To miss-attend is to misalign! Residual Self-Attentive Feature Alignment for Adapting Object Detectors

Vaishnavi Khindkar[1] Chetan Arora[2] Vineeth N Balasubramanian[3]
Anbumani Subramanian[1] Rohit Saluja[1] C.V. Jawahar[1]
[1]CVIT - IIIT Hyderabad, India [2]IIT Delhi, India [3]IIT Hyderabad, India
https://github.com/Vaishnvi/ILLUME
[1]vkhindkar@gmail.com, [2]chetan@cse.iitd.ac.in, [3]vineethnb@iith.ac.in,
[1]{anbumani@, rohit.saluja@research., jawahar@}iiit.ac.in

Abstract

*Advancements in adaptive object detection can lead to tremendous improvements in applications like autonomous navigation, as they alleviate the distributional shifts along the detection pipeline. Prior works adopt adversarial learning to align image features at global and local levels, yet the instance-specific misalignment persists. Also, adaptive object detection remains challenging due to visual diversity in background scenes and intricate combinations of objects. Motivated by structural importance, we aim to attend prominent instance-specific regions, overcoming the feature misalignment issue. We propose a novel resIduaL seLf-attentive featUre alignMEnt (*ILLUME*) method for adaptive object detection.* ILLUME *comprises Self-Attention Feature Map (SAFM) module that enhances structural attention to object-related regions and thereby generates domain invariant features. Our approach significantly reduces the domain distance with the improved feature alignment of the instances. Qualitative results demonstrate the ability of* ILLUME *to attend important object instances required for alignment. Experimental results on several benchmark datasets show that our method outperforms the existing state-of-the-art approaches.*

1. Introduction

Object detection has shown significant improvement in the deep learning era. Mostly the two-stage detectors such as Faster R-CNN [30] are widely used in domain adaptation. However, they usually rely on a large amount of training data, which requires task-specific annotation efforts and is also cost-intensive. Also, they work well only on label-rich domains with no domain gaps and often fail to generalize well in universal settings due to the dataset bias.

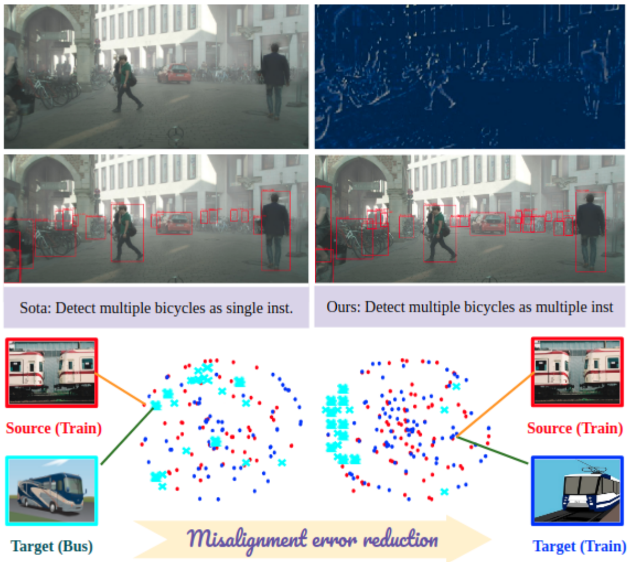

Sota: Detect multiple bicycles as single inst.
Ours: Detect multiple bicycles as multiple inst
Source (Train)
Target (Bus)
Misalignment error reduction
Source (Train)
Target (Train)

Figure 1: **Top:** Visualization of the self-attention feature map (right) for Foggy cityscapes [33] target sample (left) enhancing the prominent structural regions - instances like person/car. **Middle:** State-of-the-art (left) unsuccessfully detects multiple bicycles as a single instance compared to ours (right) which correctly detects multiple bicycles as different instances. **Bottom:** Instance-level feature visualizations using tsne. State-of-the-art (left) misalign source and target instance features, compared to our method (right). Misalignment error and domain distance are significantly reduced using our approach which improves detection performance. Refer section 4.4 for further details.

Unsupervised Domain Adaptation (UDA) attempts to handle the problem of dataset bias for different tasks. Recently UDA methods have been widely used in multiple tasks that involve deep learning. UDA also offers an appealing solution by adapting object detectors from label-rich source domains to unlabeled target domains. Among the widely used UDA methods [48, 59], adversarial training has shown significant results. The feature extractor is trained to generate domain invariant features, thereby deceiving the domain classifier, which tries to focus on correctly classifying source and target domains. In the literature, adversarial training has been well-studied for domain adaptive image classification, semantic segmentation, and object detection. Complex localization due to intricate combinations of objects and multiple scales, makes detection a challenging task as compared to others like classification. Of many domain adaptive detection methods, [3] is a representative work that trains Faster R-CNN in an adversarial manner to be domain adaptive. To address the domain shift problem, the adaptive detector aligns image and instance distributions across domains with adversarial training. Recently, the domain adaptive Faster R-CNN has rapidly evolved into successful methods [21, 40, 2, 32, 48]. We note that object detection also focuses on local regions that may contain objects of interest. Subsequently, although instance-level alignment can match object proposals in both domains, current practices do not address the problem of misaligned label spaces effectively, resulting in low detection performance.

To overcome this, we propose an adaptive object detection method based on self-attention. Our method improves instance-specific feature alignment by significantly reducing the misalignment error. The learning process is inspired by the structural importance of memorizing an object irrespective of domain/background differences. Aligning feature distributions of prominent object-related regions to train the Faster R-CNN detector helps it generalize well.

We design a novel architecture that comprises a residual self-attentive feature alignment module to generate i) self-attention feature maps, and ii) a pixel-wise domain classifier output. These components are combined to generate domain invariant attention feature maps. Finally, these attention feature maps are combined with initial feature maps of the image in a residual manner, giving instance-specific attentive feature maps as shown in Fig. 1 (top). With enhanced structural attention and improved feature alignment, our model detects multiple instances correctly. As shown in Fig. 1 (middle), state-of-the-art model [21] detects multiple instances of the bicycle as a single object, in contrast, our model correctly detects them as different objects. To summarize, following are the main contributions of our work:

- We propose a novel ILLUME method to effectively enhance the instance feature alignment. We are the first to investigate *self-attention* based feature alignment for *detection* task in domain adaptation.
- We propose a simple yet effective SAFM module that focuses on attending the instances that are necessary for adaptation with prominence to object-related structural regions, without any need of regularisation.
- We tackle the instance-specific misalignment issue with the incorporation of multi-stage residual self-attentive alignment that significantly reduces domain distance between source and target instance features.
- The proposed method is evaluated on several benchmark datasets and outperforms recent domain adaptive detection approaches. Additionally, we conduct detailed ablation studies to disambiguate the role of SAFM for achieving improved detection performance.

2. Related Work

Object detection. Object detection problems are widely studied in computer vision. In the era of deep learning and CNNs' success, object detection can be categorized into two classes: one-stage detectors [29, 24] and two-stage ones [30]. Although one-stage detectors have high efficiency and have become popular, two-stage detectors are widely adopted for pursuing much higher performance. In particular, Faster R-CNN [30] is a classical two-stage object detector and is widely adopted for domain adaptive detection.

Unsupervised domain adaptation. Domain shift problem is frequently encountered in the real world when models are deployed in slightly different conditions compared to the training data. This issue is addressed using UDA approaches resulting in better generalization. A lot of UDA methods based on adversarial learning have been proposed. Previous work in the UDA setting are investigated in different topics including image classification [26, 9, 14, 22, 25, 37, 51], semantic segmentation [6, 16, 28, 41, 42, 54].

Domain adaptation for object detection. Researches on domain adaptation in object detection are still in the early stage. With the achievement of domain adaptation in classification, Chen *et al*. [3] proposed a pioneering framework of domain adaptive Faster R-CNN. For domain adaptive Faster R-CNN, [32] focuses the adversarial alignment loss on globally similar images. As stated in the paper, the method does not work right when the appearance of the objects is largely different. The unsuccessful detection results also suggest that structural importance is necessary for alignment. Xu *et al*. [48] recently explored categorical consistency between image and instance-level prediction, aligning relevant object regions. But they use an additional instance-level classifier as a regulariser. Chen *et al*. [2] proposed a method for harmonizing transferability and discriminability of features. CycleGAN was used to generate interpolated images to reduce global discriminability. Vibashan *et al*. [40] proposed a method to ensure category-aware feature alignment for learning domain-invariant features. They generate category-specific attention maps since category information is not available for target samples. Li *et al*. [21] proposed a self-training paradigm to reduce the domain gap. This requires pseudo-labels as ground truth to exploit the unlabeled

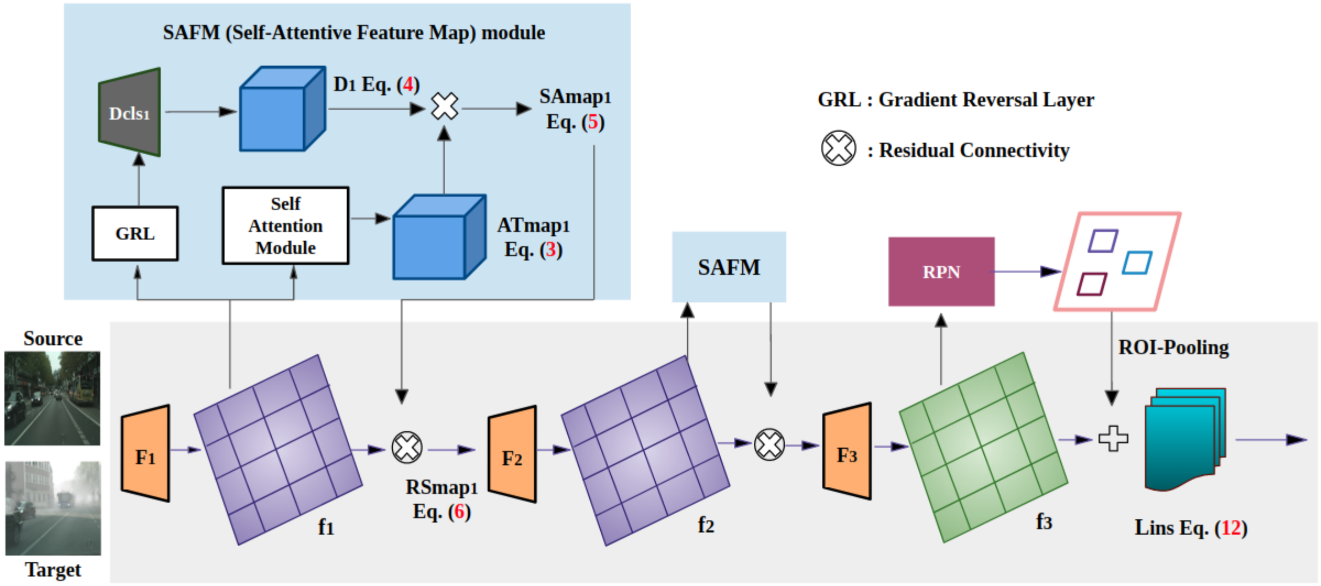


Figure 2: Proposed architecture of the ILLUME. Initial image features are fed to the SAFM module to generate self-attentive feature maps and pixel-wise domain classifier output, combined to form final attention feature maps. They are combined with initial feature maps of the image in a residual manner. The SAFM module is incorporated at multiple stages in our network thereby generating domain invariant features and enhancing structural importance during feature alignment.

target data. These state-of-the-art methods require additional modeling or regularisation for better alignment of instances. Comparatively, our method is simple yet efficient, with focus on aligning important instances using residual self-attention and does not require any regularisation.

Attention in domain adaptation. The self-attention mechanism hasn't been explored in domain adaptive object detection space. Prior works [2, 40, 41, 34, 44, 19, 43, 27, 20] have focused on using attention or entropy maps for various tasks like classification, detection, and segmentation. Such works require additional pseudo-labeling in the target domain [34, 2, 43] or use attention only at the local level [44] or to optimize the relationships of multiple domain samples for better semantic transfer [50]. [19] uses cycleGAN for domain translation. [40] use category-specific attention to route features to the category discriminator. [20] employ the task-specific semantic information to guide pyramid attention. Some works use *self-attention in classification/segmentation* task - to augment source features by selectively aggregating target features based on their similarities [49] or to eliminate the influence of untransferable features [53] or just as a regularizer independent of the adaptation [45]. These are entirely different from the way we use *self-attention for detection* task to improve instance feature alignment. We also don't require additional regularisation or augmentation, given the ability of ILLUME to align the instances well.

Some of the prior works in UDA also focus on the feature misalignment issue. [18] performs class-aware domain alignment. [47] weights the learning losses of alignment to guarantee information balance. [10] proposed an uncertainty metric that assesses the alignment of each sample. ILLUME effectively outperforms these works as well as methods using attention for classification as shown in Table 6.

3. Approach

We design a novel architecture that comprises the self-attention feature alignment module. We aim to align feature distributions of instances that are prominent for the adaptive detection instead of background as discussed in Section 1. Initially, the features are extracted from a batch of training data, comprised of one source and one target image. These features are then fed to the Self-Attentive Feature Map (SAFM) module. SAFM generates self-attentive feature maps and pixel-wise domain classifier output using the Gradient Reversal Layer (GRL) [8] used by the domain classifier. These features are further combined to form domain invariant feature maps and then combined in a residual manner where the final attention feature maps are combined with the initially extracted feature maps of the image. Further SAFM module is incorporated at multiple stages in our network thereby generating domain invariant features.

3.1. Problem Formulation

In Unsupervised Domain Adaptation (UDA), we have labeled source data $D_s = (X_i^s, Y_i^s)$ where X represents the source images and $Y \in R^{m\times 5}$ represents the list of bounding boxes and corresponding class labels. Similarly, we have unlabelled target data $D_t = X_i^t$ for training the detector. Despite sharing similar label space, the source and target data are sampled from different distributions which contribute to the large domain gap between them. ILLUME helps to generate self-attentive domain invariant features, necessary for alignment in DA. Our proposed method is based on the Faster R-CNN framework.

3.2. Residual Self-attentive Feature Alignment

Our proposed method comprises a novel residual self-attentive feature alignment module shown in the architecture diagram (Fig. 2). This method generates self-attention feature maps and a pixel-level domain classifier output, discussed in detail in the next section.

3.2.1 Self-Attentive Feature Map Module (SAFM)

Initially, the image is fed to the feature extractor F_1 as shown in Fig. 2. F_1 extracts image-level features f_1, which are then passed to the SAFM module. In the SAFM, we use a self-attention module and a domain classifier to produce the combined self-attentive domain invariant feature maps. The domain classifier is trained in an adversarial manner with loss function L_1 as shown in the equation below,

$$L_1 = \sum_{i=1}^{n_s} \log(D_1(F_1(x_i^s)))^2 + \sum_{i=1}^{n_t} \log(1 - D_1(F_1(x_i^t)))^2 \tag{1}$$

where $F_1(x^s)$ denotes extracted features of source training data and $F_1(x^s)$ denotes extracted features of target training data respectively. D_1 is a domain classifier that produces a pixel-wise probability output map for the source and target input features. The gradient reversal layer is used in between the feature extractor and domain classifier of an adaptive object detector in an adversarial manner, which reverses the gradients during backpropagation. The self-attentive feature maps are generated using the self-attention module where we use key, query, and value vectors. The Q (query) and K (key) undergo a matrix multiplication and pass through a softmax which converts the resultant vector into a probability distribution, and then it finally gets multiplied by V (value). In self-attention, the query, the key and, the value is all same. Here we pass the features f_1 extracted from feature extractor F_1 to the SAFM module to generate self-attention feature maps enhancing structural importance, necessary for alignment, where $Q_1 = K_1 = V_1 = f_1$. The Query Q_1 and Key K_1 undergo matrix multiplication as shown in Eq.(2). Further, they undergo softmax and finally are multiplied to value vector V_1 as shown in Eq.(3). which gives the final self-attentive feature maps.

$$QK_1 = Q_1 \times K_1 \tag{2}$$

$$AT_{map_1} = softmax(QK_1) \times V_1 \tag{3}$$

The pixel-wise probability output of discriminator D_1 is denoted as shown in Eq.(4). Finally the features from SAFM module AT_{map_1} and domain classifier output D_1 are combined to form the final domain invariant self-attention feature maps SA_{map_1} as shown in equation below,

$$D_1 = D_{cls_1}(f_1) \tag{4}$$

$$SA_{map_1} = AT_{map_1} \times D_1 \tag{5}$$

3.2.2 Residual Connectivity and Multi Stage SAFM

We now combine the self-attentive feature maps SA_{map_1} and the initial image features f_1 in a residual manner, as shown in the Eq. (6). Resultant features RS_{map_1} are domain invariant. Such incorporation of SAFM module to generate domain invariant features at multiple stages of the network can be very efficient in terms of the transferability of features. Hence we incorporate SAFM module in our network at multiple stages as shown in the Eq. (7) and (8).

$$RS_{map_1} = SA_{map_1} \times f_1 \tag{6}$$

We show the attention feature maps RS_{map_1} in Fig. 1. Attention feature maps quantify the importance of structural alignment necessary for aligning features of instances instead of background like road/sky. Final features f_3 which are the output of the detection backbone network are then passed to the RPN layer to train the Faster R-CNN for UDA.

$$f_2 = F_2(RS_{map_1}) \tag{7}$$

$$RS_{map_2} = SA_{map_2} \times f_2 \tag{8}$$

$$f_3 = F_3(RS_{map_2}) \tag{9}$$

The adversarial loss L_2 can be derived as,

$$L_2 = \mathbb{E}[\log(D_2(F_2(RS^s_{map_1}))) + \log(1 - D_2(F_2(RS^t_{map_1})))] \tag{10}$$

RS_{map_1} and RS_{map_2} are the final domain invariant attention feature maps passed to feature extractors F_2 and F_3, respectively. The transformed features f_3 form the final output of the backbone detection network which are fed to the Region Proposal Network (RPN) of Faster R-CNN detector for object detection. Finally, the instance level features f_{ins} are fed to the instance level domain classifier D_{ins} giving the instance level domain classifier loss. The instance-level adversarial domain classifier loss is as follows,

$$L_{ins} = -\sum_{i=1}^{n_s}\sum_{j} \log(D_{ins}(f^s_{ins})_j) - \sum_{i=1}^{n_t}\sum_{j} \log(1 - D_{ins}(f^t_{ins})_j) \tag{11}$$

3.3. Training Loss

The training loss is the combination of the detection loss and the adversarial loss. The detection loss is as shown,

$$L_{det} = L_{cls} + L_{reg} \tag{12}$$

where L_{cls} and L_{reg} denotes the classification and regression losses respectively for object detection. Note that this detection loss is calculated only on source samples and not the target samples as in UDA we consider unlabelled target samples. The adversarial loss is a combination of multistage detector training losses given as follows,

$$L_{adv} = L_1 + L_2 + L_{ins} \tag{13}$$

where, L_1, L_2 and L_{ins} are the multi-stage losses of the domain classifiers D_1, D_2 and D_{ins} as discussed in previous section. Finally, the overall training loss can be derived as,

$$Loss = \max_{D_i} \min_{F_i} L_{det} + \lambda \cdot L_{adv} \tag{14}$$

where λ is a hyperparameter used for balancing detection and adversarial losses while training.

4. Experiments

4.1. Datasets

To perform different experiments we utilized six different datasets. **Cityscapes** [4] has variational data of outdoor street scenes captured in normal weather conditions from different cities. It contains 2975 training and 500 validation images. The bounding box annotations for object detection have been generated by transforming the instance segmentation annotations in the dataset for our experiments. **Foggy Cityscapes** [33] is inherited from the Cityscapes dataset by using depth information to simulate the foggy weather with three levels of foggy weather suitable for weather adaptation experiments. **BDD100K** [52] contains 100K images, 70K training, and 30K validation images with annotated bounding boxes. We use a daytime subset of this data for scene adaptation experiments consisting of 36,728 training and 5,258 validation images. **Sim10K** [17] is a synthetic dataset that is rendered from the game Grand Theft Auto V (GTA V). It consists of 10,000 images of street scenes with 58,701 bounding box annotations for cars. **Pascal VOC** [7] is a dataset containing real-world image data of 20 different common object categories along with bounding box annotations. We use both Pascal VOC 2007 as well as 2012 training and validation data (16,551 training images in total) for our experiments. **Clipart1K** [15] consists of 1K training images with the same classes as Pascal VOC but exhibits a significant domain gap between them. We use all the training and testing images in clipart1K for experiments.

4.2. Implementation Details

We use the PyTorch framework for all the tasks. Following the practices in [38, 48] we employ VGG-16 [35] as backbone network where their weights are pre-trained on the Imagenet Dataset [5], but for dissimilar domain adaptation experiments for Pascal VOC [7] to Clipart1K [15], we follow practices of [32, 48] and use ResNet-101 [12] as backbone network for detection. In all our experiments the shorter side of all the training and testing images is resized to 600 following practices in [3, 32, 48]. Each batch in our training data is composed of two images, one from the source and another from target samples. We fine-tune the detection network with a learning rate of 1×10^{-3} for 50K iterations and then reduce the learning rate to 1×10^{-4} for the other 20K iterations. The momentum of 0.9 and the weight decay of 5×10^{-4} is used for VGG-16 [35] based detectors, while for ResNet-101 [12] based detectors, we set it to 1×10^{-4}. In all experiments, we use RoIAlign [11] for RoI feature extraction. The hyperparameter λ is set to 0.1 for synthetic to real adaptation task (Sim10K [17] $\rightarrow$ Cityscape [4]) and $\lambda = 1$ for all other adaptation tasks. We compare our method with Source-Only baseline [3] (Faster R-CNN trained using only source images) as well as the other state-of-the-art methods [3, 1, 2, 48, 59, 56, 13, 36, 58, 23, 46, 40, 55, 21]. We use mean average precision (mAP) metrics for evaluation.

4.3. Comparisons with State-of-the-Art

Synthetic to Real Adaptation. Synthetic data is profoundly available and can be processed to create large datasets instead of creating real-time data. Results of adaptation from Sim10K [17] as source to Cityscapes [4] as target are shown in Table 3. Our method's performance boosts the AP by 19% over the Source-Only baseline. Moreover, it outperforms all state-of-the-art methods with a significantly higher margin and improvement in mAP of over 4.3%.

Weather Adaptation. We use Cityscapes [4] as source and Foggy Cityscapes [33] as target for weather adaptation. ILLUME boosts the performance of Faster R-CNN detector with an improvement of 1.5% and 22% in mAPs as compared to the state-of-the-art and source-only baseline, respectively, as shown in Table 1. It also improves performance over Oracle detector (Supervised). We note that for most of the classes, our model outperforms oracle reasonably cause of our adaptive detection method compared to the full supervision in foggy weather. Our results are highest for all the categories except for only two - person and rider, for which DSA [46] have better accuracies. Yet, their method restrains the learning of final task, requiring an additional pre-training process to overcome the limitation.

Scene Adaptation. We choose Cityscapes [4] training set as source and a subset of BDD100K as target to study the effectiveness of ILLUME for scene adaptation. We choose

Table 1: **Weather Adaptation:** Results on Foggy Cityscapes, using models trained on Cityscapes

Method	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
Faster R-CNN (Source) [3]	24.4	25.4	32.6	10.8	30.5	9.1	15.2	28.3	22.0
MTOR (ICCV'19) [1]	30.6	41.4	44.0	21.9	38.6	40.6	28.3	35.6	35.1
AFL (AAAI'21) [23]	34.1	44.4	51.9	30.4	41.8	25.7	30.3	37.2	37.0
DMLP (ECCV'20) [56]	32.0	42.1	43.9	31.3	44.1	43.4	37.4	36.6	38.8
ATFR (ECCV'20) [13]	34.6	43.3	50.0	23.7	47.0	38.7	33.4	38.8	38.7
PDA (ECCV'20) [36]	36.4	47.3	51.7	22.8	47.6	34.1	36.0	38.7	39.3
SDA (CVPR'19) [58]	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
ICR-CCR (CVPR'20) [48]	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
HTCN (CVPR'20) [2]	33.2	47.4	47.9	31.6	47.5	40.9	32.3	37.1	39.8
RPN-PR (CVPR'21) [55]	33.3	45.6	50.5	30.4	43.6	42.0	29.7	36.8	39.0
DSA (CVPR'21) [46]	**42.9**	**51.2**	53.6	33.6	49.2	18.9	36.2	41.8	40.9
MeGA-CDA (CVPR'21) [40]	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8
CGD (AAAI'21) [21]	38.0	47.4	53.1	34.2	47.5	41.1	38.3	38.9	42.3
ILLUME **(Ours)**	35.8	45.1	**54.3**	**34.5**	**49.7**	**50.3**	**38.7**	**42.0**	**43.8**
Faster R-CNN (Oracle)	36.2	47.7	53.0	34.7	51.9	41.0	36.8	37.8	42.4

Table 2: **Scene Adaptation:** Results of 7 common categories of BDD100K, using models trained on Cityscapes.

Method	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
Faster R-CNN (Source) [3]	26.9	16.7	44.7	17.4	22.1	-	17.1	18.8	23.4
ICR-CCR [48]	31.4	18.9	46.3	19.5	31.3	-	17.3	23.8	26.9
AFL [23]	32.4	**32.6**	**50.4**	20.6	23.4	-	18.9	25.0	29.0
ILLUME **(Ours)**	**33.2**	20.5	47.8	**20.8**	**33.8**	-	**24.4**	**26.7**	**29.6**

Table 3: **Synthetic to Real Domain Adaptation:** Results on Sim10K to Cityscapes (%).

Methods	AP on Car
Faster R-CNN (Source) [3]	34.6
ATFR [13]	42.8
HTCN [2]	42.5
AFL [23]	43.1
DSA [46]	44.5
MeGA-CDA [40]	44.8
RPN-PR [55]	45.7
iFAN [59]	47.1
CGD [21]	48.8
ILLUME **(Ours)**	**53.1**

only daytime annotated subset of BDD100K as target since there exists only daytime data in the Cityscapes [4]. We report the detection results on seven common categories for both datasets. As shown in Table 2, we achieve improved performance by 0.6% and 6% in mAPs as compared to the current state-of-the-art (AFL) and source-only baseline.

Dissimilar Domain Adaptation. We utilize Pascal VOC [7] as source and Clipart1K [15] as target for dissimilar domain adaptation from real to artistic images. Our method's performance boosts the AP by 14% over the Source-Only baseline as shown in Table 4. Moreover, it outperforms most of the state-of-the-art methods with comparable performance to AFTR [13]. We perform equally better for all categories. We do not have a degrading performance on any, as can be seen for the dog, cat, or sheep categories using AFTR or sheep and dog categories using MEAA, that have very low mAPs overall. Notably our framework comprises an independent multi-stage SAFM, while ATFR uses an additional ancillary net for source risk bounding.

4.4. Visualisation and Analysis

Ablation Study. We conduct ablation in the absence of SAFM at multiple stages of the network, as shown in Table 5. We can see a high-performance drop-in mAP by 4.8% without any SAFM (row 1) and a 2.7% drop with a single first-layer SAFM module (row 2), compared to ILLUME. This drop denotes significance of multi-stage residual self-attentive alignment that substantially improves detection performance. We choose a design to use 2-SAFM(s) because with additional SAFM(s), we got slightly degraded performance due to overfitting as shown in last row of table. We consciously feed high-level features to SAFM to enhance structural object-related regions.

Visualisations of the Transformed Features. As shown in Fig. 3 (top), we visualize the transformed features that are the output of the detection backbone, using ILLUME. The enhanced features depict the efficiency of our method to transform features such that only important instance features would be considered by Faster R-CNN to learn domain invariance essential for alignment. We show visualizations

Table 4: **Dissimilar Domain Adaptation:** Results on the Clipart1K dataset, using models trained on the Pascal VOC

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Source only [3]	21.9	42.2	22.9	19.0	30.8	43.1	28.9	10.7	27.4	18.1	13.5	10.3	25.0	50.7	39.0	37.4	6.9	18.1	39.2	34.9	27.0
ICR-CCR [48]	28.7	55.3	31.8	26.0	40.1	63.6	36.6	9.4	38.7	49.3	17.6	14.1	33.3	74.3	61.3	46.3	22.3	24.3	49.1	44.3	38.3
HTCN [2]	33.6	58.9	34.0	23.4	45.6	57.0	39.8	12.0	39.7	51.3	21.1	20.1	39.1	72.8	63.0	43.1	19.3	30.1	50.2	51.8	40.3
ATFR [13]	41.9	67.0	27.4	36.4	41.0	48.5	42.0	13.1	39.2	75.1	33.4	7.9	41.2	56.2	61.4	50.6	42.0	25.0	53.1	39.1	42.1
ILLUME **(Ours)**	34.4	58.5	33.3	27.5	37.8	59.9	36.8	25.2	38.9	47.6	26.2	24.5	21.9	72.4	67.4	47.6	27.4	35.9	51.4	58.4	41.6

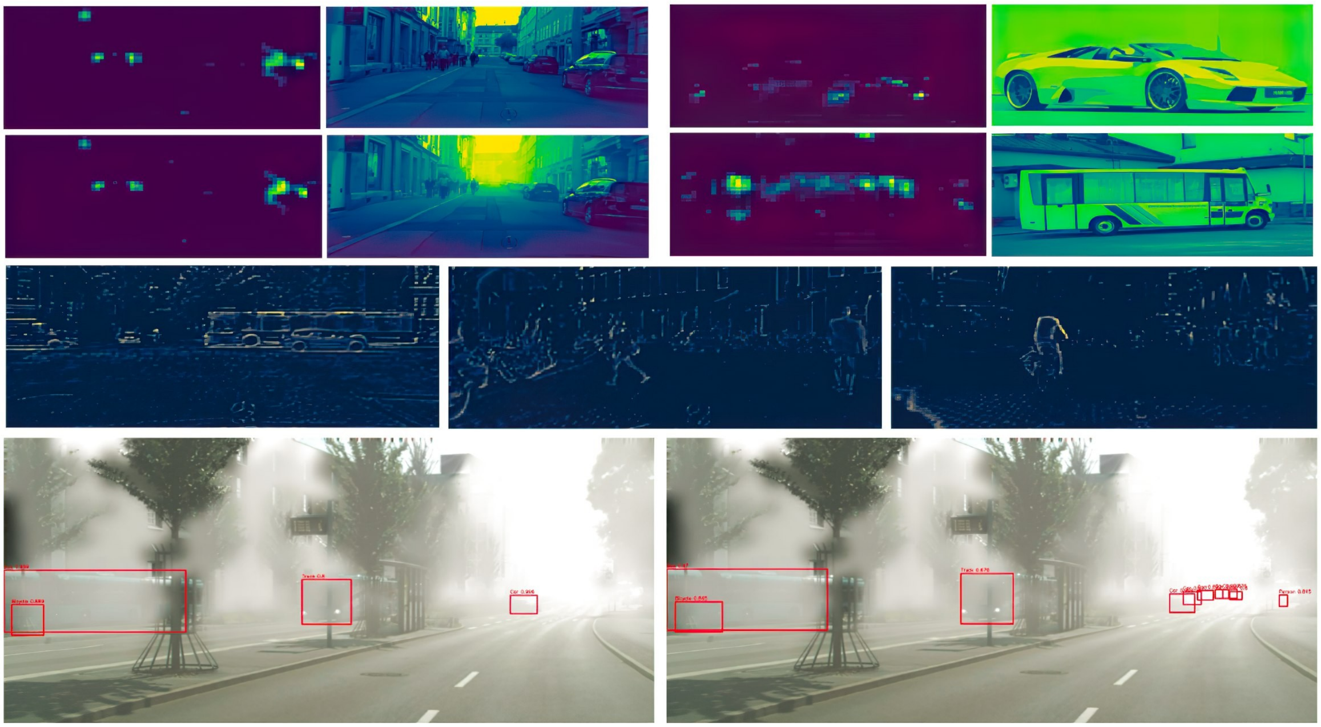


Figure 3: **Top:** Visualizations of the transformed features using our ILLUME method that enhance the importance instances. It is worth noting that, the (top left) feature maps for source (Cityscapes) and target (Foggy Cityscapes) images are similar irrespective of the difference in the domains due to foggy weather. The top right are the similar visualizations results for dissimilar DA tasks (Pascal VOC to clipart). **Middle:** Visualisations of self-attentive feature maps enhancing instances like car or bus in the image obtained using SAFM module. **Bottom:** Detection results on target images (Foggy Cityscapes), comparing the state-of-the-art [21] (left) and our method (right).

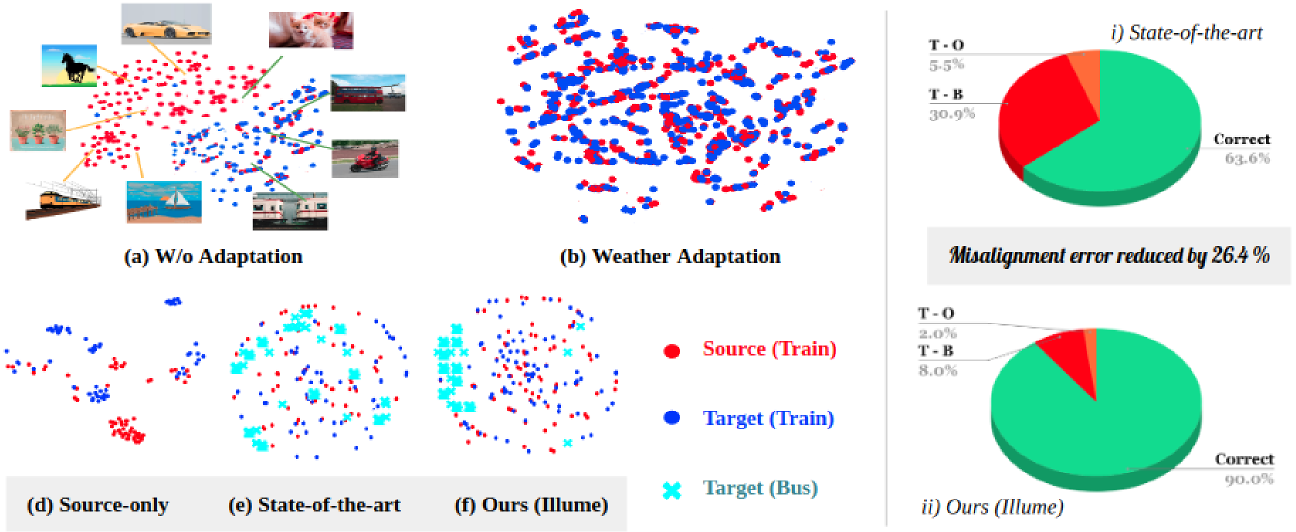
Table 5: Ablation on Cityscapes to Foggy Cityscapes.

Method	per	rid	car	tru	bus	tra	mcy	bic	mAP
Source-Only [3]	24.4	25.4	32.6	10.8	30.5	9.1	15.2	28.3	22.0
ILLUME w 0-SAFM	34.7	45.3	51.9	31.8	47.2	30.6	32.3	38.2	39.0
ILLUME w 1-SAFM	35.6	47.8	52.7	33.7	45.2	40.3	34.5	38.7	41.1
ILLUME **(Ours)**	35.8	45.1	54.3	34.5	49.7	50.3	38.7	42.0	**43.8**
ILLUME w 3-SAFM	35.0	44.8	54.1	33.8	47.9	48.4	38.2	41.6	43.0
Oracle	36.2	47.7	53.0	34.7	51.9	41.0	36.8	37.8	42.4

for two DA tasks. For the weather adaptation (left), it is worth noting that the feature map for both source and target images highlighting the important instances is similar irrespective of the domain gap due to foggy weather. This proves the efficiency of our method. For dissimilar adaptation (right), a similar visualization analysis can be seen.

Self-attentive Feature Maps and Detection Results. We visualize self-attention feature maps, enhancing important instances in the image for alignment (using SAFM). We show improved detection results using ILLUME (right) as compared to state-of-the-art [21] (left) on target images (Foggy Cityscapes) for weather adaptation task in Fig. 3.

Feature Visualisation using t-sne. We visualize the improved feature alignment for weather adaptation task using t-sne as shown in Fig.4 (a). For this experiment, we randomly sample 500 images from the source and target domain each. The image features are extracted by applying global average pooling on the output of the detection backbone. We also show the instance-level visualizations for *train* instance. For this, we randomly sample 100 ground truth instances for each category, extracted by ROIAlign. State-of-the-art method [21] misalign the features for source - *train* with target - *bus*. Comparatively, with our approach the target instance features for *bus* form a separate cluster and source features for *train (red)* align perfectly with


(a) W/o Adaptation

(b) Weather Adaptation

(d) Source-only

(e) State-of-the-art

(f) Ours (Illume)

Figure 4: Visualization of features using t-sne [39]. (b) shows the adaptation results for Cityscape (source - red) to Foggy Cityscape (target - blue). Our method match feature distributions strictly well. The bottom row shows the instance-level feature visualizations using t-sne for *train* as the source with its aligned target instance. State-of-the-art method misalign the *train* instance features with *bus* (cyan), compared to ours where instance features for source - *train* are correctly aligned with target - *train* instead of *bus* that are clustered separately. The domain distance is reduced significantly using our approach. We also show graphs for misalignment error reduction using our approach. *T - train. O - other instances. B - bus.*

the correct target - *train (blue)* instead of some other instance like *bus*. **Instance feature misalignment.** We calculate misalignment error for this experiment as shown in figure (right). The combined error for train instance misaligned with bus *(T-B)* and other instances *(T-O)* is reduced by large margin of 26.4% as compared to state-of-the-art. **Domain distance.** We also calculate quantitative metric for domain distance, where both domains are represented by object instances. We use the same number of instance samples as the feature visualization experiment for all categories. Specifically, we adopt euclidean distance as the metric for measuring domain distance. With this metric, domain distance computed are 8.1 and 6.3 using state-of-the-art and our method respectively. The consistency between domain distance, improvement in the accuracy of the instances and better instance-level feature alignment verifies the motivation for our work and effectiveness of ILLUME.

Application on UDA for classification. The effectiveness of our method increases the scope of exploring self-attentive feature alignment for other tasks in domain adaptation. We show the applicability of our method for other tasks like classification and perform experiments on the widely used Office-31 dataset [31]. It contains 4,110 images of 31 categories in three domains: Amazon (A), Webcam (W), DSLR (D). With incorporation of our multi-stage residual SAFM(s) and training with combined adversarial as well as classification loss; we evaluated our method on six adaptation tasks as shown in Table 6. We achieve competitive performance with an accuracy of 91%, without additional domain augmentation like FixBi [26]. Notably, we compare against approaches using attention [44, 34, 53] as well as the works focusing on feature misalignment [47, 18] and recent state-of-the-arts in classification task for UDA. We briefly discuss these prior works in Section 2. This proves the efficiency and applicability of ILLUME.

Table 6: Accuracy (%) on Office-31 for classification in DA

Method	A-W	D-W	W-D	A-D	D-A	W-A	Avg
TADA (AAAI'19) [44]	94.3	98.7	99.8	91.6	72.9	73.0	88.4
EADA [34]	94.0	97.9	100.0	94.2	74.6	74.9	89.3
TAN [53]	95.4	98.7	100.0	93.3	73.7	75.1	89.3
CRA (AAAI'21) [57]	93.0	99.0	100.0	95.6	78.9	74.7	90.2
DWL (CVPR'21) [47]	89.2	99.2	100.0	91.2	73.1	69.8	87.1
CAN (CVPR'19) [18]	94.5	99.1	99.8	95.0	78.0	77.0	90.6
FixBi (CVPR'21) [26]	96.1	99.3	100.0	95.0	78.7	79.4	91.4
ILLUME (Ours)	96.2	99.4	99.9	96.8	76.4	77.7	91.0

5. Conclusion

We presented a novel ILLUME method for adaptive object detection. Specifically, we explored self-attention mechanism for enhancing prominent object-related regions to improve feature alignment. Our method significantly reduces misalignment error and domain distance. ILLUME outperforms the performance of existing adaptive Faster R-CNN detectors and set new benchmarks.

6. Acknowledgement

This work was partly funded by IHub-Data at IIIT Hyderabad, and DST through the IMPRINT program.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019.

[2] 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 (CVPR)*, June 2020.

[3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain Adaptive Faster R-CNN for Object Detection in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. What Can be Transferred: Unsupervised Domain Adaptation for Endoscopic Lesions Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[8] Yaroslav Ganin and Victor Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189. PMLR, 07–09 Jul 2015.

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[10] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao. Uncertainty-aware unsupervised domain adaptation in object detection. *IEEE Transactions on Multimedia*, pages 1–1, 2021.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] Zhenwei He and Lei Zhang. Domain Adaptive Object Detection via Asymmetric Tri-Way Faster-RCNN. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 309–324, Cham, 2020. Springer International Publishing.

[14] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[15] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-Domain Weakly-Supervised Object Detection Through Progressive Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[16] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Perez. xMUDA: Cross-Modal Unsupervised Domain Adaptation for 3D Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[17] 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)*, pages 746–753. IEEE, 2017.

[18] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G. Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[19] Guoliang Kang, Liang Zheng, Yan Yan, and Yi Yang. Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[20] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 481–497. Springer, 2020.

[21] Shuai Li, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Category dictionary guided unsupervised domain adaptation for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1949–1957, 2021.

[22] Shuang Li, Chi Harold Liu, Binhui Xie, Limin Su, Zhengming Ding, and Gao Huang. Joint Adversarial Domain Adaptation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 729–737. Association for Computing Machinery, 2019.

[23] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8474–8481, 2021.

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C

Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[25] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning Transferable Features with Deep Adaptation Networks. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 97–105. PMLR, 07–09 Jul 2015.

[26] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1094–1103, June 2021.

[27] Dang-Khoa Nguyen, Wei-Lun Tseng, and Hong-Han Shuai. Domain-Adaptive Object Detection via Uncertainty-Aware Distribution Alignment. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2499–2507, 2020.

[28] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised Intra-Domain Adaptation for Semantic Segmentation Through Self-Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[31] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[32] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-Weak Distribution Alignment for Adaptive Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[33] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic Foggy Scene Understanding with Synthetic Data. *International Journal of Computer Vision*, 126(9):973–992, 2018.

[34] Kekai Sheng, Ke Li, Xiawu Zheng, Jian Liang, Weiming Dong, Feiyue Huang, Rongrong Ji, and Xing Sun. On evolving attention towards domain adaptation. *arXiv preprint arXiv:2103.13561*, 2021.

[35] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, May 7-9, 2015, Conference Track Proceedings*, 2015.

[36] 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*, pages 763–780. Springer, 2020.

[37] Baochen Sun and Kate Saenko. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 443–450. Springer International Publishing, 2016.

[38] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to Adapt Structured Output Space for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[39] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[40] Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A. Sindagi, and Vishal M. Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4516–4526, June 2021.

[41] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[42] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. DADA: Depth-Aware Domain Adaptation in Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[43] Jinghua Wang. Exploring category attention for open set domain adaptation. *IEEE Access*, 9:9154–9162, 2021.

[44] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5345–5352, 2019.

[45] Yimu Wang, Renjie Song, Xiu-Shen Wei, and Lijun Zhang. An adversarial domain adaptation network for cross-domain fine-grained recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1228–1236, 2020.

[46] 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 (CVPR)*, pages 9603–9612, June 2021.

[47] Ni Xiao and Lei Zhang. Dynamic weighted learning for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15242–15251, June 2021.

[48] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring Categorical Regularization for Domain Adaptive Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[49] Jinyu Yang, Weizhi An, Chaochao Yan, Peilin Zhao, and Junzhou Huang. Context-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 514–524, January 2021.

[50] Xu Yang, Cheng Deng, Tongliang Liu, and Dacheng Tao. Heterogeneous graph attention network for unsupervised multiple-target domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[51] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Universal Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[52] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[53] Changchun Zhang, Qingjie Zhao, and Yu Wang. Transferable attention networks for adversarial domain adaptation. *Information Sciences*, 539:422–433, 2020.

[54] Yang Zhang, Philip David, and Boqing Gong. Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[55] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12425–12434, June 2021.

[56] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye. Adaptive Object Detection with Dual Multi-label Prediction. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 54–69, Cham, 2020. Springer International Publishing.

[57] Li Zhong, Zhen Fang, Feng Liu, Jie Lu, Bo Yuan, and Guangquan Zhang. How does the combined risk affect the performance of unsupervised domain adaptation approaches? In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021.

[58] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2019.

[59] Chenfan Zhuang, Xintong Han, Weilin Huang, and Matthew Scott. iFAN: Image-Instance Full Alignment Networks for Adaptive Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13122–13129, Apr. 2020.