of local-global contrastive learning can be used to improve image-to-image translation for object detection: implicitly differentiating between objects and background image regions gives rise to robust translation of object image regions, amenable for detection tasks.
- We conduct extensive experiments in common domain adaptation and detection settings, reporting state-of-the-art performance under three visual adaptation scenarios.

2 Related work

We briefly review topics most relevant to our core ideas and refer to [35, 36] for extensive surveys on both image-to-image translation and domain adaptation for object detectors.

Instance-aware I2I translation. Instance-aware I2I translation has recently garnered interest, towards enabling models to translate objects and background areas separately. Shen et al. [46, 47] perform translation with distinct encoder-decoder blocks to generate separate object, background and global image style codes and provide the model with object-specific guidance in relation to translation. Following the idea of distinct instance region encodings, Bhattacharjee et al. [2] propose to jointly learn image translation and detection, therefore focussing on certain objects during translation. A class-aware memory network was used in [20] to store features and retain individual object styles, thus improving translation for images with multiple objects. The recent work of [25] performs instance-aware I2I using a transformer model, trained with contrastive learning. All of these works crucially assume access to object annotations during training in order to guide the translation.

Trainable cross-domain object detectors. In contrast to I2I, cross-domain detection solutions integrate Unsupervised Domain Adaptation (UDA) techniques, within object detection pipelines. An extensive set of cross-domain object detector training strategies exist; adversarial feature learning for domain invariant representations, pseudo-labels for self-training, graph reasoning and domain randomization, among others [5, 6, 10, 17, 24, 28, 29, 30, 31, 40, 41, 43, 46, 48, 54, 57, 58, 61, 68]. This tranche of works fundamentally involve training or otherwise adapting a detector model. However, adaptation strategies typically assume that source domain object labels remain available and such methods are also known to suffer from catastrophic forgetting problems [53]. We consider direct comparison with this

[1] Despite not using object annotations, our empirical results demonstrate effective separation of content, leading us to adopt this terminology.

class of methods under common experimental settings to offer useful insight, with respect to investigation of method efficacy and related trade-offs (further details are found in Sec. 4).

Contrastive learning. Contemporary self-supervised learning aims to exploit the underlying structures in the data and build unsupervised visual representations; either by solving generative pretext tasks (e.g. colourisation, inpainting, jigsaw puzzles) or through contrastive learning. Contrastive learning has shown great potential when performing instance discrimination tasks [7, 8, 73], where the objective is formed by generating different views of an image and maximising their similarity through data augmentation. The success of such methods is mainly attributed to the ability of contrastive learning to encode semantic priors across different images [38]. More recently, there has been increased interest in region-based representation learning, shifting the focus to learning local descriptors that are relevant for dense prediction tasks such as image segmentation and object detection. Indeed, global-local and multi-scale crop strategies have proven popular in self-supervised and unsupervised (contrastive) learning scenarios where prevalent works include BYOL [13] and DINO [64]. Learning region-level representations has been realised through image segmentation masks [14, 51, 67] and by contrasting between local patches and global image views [59]. These recent successes lead contrastive learning to become a prevailing component of self-supervised learning and particularly successful in pre-training a strong feature extractor for several local and global discriminative tasks.

Park et al. [37] employed a contrastive approach to the I2I translation task and enforce their model to preserve structure in corresponding input and output spatial locations. Zheng et al. [69] further improve the structure consistency constraint by contrasting self-similarity patches. Huet et al. [18] proposed an entropy based query selection mechanism, towards enabling feature selection that better reflects domain specific characteristics. Jung et al. [22] enable semantic awareness in a contrastive setting through the exploitation of semantic relation consistencies across image patches. However we note that these methods lack in-built mechanisms to exploit *instance-level* information, specifically relating to semantic objects in a scene. We foresee this as a potential shortcoming for downstream detection tasks and experimentally evidence this conjecture (Sec. 4.2).

3 Method

3.1 Preliminaries

Self-supervised representations are realised under a contrastive learning regime by considering a dictionary look-up task. Specifically, given an encoded query q, the task is to identify which single positive key k_+ matches the query q among a set of encoded keys $\{k_0, k_1, ...\}$. The InfoNCE loss function [34] is employed to attract q close to k_+ while pushing it away from a set of alternative negative keys $\{k_-\}$:

$$\mathcal{L}^{\text{NCE}} = -\log \frac{\exp(q \cdot k_+ / \tau)}{\exp(q \cdot k_+ / \tau) + \sum_{k_-} \exp(q \cdot k_- / \tau)}, \tag{1}$$

with τ a temperature hyperparameter. For vision tasks, positive pairs q and k_+ can be formed by generating two different views from the same image or different views that pertain to a global image and a local patch [59].

Contrastive learning for unpaired I2I translation. Contrastive techniques can be leveraged for the I2I translation task by constraining matching spatial locations (image patches) between the input image and translated output image to have high mutual information. In this case, the query patch q is created by encoding a local region of the output image. The positive key k_+ refers to the corresponding region of the input, while the set of negative keys $\{k_-\}$ are selected by encoding different regions of the input image. As such, the contrastive loss of Eq. (1) ensures content and structure consistency between input and translated patches, while the appearance of the output image is enforced using a discriminator, trained with an adversarial loss [12].

3.2 Spatial attention for I2I translation

We assume a performant pre-trained object detector to be available, trained using images from a source domain Y, with the aim of applying the detector to a new target *detection* domain X. Our goal is to learn a function capable of performing the (inverse) image translation task $\mathcal{G}: X \rightarrow Y$ such that detection performance is significantly improved for images originally belonging to domain X. In contrast to previous works that extract separate representations to encode global and instance-level information, respectively [2, 25, 47], our approach alternatively guides translation to focus on the relevant instance regions, using spatial attention masks. The spatial attention masks are generated by a dedicated trainable module and weight the influence of separate image features in a final translation step.

We propose an attention-driven scheme that learns to decompose input image x into foreground and background regions and encourages the translation model to focus on optimising appearance of foreground objects. We adopt an encoder-decoder architecture where the encoder E_B acts as a feature extractor, and generates image representations of lower dimensionality. We decompose our decoder into two components: a content generator G_C that generates multiple image content maps and an attention generator G_A that outputs attention masks. Attention masks enable combination of the generated content maps in a learnable fashion to obtain a final translated image. Fig. 2 depicts an overview of the proposed method.

More formally, input image x is first converted into a latent representation via feature extractor E_B: $m_E = E_B(x)$. This representation serves as input to the content and attention generators. The content generator G_C generates a set of n content maps $\{C_t \,|\, t \in [0, n-1]\}$. Each layer l of G_C comprises a group of n convolutional filters, such that each filter is associated with a specific content map. Content map t at layer l can be expressed as: $C_l^t = \sigma(\texttt{Conv}^t(C_{l-1}^t))$ with $t = 0, \ldots, n-1$ and $C_0^t = m_E$, $\forall t$. The activation function $\sigma(\cdot)$ is selected as a $\texttt{ReLU}$ [1] in the intermediate layers, and $\sigma(\cdot) = \tanh(\cdot)$ in the final layer. Similarly, the attention generator outputs a set of $n+1$ attention maps $\{A_t \,|\, t \in [0, n]\}$, using $n+1$ convolutional filters per layer, with $\sigma(\cdot) = \texttt{softmax}(\cdot)$ for the last layer. Finally, the translated image $G(x)$ is recovered through summation of the generated foreground content maps, weighted by the respective attention masks, with an additional explicit component, representing image background content:

$$G(x) = \sum_{t=1}^{n} \underbrace{(C^t \odot A^t)}_{\text{foreground}} + \underbrace{(x \odot A^{n+1})}_{\text{background}}. \tag{2}$$

By disentangling the translation task in this manner and explicitly modelling distinct spatial regions, we enable our model to actively focus on discriminative image locations containing objects. We find that accurate and precise translation of such image regions to be of crucial importance for strong detection performance (see Sec. 4.2 for further details).

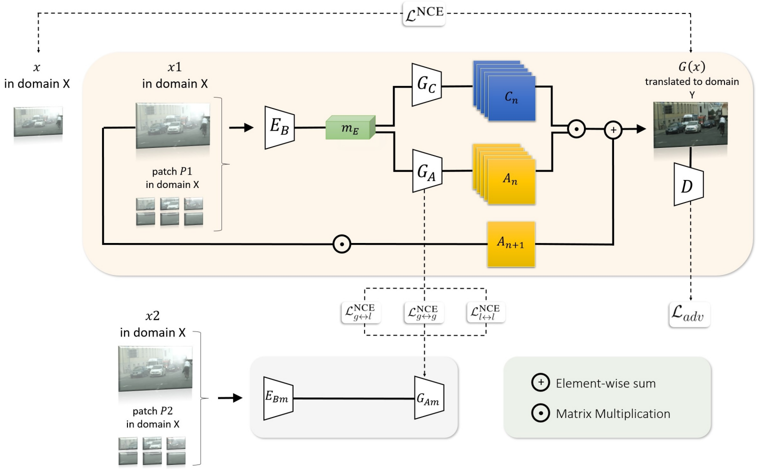


Figure 2: Overview of the proposed method - see text for further details.

We follow previous contrastive I2I work [37] and train our generator using the InfoNCE loss found in Eq. (1), which enforces content and structure consistency at the image patch level. To ensure that translated images match the appearance of source domain Y, we further use a discriminator module D trained with a standard adversarial loss:

$$\mathcal{L}_{adv} = -\mathbb{E}_{y\sim Y} \log D(y) - \mathbb{E}_{x\sim X} \log(1 - D(G(x))). \tag{3}$$

The discriminator then minimises the negative log-likelihood for a standard binary classification task and this is equivalent to minimising the Jensen-Shannon (JS) divergence between the model output distribution and real source domain distribution Y.

Our scheme, thus far, has equipped the generator model with the ability to treat different image regions non-uniformly by decomposing the image into foreground and background content. In order to further guide the translation task to attend to image regions containing semantically meaningful content, we introduce an additional loss on the attention generator $\mathcal{L}_{G_A}$, which exploits the relationship between local and global image patches. The full optimisation objective can then be expressed as follows:

$$\mathcal{L}_G = \mathcal{L}_{adv} + \mathcal{L}^{\mathrm{NCE}} + \mathcal{L}_{G_A}. \tag{4}$$

3.3 Local-global contrastive learning

The contrastive loss of Eq. (1) maximizes mutual information between corresponding input and output patches, ensuring structural consistency between the input and the translated image. We further encourage our model to consistently encode local information, of benefit to object detection tasks, and global information representing content that spans beyond objects to uncountable amorphous regions of similar texture. The goal is to learn representations that result in globally consistent translations while also remaining sensitive to accurate representation of local details and structure. To this end, we design an objective that (i) encourages local representations of an image to be closer to the global representation of the same image, (ii) encourages local representations of an image to be close to one another, (iii) pulls the global representations of an image's distinct augmentations close to one another. This provides the model with an ability to discriminate local representations that describe different content while encouraging patch representations of a common scene to cluster in the latent space. Similar local, global strategies have previously proved successful in standard unsupervised object detection settings [59] and here we alternatively explore the potential benefits when translating content-rich scenes in a cross-domain object detection setup.

We decompose our input image x into 16 non-overlapping patches: $P(x) = \{x^p \mid p \in [1,16]\}$. We apply two different random augmentations to x, yielding transformed images x_1 and x_2 and correspondingly; two sets of transformed patches P_1 and P_2. To obtain local and global representations for contrastive learning, we attach two projection heads to our attention generator G_A, one assigned for local patches: MLP_{local} and one assigned to the full global image: MLP_{global}. Following a common protocol, we additionally introduce momentum copies of E_B and G_A, denoted G_{Am}, E_{Bm}, where their weights are updated using an exponential moving average. Forwarding images x_1, x_2 and patch sets P_1 and P_2 through encoders E_B, E_{Bm} and then decoders G_A and G_{Am}, generates two sets of global and local feature representations $\{f_A x_1, f_A x_2, f_A x^p_1, f_A x^p_2\}$ and $\{f_{Am} x_1, f_{Am} x_2, f_{Am} x^p_1, f_{Am} x^p_2\}$, pertaining to the outputs of G_A and G_{Am}, respectively. We then compute $\mathcal{L}^{NCE}$ between all pairs in these feature sets to optimise model discriminative power, with negative pairs drawn from a memory bank.

Finally, we define multi-scale supervision to improve the model's ability to identify salient regions. We introduce additional local and global MLP layers at the output of each layer in G_A and compute the infoNCE loss for each new set of features. As a result, our unsupervised loss for G_A can be expressed as follows:

$$\mathcal{L}_{G_A} = \sum_{i=1}^{L} w_i \mathcal{L}^{NCE}_{g\leftrightarrow g} + \sum_{i=1}^{L} w_i \mathcal{L}^{NCE}_{g\leftrightarrow l} + \sum_{i=1}^{L} w_i \mathcal{L}^{NCE}_{l\leftrightarrow l}, \tag{5}$$

where L is the number of layers in G_A, and w_i is a weight parameter controlling the importance of each layer contribution. The first term in our objective defined in Eq. (5), denoted $g \leftrightarrow g$, computes the loss between *global* representations, while objective terms $g \leftrightarrow l$ and $l \leftrightarrow l$ indicate that the infoNCE loss is considering *local to global* and *local to local* representations, respectively. We clarify the number of hyperparameters introduced in Eq. (5) as follows: the three component loss terms, $\mathcal{L}^{NCE}_{g\leftrightarrow g}, \mathcal{L}^{NCE}_{g\leftrightarrow l}$ and $\mathcal{L}^{NCE}_{l\leftrightarrow l}$, are computed using multi-stage features from individual network layers. We use L= 4, leading to a total of four weights w_i, for each of the three loss terms. In practice, the weighting for each stage i is common across the three losses, resulting in only four additional distinct hyperparameters in total. We follow convention [59] and apply smaller weights to shallow layers and larger weights to deeper layers.

By attaching the aforementioned loss to the features of the G_A module, we conjecture that we are able to encourage the attention generator to develop an enhanced sensitivity to semantic content and attend to translation regions of importance for the object detection task. For completeness, we additionally consider a scenario where annotation labels are available and replace the local-global contrastive loss of Eq. (5) with a supervised object saliency loss. We add a simple auxiliary object saliency task that uses a binary object mask $k(x)$, explicitly separating all foreground objects from the background, as a target ground truth. We introduce a new convolutional layer $\texttt{Conv}^S$ in G_A that receives the predicted attention maps before softmax and outputs a binary object mask prediction $m(x)$. The resulting supervised loss can be defined as:

$$\mathcal{L}_{G_{A_{sup}}} = -\mathbb{E}_{x\sim X}\left[k(x)\log m(x) + (1-k(x))\log(1-m(x))\right]. \tag{6}$$

Method	D.A.	I2I	Backbone	person	rider	car	truck	bus	train	motor	bike	mAP ↑
FGRR [1] *(TPAMI'23)*	✓		Vgg-16	34.4	47.6	51.3	30.0	46.8	42.3	35.1	38.9	40.8
DAF+NLTE [31] *(CVPR '22)*	✓		Res-50	37.0	46.9	54.8	32.1	49.9	43.5	29.9	39.6	41.8
TIA [58] *(CVPR '22)*	✓	✓	Res-50	34.8	46.3	49.7	31.1	52.1	48.6	37.7	38.1	42.3
SCAN [27] *(AAAI '22)*	✓		Vgg-16	41.7	43.9	57.3	28.7	48.6	48.7	31.0	37.3	42.1
SIGMA [28] *(CVPR '22)*	✓		Res-50	46.9	48.4	63.7	27.1	50.7	35.9	34.7	41.4	43.5
SDA [40] *(CVPR '21)*	✓		Res-50	38.8	45.9	57.2	29.9	50.2	51.9	31.9	40.9	43.3
MGA [70] *(CVPR '22)*	✓		Vgg-16	43.9	49.6	60.6	29.6	50.7	39.0	38.3	42.8	44.3
DA-DETR [53] *(CVPR '23)*	✓		Res-50	49.9	50.0	63.1	24.0	45.8	37.5	31.6	46.3	43.5
memCLR [52] *(WACV'23)*	✓		Vgg-16	37.7	42.8	52.4	24.5	40.6	31.7	29.4	42.2	37.7
MIC [16] *(CVPR 23)*	✓		Vgg-16	52.4	47.5	67.0	40.6	50.9	55.3	33.7	33.9	**47.6**
CDAT [3] *(CVPR 23)*	✓		Vgg-16	42.3	51.7	64.0	26.0	42.7	37.1	42.5	44.0	43.8
Ours - supervised ($\mathcal{L}_{G_{A_{sup}}}$)		✓	Res-50	44.4	49.5	61.4	32.6	50.8	52.2	38.3	44.0	<u>46.7</u>
CUT*† [37] *(ECCV '20)*		✓	Res-50	39.6	45.3	59.4	27.9	47.4	45.4	35.3	39.2	42.4
FeSeSim*† [59] *(CVPR '21)*		✓	Res-50	40.9	47.2	58.4	28.4	48.6	49.8	34.3	42.7	43.8
Qs-Att.*† [18] *(CVPR '22)*		✓	Res-50	42.2	49.0	60.3	23.5	50.5	52.0	36.6	41.4	44.4
NEGCUT*† [55] *(CVPR '21)*		✓	Res-50	42.2	48.2	58.8	27.9	47.8	50.2	34.9	43.7	44.2
Hneg_SCR*† [23] *(CVPR '22)*		✓	Res-50	42.8	46.9	59.7	32.3	48.4	48.9	36.8	43.4	44.9
Santa*† [50] *(CVPR '23)*		✓	Res-50	42.3	47.9	59.4	34.4	49.3	49.1	36.4	42.3	<u>45.1</u>
Source			Res-50	35.5	38.7	41.5	18.4	32.8	12.5	22.3	33.6	29.4
Target Oracle			Res-50	47.5	51.7	66.9	39.4	56.8	49.0	43.2	47.3	50.2
Ours - local-global † ($\mathcal{L}_{G_A}$)		✓	Res-50	43.2	50.1	61.7	33.3	48.6	47.8	35.2	42.6	**45.3**

Table 1: The Foggy Cityscapes → Cityscapes adaptation scenario. We report object detection (mAP) per class. Previous works utilises detector adaptation (D.A), image-to-image translation (I2I) components. Locally reproduced methods using publicly available codes are indicated by *. We separate methods that **do** (upper) and **do not** (lower) have access to object annotations at training time, with the latter methods denoted †. Results are denoted **best** and <u>second best</u> for upper and lower table sections.

Method	D.A.	I2I	car	person	mAP ↑
DARL [26] *(CVPR '19)*	✓	✓	58.7	46.4	52.5
DAOD [illegible] *(BMVC '19)*	✓	✓	59.1	47.3	62.9
DUNIT [illegible] *(CVPR '20)*	✓	✓	65.1	<u>60.7</u>	62.9
MGUIT [19] *(CVPR '21)*		✓	68.2	58.3	63.2
InstaFormer [25] *(CVPR '22)*		✓	<u>69.5</u>	**61.8**	**65.6**
DA-DETR [53] *(CVPR '22)*		✓	48.9	-	-
Source			63.4	55.0	59.2
Target Oracle			77.4	66.3	71.8
Ours - supervised ($\mathcal{L}_{G_{A_{sup}}}$)		✓	**71.0**	59.5	<u>65.2</u>
Ours - local-global † ($\mathcal{L}_{G_A}$)		✓	67.5	<u>60.7</u>	64.1

Method	D.A.	I2I	Backbone	AP_{car} ↑
HTCN [illegible] *(CVPR '20)*	✓	✓	Vgg-16	42.5
UMT [illegible] *(CVPR '21)*	✓	✓	Vgg-16	43.1
FGRR [1] *(PAMI '23)*	✓		Vgg-16	44.5
DSS [illegible] *(CVPR '21)*	✓		Res-50	44.5
SWDA [illegible] *(CVPR '19)*	✓		Res-50	44.6
SCDA [illegible] *(CVPR '19)*	✓		Res-50	45.1
AFAN [illegible] *(TIP '21)*	✓	✓	Res-50	45.5
GPA [illegible] ($CVPR'20$)	✓		Res-50	47.6
SDA [illegible] *(CVPR '21)*	✓		Vgg-16	49.3
MGA [illegible] *(CVPR '22)*	✓		Vgg-16	49.8
KTNet [19] *(ICCV '21)*	✓		Vgg-16	50.7
SSAL [33] *(NeurIPS '21)*	✓		Vgg-16	51.8
SCAN [27] *(AAAI '22)*	✓		Vgg-16	52.6
SIGMA [28] *(CVPR '22)*	✓		Vgg-16	53.4
DA-DETR [53] *(CVPR '23)*	✓		Vgg-16	**54.7**
Source			Res-50	41.7
Ours - supervised ($\mathcal{L}_{G_{A_{sup}}}$)		✓	Res-50	<u>53.6</u>
Ours - local-global † ($\mathcal{L}_{G_A}$)		✓	Res-50	52.1

Table 2: Adaptation results for KITTI → Cityscapes (left) and Sim10K → Cityscapes (right).

4 Experiments

4.1 Datasets

Foggy Cityscapes → Cityscapes. Cityscapes [9] was collected by capturing images from outdoor urban street scenes, containing 2,975 images for training and 500 images for testing with eight annotated object categories, namely: *person, rider, car, truck, bus, train, motorcycle and bicycle*. Foggy Cityscapes [45], analogously, is a synthetic foggy dataset rendered using Cityscapes, using aligned depth information to simulate synthetic fog on the original clear weather scenes. We firstly evaluate our method under this adversarial weather scenario.

KITTI → Cityscapes. KITTI [11] is a widely used autonomous driving dataset containing videos of traffic scenarios recorded with different sensors. The dataset consists of 7,481 training images and 7,518 test images, with a total of 80,256 annotated objects which span eight different categories: *car, van, truck, pedestrian, person sitting, cyclist, tram, misc*. In this challenging real-to-real translation scenario, we study cross-camera adaptation by

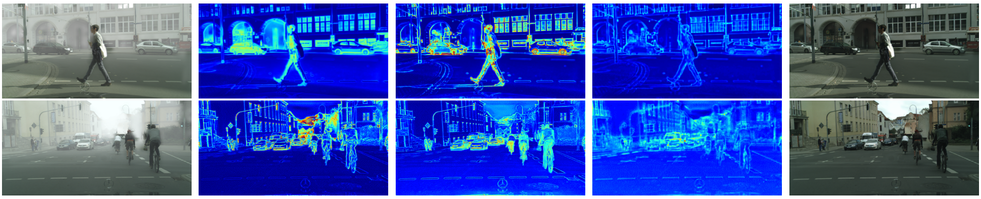


Figure 3: Col 1: input images. Cols $2-4$: learned foreground attention masks. Local-global self-supervision accentuates semantic object content regions and improves translation in areas critical for object detection (e.g. people, cars). Col 5: translated output images.

performing translation from Cityscapes [9] imagery and evaluate on classes *car* and *person*.

Sim10*k* → Cityscapes. Sim10*k* [21] is a simulated dataset generated using the GTA-V game engine. It consists of $10,000$ images of synthetic driving scenes and $58,701$ annotated object instances. We perform domain adaptation between the synthesized imagery and the real-world images of the Cityscapes [9] dataset. Here we evaluate our proposed approach by considering detection performance using the *car* class, following a common protocol.

4.2 Comparison with State-of-the-Art

Foggy Cityscapes → Cityscapes. We report object detection results under our initial adaptation scenario in Tab. 1. We present detection performance in terms of per-class average precision (AP) and mean average precision (mAP). With respect to the subset of methods that do not have annotations, our self-supervised local-global configuration achieves state-of-the-art performance of 45.3% mAP with the supervised counterpart offering additional further improvement of 46.7%. We additionally probe framework efficacy by replacing our specific I2I translation model with three alternative state-of-the-art I2I approaches, whilst keeping the object detector component fixed. Namely we consider I2I approaches CUT [37], FeSeSim [69] and Qs-Attn [18] (see Tab. 1, lower). Our translation model can show detection gains c.f. these recent I2I translation models, in each case. We attribute improvements to our local-global framework, capable of accurate object region translation. We provide visualisations of the learned attention masks in Fig. 3. These intend to highlight the ability to delineate semantically meaningful regions and attend to relevant discriminative areas in local regions. Specifically, we observe that generated attention masks focus on semantic regions that contain objects, enhancing the discriminative ability of the model.

KITTI → Cityscapes. In Tab. 2 (left) we report results for this adaptation scenario in comparison with several instance-aware translation methods. Following [2], we present the per-class (AP) in addition to the mAP for classes *car* and *person*. Our supervised model achieves 65.2% mAP, while our self-supervised local-global strategy again exhibits competitive performance. Our local-global approach stands out as the only method in Tab. 2 that does not rely on object annotations during training for this challenging real-to-real scenario.

Sim10k → Cityscapes. In Tab. 2 (right) we report detection results for a further adaptation scenario. In this setting our approach is able to achieve 52.1% and 53.6% AP_{50} in self-supervised and supervised settings, respectively. Compared with recent strategies that specifically employ an image translation module and use an identical detector backbone (namely AFAN [54]) our supervised model achieves gains of over 8%. Our self-supervised variant trained with local-global contrastive learning can also achieve performance competitive with detector adaptation based methods, yet without access to object label information.

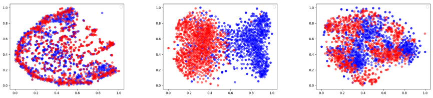
4.3 Ablative study

We examine the impact of the proposed components in detail and report results in Tab. 3. We select the Foggy Cityscapes → Cityscapes adaptation scenario and train the proposed model architecture under the following ablations: (i) *without* the G_A network and *without* the proposed attention module; (ii) *with* the G_A network, *with* the proposed attention module and *without* loss $\mathcal{L}_{G_A}$; (iii) *with* the G_A network, *with* the proposed attention module and *with* an unsupervised $\mathcal{L}_{G_A}$ loss; and finally (iv) *with* the G_A network, *with* the proposed attention module and *with* a supervised $\mathcal{L}_{G_{A_{sup}}}$ loss. All models are trained under identical settings which are reported in our supplementary materials. In all cases we use the adversarial loss $\mathcal{L}_{adv}$ found in Eq. (3) and the InfoNCE loss found in Eq. (1). We observe that detection performance is lower in case (i) where all method components under consideration are absent. In case (ii) performance is improved by 0.5–1.7% mAP@.5 which we attribute to the addition of the proposed attention module, trained without any guidance (i.e. $\mathcal{L}_{G_A} = 0$). Inclusion of all components results in 2.3–2.6% mAP@.5 gains for the unsupervised model (case (iii)) and 2.3–4% mAP@.5 gains for the supervised model (case (iv)). We provide further ablative comparisons in our supp. materials.

Det. backbone	G_A	$\mathcal{L}_{G_A}$	Supervision	Attention	mAP@[.5:.95]	mAP@.5	mAP@.75	mAP@[.5:.95]	mAP@[.5:.95]	mAP@[.5:.95]
								small	medium	large
Res-50			-		23.0	42.7	21.8	2.2	20.8	47.4
	✓		-	✓	23.5	44.4	20.8	2.5	22.3	46.3
	✓	✓	local–global $\mathcal{L}_{G_A}$	✓	24.1	45.3	23.2	2.6	23.3	47.1
	✓	✓	supervised $\mathcal{L}_{G_{A_{sup}}}$	✓	24.5	46.7	22.9	2.7	23.4	47.1

Table 3: Ablation on method components (Foggy Cityscapes → Cityscapes).

We additionally present a feature-level visualization via t-SNE [50] in Fig. 4, towards evidencing method effectiveness in terms of identifying relevant salient object regions. We visualise two classes, defined using object bounding box labels, as {object, background} and randomly sample 1000 feature points. We observe that in case (a) the learned representations do not afford discriminability between these two classes. Adding the supervised object saliency signal in case (b) results in a clearly separable learned feature embedding. Finally, in case (c) we evidence that our self-supervised local-global model, using Eq. (5), can enhance separability c.f. case (a), which concurs with our empirical observations that manifest as object, background disentanglement behaviour.



(a) Baseline without G_A (b) supervised $G_{A_{sup}}$ (c) self-supervised G_A

Figure 4: t-SNE feature visuliazation; we randomly sample object features corresponding to salient objects (red) and image background regions (blue).

5 Conclusion

We propose a novel approach for cross-domain object detection using unpaired image-to-image translation. Our contrastive-learning based attention mechanism endows the model with object awareness and steers feature representations to be discriminative in terms of benefit to downstream detection tasks, post image translation between source and target domains. We explore generation of attention masks in fully unsupervised regimes and evidence competitive detection results in comparison with numerous state-of-the-art methods, whilst requiring neither domain-paired image data nor access to object labels.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *CoRR*, abs/1803.08375, 2018.

[2] Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, and Mathieu Salzmann. Dunit: Detection-based unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4787–4796, 2020.

[3] Shengcao Cao, Dhiraj Joshi, Liangyan Gui, and Yu-Xiong Wang. Contrastive mean teacher for domain adaptive object detectors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23839–23848, 2023.

[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.

[5] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2020.

[6] Chaoqi Chen, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. I3net: Implicit instance-invariant network for adapting one-stage object detectors. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.

[8] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. In *ArXiv*, volume abs/2003.04297, 2020.

[9] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[11] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[14] Olivier J. Henaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and Joao Carreira. Efficient visual pretraining with contrastive detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, volume 30, 2017.

[16] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11721–11732, 2022.

[17] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *European Conference on Computer Vision*, pages 733–748. Springer, 2020.

[18] Xueqi Hu, Xinyue Zhou, Qiusheng Huang, Zhengyi Shi, Li Sun, and Qingli Li. Qs-attn: Query-selected attention for contrastive learning in I2I translation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[19] S. Jeong, Y. Kim, E. Lee, and K. Sohn. Memory-guided unsupervised image-to-image translation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6554–6563, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society.

[20] Somi Jeong, Youngjung Kim, Eungbean Lee, and Kwanghoon Sohn. Memory-guided unsupervised image-to-image translation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[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? In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[22] Chanyong Jung, Gihyun Kwon, and Jong Chul Ye. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[23] Chanyong Jung, Gihyun Kwon, and Jong-Chul Ye. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18239–18248, 2022.

[24] 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*, 2019.

[25] Soohyun Kim, Jongbeom Baek, Jihye Park, Gyeongnyeon Kim, and Seungryong Kim. Instaformer: Instance-aware image-to-image translation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18321–18331, 2022.

[26] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12448–12457, 2019.

[27] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan. SCAN: cross domain object detection with semantic conditioned adaptation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 1421–1428. AAAI Press, 2022.

[28] 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.

[29] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–18, 2023.

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

[31] Xinyu Liu, Wuyang Li, Qiushi Yang, Baopu Li, and Yixuan Yuan. Towards robust adaptive object detection under noisy annotations. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[32] T. Breuel M.-Y. Liu and J. Kautz. Unsupervised image-to-image translation networks. In *31st International Conference on Neural Information Processing Systems*, 2017.

[33] Muhammad Akhtar Munir, M. H. Khan, M. Saquib Sarfraz, and Mohsen Ali. Synergizing between self-training and adversarial learning for domain adaptive object detection. In *ArXiv*, volume abs/2110.00249, 2021.

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, Jan 2019.

[35] Poojan Oza, Vishwanath A Sindagi, Vibashan VS, and Vishal M Patel. Unsupervised domain adaptation of object detectors: A survey. In *arXiv preprint arXiv:2105.13502*, 2021.

[36] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. In *IEEE Transactions on Multimedia*, volume 24, pages 3859–3881. IEEE, 2021.

[37] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020.

[38] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. In *Advances in Neural Information Processing Systems*, volume 33, pages 3407–3418, 2020.

[39] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pages 435–446. Springer, 2012.

[40] Farzaneh Rezaeianaran, Rakshith Shetty, Rahaf Aljundi, Daniel Olmeda Reino, Shanshan Zhang, and Bernt Schiele. Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9204–9213, 2021.

[41] A. L. Rodriguez and K. Mikolajczyk. Domain adaptation for object detection via style consistency. In *British Machine Vision Conference*, 2019.

[42] Adrian Lopez Rodriguez and Krystian Mikolajczyk. Domain adaptation for object detection via style consistency. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 232. BMVA Press, 2019.

[43] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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

[45] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. In *International booktitle of Computer Vision*, volume 126, page 973–992, 2018.

[46] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Zechun Liu, Harsh Maheshwari, Yutong Zheng, Xiangyang Xue, Marios Savvides, and Thomas S Huang. Cdtd: A large-scale cross-domain benchmark for instance-level image-to-image translation and domain adaptive object detection. In *International booktitle of Computer Vision*, pages 761–780, 2021.

[47] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas S Huang. Towards instance-level image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3683–3692, 2019.

[48] Vishwanath A. Sindagi, Poojan Oza, Rajeev Yasarla, and Vishal M. Patel. Prior-based domain adaptive object detection for hazy and rainy conditions. In *European Conference on Computer Vision*, page 763–780, 2020.

[49] Kun Tian, Chenghao Zhang, Ying Wang, Shiming Xiang, and Chunhong Pan. Knowledge mining and transferring for domain adaptive object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9113–9122, 2021.

[50] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. In *booktitle of Machine Learning Research*, volume 9, pages 2579–2605, 11 2008.

[51] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[52] VS Vibashan, Poojan Oza, and Vishal M. Patel. Towards online domain adaptive object detection. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 478–488, 2022.

[53] Riccardo Volpi, Diane Larlus, and Grégory Rogez. Continual adaptation of visual representations via domain randomization and meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4443–4453, 2021.

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

[55] Weilun Wang, Wen gang Zhou, Jianmin Bao, Dong Chen, and Houqiang Li. Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14000–14009, 2021.

[56] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[57] 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.

[58] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2021.

[59] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[60] Shaoan Xie, Yanwu Xu, Mingming Gong, and Kun Zhang. Unpaired image-to-image translation with shortest path regularization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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

[62] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[63] Dan Zhang, Jingjing Li, Lin Xiong, Lan Lin, Mao Ye, and Shangming Yang. Cycle-consistent domain adaptive faster rcnn. In *IEEE Access*, volume 7, pages 123903–123911. IEEE, 2019.

[64] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[65] Jingyi Zhang, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, Xiaoqin Zhang, and Shijian Lu. Da-detr: Domain adaptive detection transformer with information fusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23787–23798, 2021.

[66] Liyun Zhang, Photchara Ratsamee, Bowen Wang, Zhaojie Luo, Yuki Uranishi, Manabu Higashida, and Haruo Takemura. Panoptic-aware image-to-image translation. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[67] Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS 2020, Red Hook, NY, USA, 2020. Curran Associates Inc.

[68] Liang Zhao and Limin Wang. Task-specific inconsistency alignment for domain adaptive object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[69] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. The spatially-correlative loss for various image translation tasks. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[70] Wenzhang Zhou, Dawei Du, Libo Zhang, Tiejian Luo, and Yanjun Wu. Multi-granularity alignment domain adaptation for object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9571–9580, 2022.

[71] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[72] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[73] Chengxu Zhuang, Alex Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6001–6011, 2019.