MeGA-CDA: Memory Guided Attention for Category-Aware Unsupervised Domain Adaptive Object Detection

Vibashan VS[1*†], Vikram Gupta[2*], Poojan Oza[1*], Vishwanath A. Sindagi[1*], Vishal M. Patel[1]

[1] Johns Hopkins University, Baltimore, MD, USA

[2] Mercedes-Benz Research and Development India

{vvishnu2,poza2,vishwanathsindagi,vpatel36}@jhu.edu,vikram.gupta@daimler.com

Abstract

*Existing approaches for unsupervised domain adaptive object detection perform feature alignment via adversarial training. While these methods achieve reasonable improvements in performance, they typically perform category-agnostic domain alignment, thereby resulting in negative transfer of features. To overcome this issue, in this work, we attempt to incorporate category information into the domain adaptation process by proposing **Me**mory **G**uided **A**ttention for **C**ategory-**A**ware **D**omain **A**daptation (MeGA-CDA). The proposed method consists of employing category-wise discriminators to ensure category-aware feature alignment for learning domain-invariant discriminative features. However, since the category information is not available for the target samples, we propose to generate memory-guided category-specific attention maps which are then used to route the features appropriately to the corresponding category discriminator. The proposed method is evaluated on several benchmark datasets and is shown to outperform existing approaches.*

1. Introduction

Object detectors [49, 11, 15, 14, 28, 38] are a critical part in the inference pipeline of several applications like autonomous navigation, video surveillance, image analysis *etc*. Due to this, object detection has received significant interest from the research community. Recent works like [15, 28, 38] have achieved exceedingly good performance on several benchmark datasets [9, 8, 13, 27]. However, these approaches suffer from severe degradation of performance when evaluated on images that are sampled from a different distribution as compared to that of training images. Such scenarios are encountered frequently in the real world.

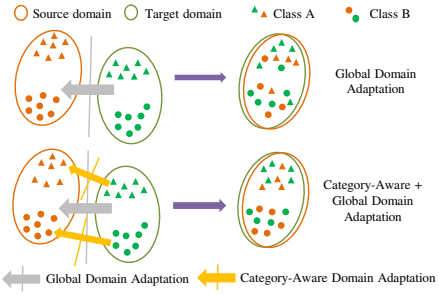

Figure 1. Performing global domain adaptation alone results in potential negative transfer of features. To mitigate this issue, we employ additional category-aware adaptation.

For example, consider the case of self-driving cars where the detectors are typically trained on datasets obtained from one particular city or environmental condition (belonging to source domain) and are expected to be deployed in different city or environment (belonging to target domain). Due to this, it is crucial to develop approaches that enable better generalization of detectors.

One approach to address this issue is through unsupervised domain adaptation [6, 44, 40], where the goal is to utilize labeled source domain data and unlabeled target domain data to adapt object detector and improve the performance on the unlabeled target domain data. To address this issue, typically these methods attempt to learn domain-invariant features by performing feature alignment between the source and target images. Based on the theoretical insights that minimizing the divergence between the domains reduces the upper-bound error on the target domain [4], they achieve the feature alignment through adversarial training. Although these approaches result in considerable improvements, they perform the domain alignment in a category-agnostic way. That is, they match the global marginal distributions of the two domains without considering the category information. This may lead to cases where the target domain samples are incorrectly aligned with the source-

*Equal contribution

†Work performed during internship at Mercedes-Benz Research and Development India

domain samples of a different class (see Fig. 1), thereby resulting in sub-optimal adaptation performance. The task of adapting object detectors is especially prone to this problem due to the presence of multiple categories of objects.

Considering this issue, we focus on incorporating category information into the domain adaptation process by matching the local joint distribution of features in addition to the global alignment. In particular, we perform category-wise alignment of features by employing category-specific discriminators in the training process. Note that this requires pixel-wise category labels so that the features can be explicitly routed to the respective category-specific discriminator. However, in the case of unsupervised domain adaptation, annotations are not available in any form (bounding boxes/category labels/pixel-wise labels) for the target dataset. This lack of annotations makes it difficult to use category-specific discriminators.

In order to overcome this challenge, we propose memory-guided attention maps for enabling the category-aware feature alignment. The objective of these attention maps is to focus on objects of specific categories, and hence can be used to route the backbone features into the appropriate category-specific discriminators. For generating these attention maps, we propose the use of memory networks [51, 47, 26, 23, 10, 34, 16]. During the training process, these memory banks are used to store prototypes of the objects of different categories, where individual items in the memory correspond to prototypical features of a particular object category. The use of memory network is inspired by their ability to store patterns over longer periods of time. Additionally, the ability to update the patterns using explicit write operations makes them especially useful in domain adversarial training since the features change over the training process. For determining the attention at a particular location, we use the feature at this location as a query to retrieve relevant items from the different category-specific memory networks. The retrieved items are then compared with the query item and based on the similarity, we compute the category-specific attention map. Furthermore, in order to improve the effectiveness of the memory module and the attention map generation process, we propose a metric-learning based approach that involves learning an appropriate similarity metric based on the available weak-supervision in the source domain. In order to demonstrate the effectiveness of the proposed method, we evaluate it on several benchmark datasets and adaptation protocols. Furthermore, we show that the memory-guided attention maps play an important role in achieving category-wise distribution matching, thereby mitigating the issue of incorrect feature alignment.

To summarize, following are the main contributions of our work:

- We propose memory-guided attention maps for enabling category-wise distribution matching for domain adaptive object detection.
- In addition, we improve the effectiveness of the memory modules by employing metric learning-based approach for computing the category-specific attention maps.
- The proposed method is evaluated on several benchmark datasets and is shown to outperform recent domain adaptive detection approaches by a considerable margin. Additionally, we conduct detailed ablation studies to clearly disambiguate the role of memory-guided attention for achieving category-wise alignment.

2. Related work

Object detection: The problem of object detection has attracted significant interest due to its widespread applications in several higher-level inference tasks. Recent approaches have benefited largely from the success of convolutional neural networks, where different techniques have developed anchor-based strategies for achieving high performance object detection. These approaches can be broadly categorized into (i) two-stage [38] and (ii) single-stage approaches [37, 28].

Unsupervised domain adaptation: Deep-learning based methods are affected by the domain-shift problem [35, 50], where networks trained on one distribution of data tend to perform poorly on a different distribution of data. This 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 typically using unsupervised domain adaptation approaches, where the data from different domains are aligned so that the resulting networks/models achieve good generalization performance. Recent domain adaptation approaches involve feature distribution alignment [48, 12, 45, 41], residual transfer [30, 31], and image-to-image translation approaches [21, 33, 19, 43, 2, 1, 36].

Domain adaptation for object detection: The task of domain adaptation for object detection was recently introduced by Chen *et al*. [6], where they address the problem of domain shift at both image-level and instance-level. Shan *et al*. [44] proposed to perform joint adaptation at image level using the Cycle-GAN framework [54] and at feature level using conventional domain adaptation losses. Saito *et al*. [40] showed that strong alignment of the features at global level is not necessarily optimal and proposed strong alignment of the local features and weak alignment of the global features. Kim *et al*. [25] diversified the labeled data, followed by adversarial learning with the help of multi-domain discriminators. Cai *et al*. [5] addressed this problem in the semi-supervised setting using mean teacher framework. Zhu *et al*. [55] proposed region mining and region-level alignment in order to correctly align the source and target features. Roychowdhury *et al*. [39] adapted de-

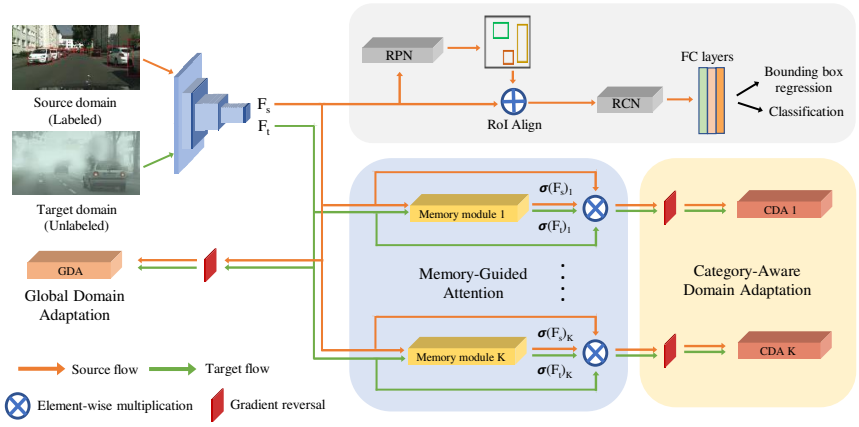

Figure 2. Overview of the proposed approach. Source and target features are aligned through global domain adaptation and category-aware domain adaptation. Global alignment is achieved by category agnostic global discriminator whereas the category-aware alignment is achieved by employing K category-specific discriminators. Since target labels are unavailable, the features to these discriminators are routed using memory-guided category-specific attention maps.

tectors to a new domain assuming the availability of large number of video data from the target domain. These video data are used to generate pseudo-labels for the target set, which are further employed to train the network. Khodabandeh *et al*. [24] formulated the domain adaptation training with noisy labels. Specifically, the model is trained on the target domain using a set of noisy bounding boxes that are obtained by a detection model trained only in the source domain. Sindagi *et al*. [46] used additional prior about weather into the domain adaptation process. Hsu *et al*. [20] proposed center-aware feature alignment to emphasize adaptation for foreground regions. Abramov *et al*. [3] proposed a simple approach by matching the image statistics like color histograms or mean/covariance between the source and target domain. He *et al*. [18] proposed an asymmetric tri-way approach to account for the differences in labeling statistics between source and target domain. Xu *et al*. [52] added a multi-label classifier as an auxiliary loss to regularize the features. However, the added loss does not pass these category-specific information to the discriminator to help perform the feature alignment. Zhao *et al*. [53] showed that using multi-label classification loss as an auxiliary loss for the domain discriminator yields better performance. Inspired from conditional adversarial networks [29], Zhao *et al*. [53] utilizes the multi-label prediction probability to perform conditional global feature alignment.

3. Proposed method

In this section, we introduce the details of our proposed method. We assume availability of fully-labeled source domain images with bounding-box annotations and unlabeled target domain images without any annotations. For rest of the paper, we denote the source domain dataset as $D_s = \{X_s^i, b_s^i, y_s^i\}_{i=1}^{N_s}$, where X_s^i denotes i^{th}-image, b_s^i and y_s^i denotes the bounding box annotations and corresponding category labels in the i^{th} source domain image. Also, each category label indicates one of K objects present in the dataset and an extra category for the background classes, i.e., $y_s^i \in \{1, 2, .., K+1\}$. Furthermore, we denote the target domain dataset as $D_t = \{X_t^i\}_{i=1}^{N_t}$, where X_t^i denotes the i^{th} target domain image. Following the previous work [6, 44, 40, 25, 24], we use Faster-RCNN [38] as our base model. We denote the backbone feature encoder of the detection network as $\mathcal{E}$. The goal of the proposed method is to utilize the source domain label information to learn a detection network that can perform well on the target domain images. To achieve that, we follow a feature alignment approach to match the distribution of features extracted by feature encoder network, for images from source and target dataset through domain adversarial training [12].

Fig. 2 presents an overview of the proposed feature alignment approach which consists of three major modules: *1) Global discriminator* that aligns the entire feature map extracted by the feature encoder network, *2) Category-wise discriminators* that focus on respective category-specific information to align features belonging to corresponding category between source and target domain. *3) Memory-guided attention mechanism* to enable the training of category-wise discriminators by generating category-specific attention on the extracted feature maps. This attention helps to focus on category information in the extracted feature map for training the respective category-wise discriminators. The attention is generated using a category-specific memory module which stores relevant information for corresponding object category. Details of these modules for feature alignment are described in the following sections.

3.1. Global discriminator for adaptation

Following the existing works [6, 40], we also employ a global discriminator to perform feature alignment of

the feature maps at image-level. The global discriminator, denoted as $\mathcal{D}_{gda}$, takes in the entire feature map extracted from the backbone network and is trained to identify whether the feature map is extracted from source or target domain. More precisely, let us denote a feature map $F_s,\ F_t \in \mathbb{R}^{C\times H\times W}$ extracted from any source and target domain image X_s and X_t, respectively. The global discriminator $\mathcal{D}_{gda}$ provides a prediction map of size $H \times W$. The discriminator network is trained with the help of least squared loss supervised with domain label $y_d \in 0, 1$. For source data, $\forall X_s \in D_s$, and target data, $\forall X_t \in D_t$ the domain labels are set to 1 and 0, respectively. The overall loss function can be written as:

$$\mathcal{L}_{gda}(X_s, X_t) = -\sum_{h=1}^{H}\sum_{w=1}^{W} y_d(1-\mathcal{D}_{gda}(F_s^{(h,w)}))^2 + (1-y_d)(\mathcal{D}_{gda}(F_t^{(h,w)}))^2, \quad (1)$$

To match the distribution of the source and the target domain features, we utilize gradient reversal layer as proposed in [12]. The gradient reversal layer flips the gradient sign before propagating the gradients back to the feature extraction network. Hence, the discriminator network $\mathcal{D}_{gda}$ is trained to minimize Eq. 1 and feature encoder network is trained to maximize Eq. 1. This adversarial training between feature extractor and discriminator helps to reduce the domain gap between source and target image features. Furthermore, instead of utilizing binary cross entropy loss for training, we utilize least-squares loss as proposed in [32] as it is shown to work better in practice and helps to stabilize the training process. However, as we argued earlier, the global adaptation is a category-agnostic approach to perform feature alignment between source and target domain. Consequently, this results in negative transfer of features and hurts the overall performance. Hence, using global discriminator alone is not optimal and requires additional strategy to avoid negative transfer of features.

3.2. Category-wise discriminators for adaptation

As discussed in the earlier sections, existing methods only consider global feature alignment strategy. In the case of object detection, each image will likely contain multiple categories and hence the feature maps extracted from these images will have features belonging to those respective categories including background features. Hence, addressing negative transfer of features between the categories while aligning source and target domain still remains an important problem in domain adaptive object detection. We address this issue by employing category-wise discriminators (CDA) that focus on aligning respective category-specific features between source and target domains. Specifically, we employ K category-wise discriminators, each focusing on aligning the respective categories. Let us denote discriminator for k^{th} category as $\mathcal{D}^k_{cda}$, F_s and F_t as the features extracted from source and target domain images X_s and X_t respectively. To align the features of the k^{th} category between the source and target domain, we generate attention maps $\sigma(F_s)_k,\ \sigma(F_t)_k \in \{0,1\}^{H\times W}$ (see Section 3.3) to focus on the information related to only k^{th} category. The loss function for category-wise adaptation for k^{th} category can be written as:

$$\mathcal{L}^k_{cda}(X_s, X_t) = -\sum_{h=1}^{H}\sum_{w=1}^{W} y_d(1-\mathcal{D}^k_{cda}(\sigma(F_s)_k^{(h,w)}\cdot F_s^{(h,w)}))^2 + (1-y_d)(\mathcal{D}^k_{cda}(\sigma(F_t)_k^{(h,w)}\cdot F_t^{(h,w)}))^2, \quad (2)$$

where, $\sigma(\cdot)_k^{(h,w)} = 1$ and $\sigma(\cdot)_k^{(h,w)} = 0$ indicate the presence and the absence, respectively, of the k^{th} category feature at location (h, w) in the corresponding feature map (F_s or F_t), respectively. The major challenge in training with these category-wise discriminator is lack of information regarding the location of category in the feature maps, especially for the target domain data. To this end, we propose a mechanism to predict the attention maps indicating locations of each category with the help of a memory module.

3.3. Memory-guided attention mechanism

We propose memory-guided attention (MeGA) mechanism to aid the category-wise discriminators in aligning the category-specific features between the source and target domains. Specifically, we employ K memory modules corresponding to the K categories. These memory modules are used to store the class-prototypes of different objects during the training process, so that they can be retrieved for computing the category-specific attention maps. Next, we describe the details regarding memory updates and attention computation.

3.3.1 Memory module

A memory module has two operations, namely write and read. To write in to the memory, features extracted from the neural network are used to update the memory elements appropriately. Whereas, the memory read operation is used by the features extracted from the neural network to query the memory and retrieve the most similar memory element (or prototypical feature). These operations are illustrated in Fig. 3. In the proposed approach, we learn K memory modules, i.e. $M_k \in \mathbb{R}^{N_m\times C}$, corresponding to K categories of the source and target domain. Here, N_m are the number of memory items per category and C are the number of channels in the feature map.

Memory write. To update the memory elements, we consider only source domain images as we have access to the bounding box labels to locate the category-specific features in the extracted feature-map F_s. For brevity, let us denote

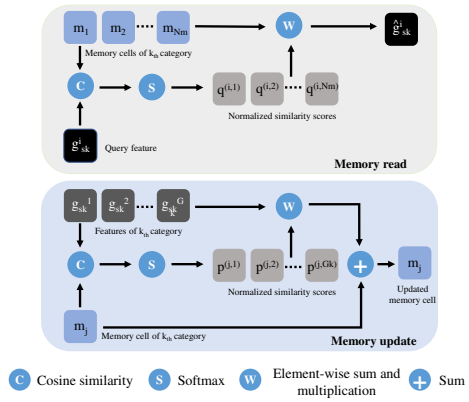

Figure 3. Read and write operations for the memory module.

$G_k = \{g^i_{s_k} \in \mathbb{R}^{1\times C}\}_{i=1}^{N_{s_k}}$ all the features belonging to k^{th} category in the feature-map F_s. Also, each memory element in memory module M_k is denoted as $m_j \in \mathbb{R}^{1\times C}$, where $j \in \{1, ..., N_m\}$. First, we compute the normalized similarity metric between the memory elements in M_k and the set of features G_k representing k^{th} category as:

$$p^{(j,i)} = \frac{\exp\left(m_j \cdot g^i_{s_k}\right)}{\sum_{l\in G_k} \exp\left(m_j \cdot g^l_{s_k}\right)}, \tag{3}$$

where, p is an $N_m \times N_k$ similarity map. We utilize this similarity between memory elements and category features to update each memory element using following equation:

$$m^j \leftarrow m^j + \sum_{i\in G_k} p^{(i,j)} g^i_{s_k}. \tag{4}$$

Also, note that if the k^{th} category is not present in the source image, we do not update the elements of the respective memory module M_k. Following [34], we further regularize the features by making sure that the memory elements should not be too far from the original features. This regularization encourages compactness in the memory module, which reduces intra-class variations. This loss is formulated in the form of L2 distance penalty as:

$$\mathcal{L}_{cmp} = \sum_{j=1}^{N_m} \left\| m_j - g^p_{s_k} \right\|_2, \tag{5}$$

where, $g^p_{s_k}$ is a function of m_j and denotes the most similar feature in the set G_k to the memory element m_j. In addition to regularizing the memory to be more compact, we enforce a uniqueness constraint to reduce the redundancy in the memory element. Following [34], we utilize a triplet loss on the memory elements such that each memory element in the memory module M_k represents unique prototype of the underlying category. This loss can be expressed as follows:

$$\mathcal{L}_{unq} = \sum_{j=1}^{N_m} max(\left\| m_j - g^p_{s_k} \right\|_2 - \left\| m_j - g^n_{s_k} \right\|_2, \alpha), \tag{6}$$

where, α denotes the triplet loss margin, $g^p_{s_k}$ and $g^n_{s_k}$ denotes respectively the most similar and the second most similar in the feature set G_k. Given these constraints, the overall loss for the memory can be defined as:

$$\mathcal{L}_{mem} = \mathcal{L}_{cmp} + \mathcal{L}_{unq}. \tag{7}$$

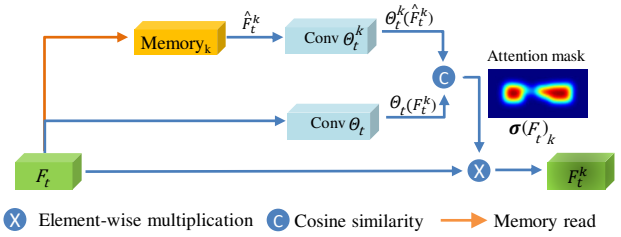

Figure 4. The feed-forward path for Memory-guided Attention (MeGA) mechanism. Input source/target feature map is queried to any k^{th}-category memory module. Through read operation closest matching elements are retrieved and used to predict attention map through learned similarity. Attention map is used to route the k-category information to the k^{th} category discriminator module.

Memory read. To retrieve the most similar memory element, we compute the similarity between each item in the memory M_k and the given query feature. Note that the query feature can be either from the source domain or target domain image, i.e. $g^i_{s_k}$ or $g^i_{t_k}$. For $g^i_{s_k}$ the normalized similarity is computed using the following equations:

$$q^{(i,j)} = \frac{\exp\left(m_j \cdot g^i_{s_k}\right)}{\sum_{l\in N_m} \exp\left(m_l \cdot g^i_{s_k}\right)}. \tag{8}$$

Given this normalized similarity q, the retrieved feature $\hat{g}^i_{s_k} \in \mathbb{R}^{1\times C}$ can be expressed as follows:

$$\hat{g}^i_{s_k} = \sum_{j\in N_m} q^{(i,j)} m_j. \tag{9}$$

Also, note that we use the same formulation to read from the memory for both the source and target domain features.

3.3.2 Attention mechanism

We utilize all the memory modules to get attention maps for category-wise discriminators. Specifically, to compute an attention map for the target feature-map F_t, we query each element $f_t \in \mathbb{R}^{1\times C}$ to the k^{th} memory module M_k and retrieve a vector $\hat{f}_t \in \mathbb{R}^{1\times C}$ to get a retrieved feature map $\hat{F}^k_t \in \mathbb{R}^{C\times H\times W}$. We compute element-wise similarity between the extracted feature map F_t and the retrieved feature map $\hat{F}^k_t$ to get the attention map for the k^{th} category-wise discriminator $\sigma(F_t)_k$. We explore two choices of similarity function to obtain the attention map.

Cosine similarity. The most commonly used similarity function in the literature is cosine similarity. We compute the element-wise cosine similarity to get $\sigma(F_t)_k$ of size $H \times W$. It can be expressed as:

$$\sigma(F_t)_k^{(h,\,w)} = \frac{F_t^{(h,w)} (\hat{F}_t^{k\ (h,w)})^T}{||F_t^{(h,w)}||_2 ||\hat{F}_t^{k\ (h,w)}||_2}, \tag{10}$$

Table 1. Quantitative results (mAP) for Cityscapes → Foggy-Cityscapes dataset.

Method	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
Source Only	25.8	33.7	35.2	13.0	28.2	9.1	18.7	31.4	24.4
DAFaster [6]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
Strong-Weak [40]	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
MAF [17]	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
D&Match [25]	30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
Selective DA [55]	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
MTOR [5]	30.6	41.4	44.0	21.9	38.6	40.6	28.3	35.6	35.1
ICR-CCR [52]	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
ATF [18]	34.6	47.0	50.0	23.7	43.3	38.7	33.4	38.8	38.7
MCAR [53]	32.0	42.1	43.9	**31.3**	44.1	43.4	**37.4**	36.6	38.8
Prior DA [46]	36.4	47.3	51.7	22.8	47.6	34.1	36.0	38.7	39.3
MeGA-CDA (ours)	**37.7**	**49.0**	**52.4**	25.4	**49.2**	**46.9**	34.5	**39.0**	**41.8**
Oracle [38]	37.2	48.2	52.7	35.2	52.2	48.5	35.3	38.8	43.5

Learned similarity. While the use of cosine similarity to compute the attention maps results in reasonable improvements in accuracy, a closer look at these maps (see Fig. 6 top-row) reveals that the attention generated using cosine similarity is not accurate. To overcome this issue, we explore a similarity metric which is parameterized with a neural network and can be learned during training. In this case, we utilize a metric learning approach where both F_t and $\hat{F}_t^k$ are first passed through a network respectively, Θ_t and Θ_t^k. To supervise the network we utilize the bounding box information available in the source dataset. In particular, we maximize the cosine similarity between $\Theta_t(F_t)^{(h,w)}$ and $\Theta_t^k(\hat{F}_t^k)^{(h,w)}$ for the location where the category k is present and minimize the similarity where there is an absence of the corresponding category as shown in Fig. 4. Then, the attention map can be expressed as:

$$\sigma(F_t)_k \;=\; Sim(\Theta_t(F_t),\; \Theta_t^k(\hat{F}_t^k)), \tag{11}$$

where, $Sim(x, y)$ indicates element-wise cosine similarity between tensor x and y of size $C \times H \times W$, similar to Eq. 10. The resulting attention $\sigma(\cdot)_k$ is of size $H \times W$. We compute this attention for both source and target images and for all K categories. Before forwarding the attention into the subsequent discriminators, we binarize it with threshold 0.5 *i.e.*, if the normalized similarity is greater than 0.5 we assign 1 to the map and vice versa.

3.4. Overall training objective for MeGA-CDA

For our final model training, we add supervised detection loss on the source domain data, which has both images and corresponding bonding box annotations with category labels as described in Sec. 3. We denote the supervised detection loss as $\mathcal{L}_{det}$, which includes both bounding box regression loss and classification loss as described in [38]. Including the global, category-wise and memory loss as described in the previous sections, the overall training objective of the proposed method can be expressed as:

$$\mathcal{L}_{cda}^{mega} \;=\; \mathcal{L}_{det}(X_s,\; b_s,\; y_s) \;+\; \beta\, \mathcal{L}_{gda}(X_s,\; X_t) \;+\; \gamma \sum_{k=1}^{K} \mathcal{L}_{cda}^{k}(X_s,\; X_t) \;+\; \lambda\, \mathcal{L}_{mem}, \tag{12}$$

where, β, γ and λ are parameters used to weight the global, category-wise and memory loss, respectively.

4. Experiments and results

4.1. Implementation details

We adopt Faster-RCNN [38] network with VGG16 backbone and train using SGD optimizer with learning rate of 0.002 and momentum 0.9 for 6 epochs and then decrease the learning rate to 0.0002. Global and Category-wise discriminators consist of four convolution layers with ReLU non-linearity[1]. The batch size is set to 2 with each batch containing one image from source domain and one from target. We use 20 memory items for each category and each memory item has a dimension of $1 \times 1 \times C$, where C denotes the number of channels in the corresponding feature map. The networks Θ_t, Θ_t^k also consist of 4 convolution layers with ReLU non-linearity. We train the network for 10 epochs and report the mean average precision (mAP) with a threshold of 0.5. The weight of the memory loss, λ and of domain adaptors, β, γ are empirically set equal to 0.1 and 0.01, respectively.

4.2. Quantitative comparison

In this section, we compare the performance of the proposed method with recent state-of-the-art approaches under three broad categories of adaptation: (i) adverse weather, (ii) synthetic-to-real adaptation, and (iii) cross-camera adaptation.

[1]Details of the architecture are included in supplementary material

4.2.1 Adverse weather conditions

Stable object detection performance in different weather conditions is critical for safety critical applications like self-driving cars. Weather conditions introduce image artifacts which can negatively impact the detection performance. To evaluate the effectiveness of proposed method in adverse weather, we utilize Foggy-Cityscapes and Cityscapes as target and source domain respectively.

Dataset: The Cityscapes dataset [7] is collected under clear weather conditions and Foggy-Cityscapes [42] is created by simulating haze on top of the Cityscapes images. Both Cityscapes and Foggy-Cityscapes have 2975 training images and 500 validation images with 8 object categories: *person, rider, car, truck, bus, train, motorcycle and bicycle.*

Results: In Table 1, we report the performance of our framework MeGA-CDA and compare with recent adaptive object detection methods. As it can be observed, MeGA-CDA, outperforms existing approaches by considerable margin, while improving over the recent best method by an average (absolute) mAP of 2.5%. Moreover, the proposed method performs consistently well across all categories, demonstrating the benefits of incorporating category-wise alignment along with global alignment of the features.

4.2.2 Synthetic data adaptation

Synthetic data offers an inexpensive alternative to real data collection as it is easier to collect and with appropriate engineering, the synthetic data can be auto-annotated. In spite of the advancements in computer graphics, photo-realistic synthetic data generated using state-of-the-art rendering engines suffer from subtle image artifacts which can result in sub-optimal performance on real-world data.

Dataset: In this experiment, Sim10k [22] is used as the source domain and Cityscapes as the target domain. Sim10k has 10,000 images with 58,701 bounding boxes of *car* category, rendered by the gaming engine *Grand Theft Auto*. We use all the Sim10k images for training and evaluate on the bounding boxes of the *car* category from the 500 images of Cityscapes validation set.

Results: In Table 2, we report the mAP of our framework trained using the Sim10K synthetic data as source and Cityscapes as target. The proposed method, MeGA-CDA, improves upon the recent best method by 1.8% mAP (absolute improvement). Since we are adapting from synthetic to real scenario, we observed better alignment when we adapted the features of the third conv layer as well. Considering that this experiment has only one category of objects, the improvements achieved by the proposed category-aware alignment demonstrates that the memory-guided attention ensures better alignment across the positive and negative (background) class of objects.

Table 2. Quantitative results (mAP) for Sim10K → Cityscapes.

Method	mAP
Source Only	34.3
DAFaster [6]	38.9
MAF [17]	41.1
Strong-Weak [40]	40.1
ATF [18]	42.8
Selective DA [55]	43.0
MeGA-CDA (ours)	**44.8**
Oracle [38]	62.7

4.2.3 Cross-camera adaptation

Differences in the intrinsic and extrinsic camera properties like resolution, distortion, orientation, location result in images which capture the objects differently from each other in terms of quality, scale and viewing angle. While the collected data can be real, these domain differences will potentially result in severe performance degradation.

Dataset: To study this effect of cross-camera domain gap, we conduct two adapataion experiments involving KITTI [13] and Cityscapes datasets. In the first experiment, we adapt from KITTI to Cityscapes, where as in the second experiment, we adapt from Cityscapes to KITTI. Note that the KITTI dataset consists of 7,481 images,

Results: The results of these two experiments are presented in Table 3. In both the experiments, proposed method is able to achieve considerable improvements over the recent best methods. From these results, one may observe that proposed memory-guided category alignment is effective in bridging the domain gap across different camera views and optical properties.

Table 3. Quantitative results (mAP) for KITTI → Cityscapes and Cityscapes → KITTI datasets.

Method	KITTI → City	City → KITTI
Source Only	30.2	53.5
DAFaster [6]	38.5	64.1
MAF [17]	41.0	72.1
Strong-Weak [40]	37.9	71.0
Selective DA [55]	42.5	-
ATF [18]	42.1	73.5
MeGA-CDA (ours)	**43.0**	**75.5**

4.3. Ablation studies

We study the impact of different components of the proposed method, MeGA-CDA, by iteratively adding each module. We use the Cityscape→Foggy-Cityscape adaptation experiment for these ablations.

Quantitative analysis: The results corresponding to ablation analysis are reported in Table 4. We observe reasonable improvement over the source only baseline by adapting the

Table 4. Ablation study on Foggy-Cityscapes. C4 and C5 indicate the adaptation loss at fourth and fifth convolutional block respectively in VGG16 backbone.

Method	C5	C4	prsn	rider	car	truc	bus	train	mcycle	bcycle	mAP
Source Only			25.8	33.7	35.2	13.0	28.2	9.1	18.7	31.4	24.4
GDA	✓		35.3	44.2	51.0	23.1	44.3	28.1	27.8	37.7	36.2
GDA+CDA+MA	✓		34.5	45.1	50.4	23.8	45.6	27.9	29.6	37.5	36.8
	✓	✓	**37.8**	47.1	**52.4**	**29.1**	48.8	29.0	**36.7**	**39.0**	40.0
GDA+CDA+MA+LS (MeGA-CDA)	✓		35.9	43.7	50.8	23.4	46.5	**48.7**	25.0	37.1	38.9
	✓	✓	37.7	**49.0**	**52.4**	25.4	**49.2**	46.9	34.5	**39.0**	**41.8**

conv5 features with a global domain discriminator (GDA). By augmenting the global domain discriminators with the proposed memory-guided category-wise discriminators (GDA+CDA+MA) trained with the cosine similarity-based attention, we obtain a further improvement of 0.6% mAP. This illustrates the benefit of adding category-wise information during domain adaptation. When GDA+CDA+MA is applied at multiple layers (both conv4 and conv5), we observe further improvement of approximately 3%. Finally, we demonstrate that strengthening the memory module with a metric learned similarity (LS) approach (MeGA-CDA) enhances the capability of the memory banks in capturing the data characteristics and results in further improvements. Specifically, when MeGA-CDA is applied at conv5, we observe an improvement of 2.1% as compared to GDA+CDA+MA baseline with cosine similarity. Additionally, applying MeGA-CDA at both conv4 and conv5 blocks results in an additional improvement of 2%. As discussed previously, this sub-network is trained using weak supervision from the ground truth bounding boxes of the source domain and hence, it does not require any additional annotations.

Qualitative analysis: We compare the detections of global alignment approach with proposed category-wise alignment in Fig. 5 for the Cityscapes→Foggy-Cityscapes adaptation experiment. As we can see from Fig. 5, global alignment-based approach results in errors such as missed-detections (false negatives) or false positives. For example, background is detected as an object (bottom row) or an object is mistakenly assigned wrong category and bounding box size (top row). The most likely reason for this is negative transfer of features, as global adaptation aligns the feature in category agnostic way. In both cases, the proposed category-wise alignment is able to rectify the error by better countering the negative transfer of features. In Fig. 6, we show attention maps generated for the car category during MeGA-CDA training. For visualization, we overlay the attention maps, generated by the memory module, on the images. The top row and bottom rows show attention maps computed using cosine similarity and metric learning-based similarity respectively. It can be observed that the cosine similarity-based attention provides reasonable focus on the car category locations. However, with the learned similarity we achieve more effectiveness where the attention spans

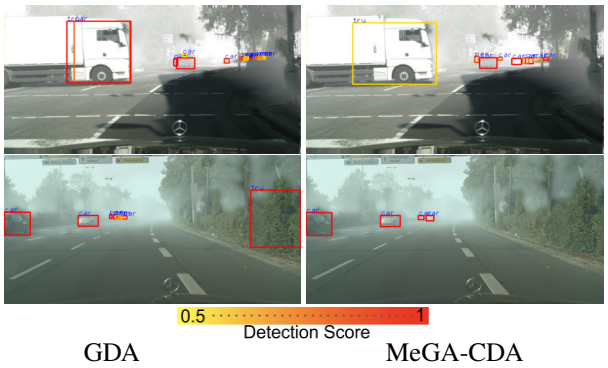

Figure 5. Qualitative detection results. Global alignment results in miss-detections. In contrast, the proposed approach reduces false-positives while achieving high-quality detections.

Figure 6. Comparison of attention maps computed using cosine similarity (top-row) and learned similarity based attention (bottom-row). Though cosine similarity based provides reasonable focus on category features, learned similarity obtains more accurate attention.

majority of the car region. This is expected as learned similarity is trained through metric learning with weak supervision from the source domain ground truth and thus results in guided learning of memory items as well.

5. Conclusions

We presented a category-aware feature alignment approach for domain adaptive object detection. Specifically, we incorporate category information into the domain alignment process by introducing category-aware discriminators. To overcome the issue of lack of category labels, especially in target domain, we propose memory-guided attention mechanism that generate category-specific attention maps for routing the features into the appropriate category-specific discriminator. By doing so, we are able to mitigate the problem of negative transfer, thereby resulting in better overall alignment. MeGA-CDA is evaluated on several benchmark datasets and is shown to outperform existing approaches by a considerable margin.

Acknowledgement

This work was supported by the NSF grant 1910141 and Mercedes-Benz Research & Development India.

References

[1] Mahdi Abavisani and Vishal M Patel. Domain adaptive subspace clustering. In *27th British Machine Vision Conference, BMVC 2016*, 2016.

[2] Mahdi Abavisani and Vishal M Patel. Adversarial domain adaptive subspace clustering. In *2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA)*, pages 1–8. IEEE, 2018.

[3] Alexey Abramov, Christopher Bayer, and Claudio Heller. Keep it simple: Image statistics matching for domain adaptation. *arXiv preprint arXiv:2005.12551*, 2020.

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[5] 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 Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019.

[6] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2018.

[7] 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*, pages 3213–3223, 2016.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[10] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1999–2007, 2019.

[11] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[15] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[16] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1705–1714, 2019.

[17] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6668–6677, 2019.

[18] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In *Proceedings of the European Conference on Computer Vision*, 2020.

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

[20] 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.

[21] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Duplex generative adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1498–1507, 2018.

[22] 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? *arXiv preprint arXiv:1610.01983*, 2016.

[23] Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. *arXiv preprint arXiv:1703.03129*, 2017.

[24] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. *arXiv preprint arXiv:1904.02361*, 2019.

[25] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019.

[26] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*, pages 1378–1387, 2016.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[28] 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.

[29] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018.

[30] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016.

[31] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2208–2217. JMLR. org, 2017.

[32] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[33] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018.

[34] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020.

[35] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[36] Pramuditha Perera, Mahdi Abavisani, and Vishal M Patel. In2i: Unsupervised multi-image-to-image translation using generative adversarial networks. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 140–146. IEEE, 2018.

[37] 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.

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[39] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 780–790, 2019.

[40] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. *CoRR*, abs/1812.04798, 2018.

[41] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018.

[42] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018.

[43] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018.

[44] Yuhu Shan, Wen Feng Lu, and Chee Meng Chew. Pixel and feature level based domain adaptation for object detection in autonomous driving. *Neurocomputing*, 367:31–38, 2019.

[45] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018.

[46] 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.

[47] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.

[48] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[49] Paul Viola, Michael Jones, et al. Rapid object detection using a boosted cascade of simple features. *CVPR (1)*, 1:511–518, 2001.

[50] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[51] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[52] 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*, pages 11724–11733, 2020.

[53] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye. Adaptive object detection with dual multi-label prediction. *arXiv preprint arXiv:2003.12943*, 2020.

[54] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[55] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 687–696, 2019.