# MIC: Masked Image Consistency for Context-Enhanced Domain Adaptation

Lukas Hoyer [1] Dengxin Dai [2] Haoran Wang [2] Luc Van Gool [1,3]
[1] ETH Zurich [2] Max Planck Institute for Informatics, Saarland Informatics Campus [3] KU Leuven
{lhoyer,vangool}@vision.ee.ethz.ch, {ddai,hawang}@mpi-inf.mpg.de

## Abstract

*In unsupervised domain adaptation (UDA), a model trained on source data (e.g. synthetic) is adapted to target data (e.g. real-world) without access to target annotation. Most previous UDA methods struggle with classes that have a similar visual appearance on the target domain as no ground truth is available to learn the slight appearance differences. To address this problem, we propose a Masked Image Consistency (MIC) module to enhance UDA by learning spatial context relations of the target domain as additional clues for robust visual recognition. MIC enforces the consistency between predictions of masked target images, where random patches are withheld, and pseudo-labels that are generated based on the complete image by an exponential moving average teacher. To minimize the consistency loss, the network has to learn to infer the predictions of the masked regions from their context. Due to its simple and universal concept, MIC can be integrated into various UDA methods across different visual recognition tasks such as image classification, semantic segmentation, and object detection. MIC significantly improves the state-of-the-art performance across the different recognition tasks for synthetic-to-real, day-to-nighttime, and clear-to-adverse-weather UDA. For instance, MIC achieves an unprecedented UDA performance of 75.9 mIoU and 92.8% on GTA→Cityscapes and VisDA-2017, respectively, which corresponds to an improvement of +2.1 and +3.0 percent points over the previous state of the art. The implementation is available at https://github.com/lhoyer/MIC.*

## 1. Introduction

In order to train state-of-the-art neural networks for visual recognition tasks, large-scale annotated datasets are necessary. However, the collection and annotation process can be very time-consuming and tedious. For instance, the annotation of a single image for semantic segmentation can take more than one hour [10,66]. Therefore, it would be beneficial to resort to existing or simulated datasets, which are easier to annotate. However, a network trained on such a source dataset usually performs worse when applied to the actual target dataset as neural networks are sensitive to domain gaps. To mitigate this issue, unsupervised domain adaptation (UDA) methods adapt the network to the target domain using unlabeled target images, for instance, with adversarial training [20,27,57,73] or self-training [30,31,72,79,97].

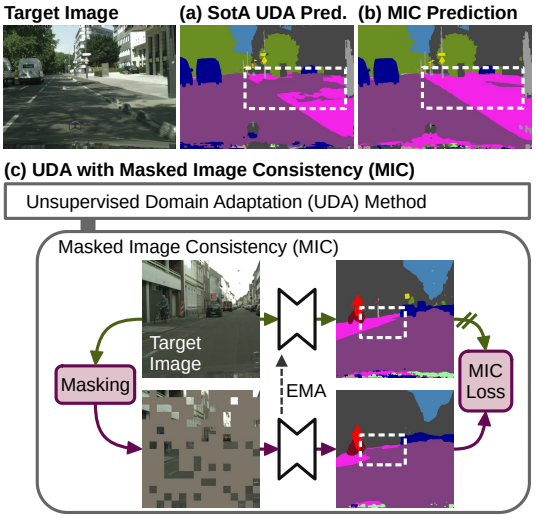



Figure 1. (a) Previous UDA methods such as HRDA [31] struggle with similarly looking classes on the unlabeled target domain. Here, the interior of the *sidewalk* is wrongly segmented as *road*, probably, due to the ambiguous local appearance. (b) The proposed Masked Image Consistency (MIC) enhances the learning of context relations to consider additional context clues such as the curb in the foreground. With MIC, the adapted network is able to correctly segment the *sidewalk*. (c) MIC can be plugged into most existing UDA methods. It enforces the consistency of the predictions of a masked target image with the pseudo-label of the original image. So, the network is trained to better utilize context clues on the target domain. Further details are shown in Fig. 3.

UDA methods have remarkably progressed in the last few years. However, there is still a noticeable performance gap compared to supervised training. A common problem is the confusion of classes with a similar visual appearance on the target domain such as *road*/*sidewalk* or *pedestrian*/*rider* as

there is no ground truth supervision available to learn the slight appearance differences. For example, the interior of the *sidewalk* in Fig. 1 is segmented as *road*, probably, due to a similar local appearance. To address this problem, we propose to enhance UDA with spatial context relations as additional clues for robust visual recognition. For instance, the curb in the foreground of Fig. 1 a) could be a crucial context clue to correctly recognize the *sidewalk* despite the ambiguous texture. Although the used network architectures already have the capability to model context relations, previous UDA methods are still not able to reach the full potential of using context dependencies on the target domain as the used unsupervised target losses are not powerful enough to enable effective learning of such information.

Therefore, we design a method to explicitly encourage the network to learn comprehensive context relations of the target domain during UDA. In particular, we propose a novel Masked Image Consistency (MIC) plug-in for UDA (see Fig. 1 c), which can be applied to various visual recognition tasks. Considering semantic segmentation for illustration, MIC masks out a random selection of target image patches and trains the network to predict the semantic segmentation result of the entire image including the masked-out parts. In that way, the network has to utilize the context to infer the semantics of the masked regions. As there are no ground truth labels for the target domain, we resort to pseudo-labels, generated by an EMA teacher that uses the original, unmasked target images as input. Therefore, the teacher can utilize both context and local clues to generate robust pseudo-labels. Over the course of the training, different parts of objects are masked out so that the network learns to utilize different context clues, which further increases the robustness. After UDA with MIC, the network is able to better exploit context clues and succeeds in correctly segmenting difficult areas that rely on context clues such as the *sidewalk* in Fig. 1 b).

To the best of our knowledge, MIC is the first UDA approach to exploit masked images to facilitate learning context relations on the target domain. Due to its universality and simplicity, MIC can be straightforwardly integrated into various UDA methods across different visual recognition tasks, making it highly valuable in practice. MIC achieves significant and consistent performance improvements for different UDA methods (including adversarial training, entropy-minimization, and self-training) on multiple visual recognition tasks (image classification, semantic segmentation, and object detection) with different domain gaps (synthetic-to-real, clear-to-adverse-weather, and day-to-night) and different network architectures (CNNs and Transformer). It sets a new state-of-the-art performance on all tested benchmarks with significant improvements over previous methods as shown in Fig. 2. For instance, MIC respectively improves the state-of-the-art performance by +2.1, +4.3, and +3.0 percent points on GTA→Cityscapes(CS), CS→DarkZurich, and

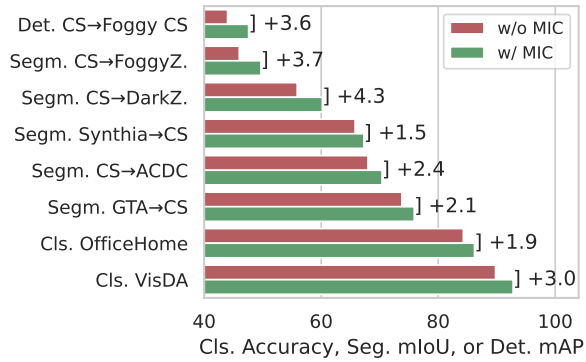



Figure 2. MIC significantly improves state-of-the-art UDA methods across different UDA benchmarks and recognition tasks such as image classification (Cls.), semantic segmentation (Segm.), and object detection (Det.). Detailed results can be found in Sec. 4.

VisDA-2017 and achieves an unprecedented UDA performance of 75.9 mIoU, 60.2 mIoU, and 92.8%, respectively.

## 2. Related Work

### 2.1. Unsupervised Domain Adaptation (UDA)

In UDA, a model trained on a labeled source domain is adapted to an unlabeled target domain. Due to the ubiquity of domain gaps, UDA methods were designed for all major computer vision problems including image classification [19, 47, 48, 55], semantic segmentation [28, 30, 73, 89], and object detection [7, 8, 42, 44]. The majority of the approaches rely on discrepancy minimization, adversarial training, or self-training. The first group minimizes the discrepancy between domains using a statistical distance function such as maximum mean discrepancy [24, 47, 50], correlation alignment [68, 69], or entropy minimization [23, 49, 76]. In adversarial training, a learned domain discriminator provides supervision in a GAN framework [22] to encourage domain-invariant inputs [21, 27], features [19, 28, 48, 73] or outputs [52, 62, 73, 76]. In self-training, pseudo-labels [39] are generated for the target domain based on predictions obtained using confidence thresholds [53, 91, 97] or pseudo-label prototypes [55, 89, 90]. To increase the robustness of the self-training, consistency regularization [63, 67, 71] is often applied to ensure consistency over different data augmentations [1, 9, 18, 54], different crops [31, 38], multiple models [88, 93, 94], or domain-mixup [30–32, 72, 95]. Further UDA strategies utilize pretext tasks [6, 32, 77, 79], follow an adaptation curriculum [11, 12, 92], exploit the increased domain-robustness of Transformers [30, 31, 70, 85], align the domains with contrastive learning [34, 82], use graph matching [4, 41, 42], or adapt multi-resolution inputs [31].

To facilitate learning domain-robust context dependencies, several UDA methods propose network components

that can capture context such as spatial attention pyramids [40], cross-domain attention [86], or context-aware feature fusion [30]. While these network modules provide the ability to capture context, the unsupervised loss on the target domain does not provide sufficient supervision to learn all relevant target context relations. To improve context learning, CrDA [35] aligns local context relations with adversarial training and HRDA [31] uses multi-crop consistency training. However, these mechanisms are not able to capture all relevant context clues as can be seen for HRDA in Fig. 1 a). Due to the random patch masking, MIC is able to learn a larger set of different context clues for robust recognition.

### 2.2. Masked Image Modeling

Predicting withheld tokens of a masked input sequence was shown to be a powerful self-supervised pretraining task in natural language processing [3, 14]. Recently, this concept was successfully transferred to self-supervised pretraining in computer vision, where it is known as masked image modeling. Given a partly masked image, the network is trained to reconstruct properties of the masked areas such as VAE features [2, 15, 43], HOG features [80], or color information [25, 84]. To sample the mask, block-wise masking [2], random patch masking [25, 84], and attention-guided masking [37, 45] have been explored.

Similarly, our method also uses masked images. However, we pursue a different purpose than previous works. Instead of aiming to learn self-supervised representations, MIC utilizes masked images in a novel way to learn context relations for domain adaptation. Due to this conceptual difference, we do not have to rely on pretext restoration targets such as VAE features but can perform the reconstruction in the actual prediction space of the relevant computer vision task such as semantic segmentation. To the best of our knowledge, MIC is the first method to exploit masked images to enhance context learning for UDA. Particularly, we show that naive masked image modeling on ImageNet does not improve the target domain performance (see Sec. 4.3).

## 3. Methods

### 3.1. Unsupervised Domain Adaptation (UDA)

A neural network $f_\theta$ can be trained on the source domain using images $\mathcal{X}^S = \{x_k^S\}_{k=1}^{N_S}$ and their labels $\mathcal{Y}^S = \{y_k^S\}_{k=1}^{N_S}$ with a supervised source loss $\mathcal{L}^S$. The specific source loss depends on the computer vision task. For image classification and semantic segmentation, the (pixel-wise) cross-entropy is typically used

$$\mathcal{L}_k^{S,cls/seg} = \mathcal{H}(f_\theta(x_k^S), y_k^S)\,, \tag{1}$$

$$\mathcal{H}(\hat{y}, y) = -\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{C} y_{ijc}\log\hat{y}_{ijc}\,, \tag{2}$$

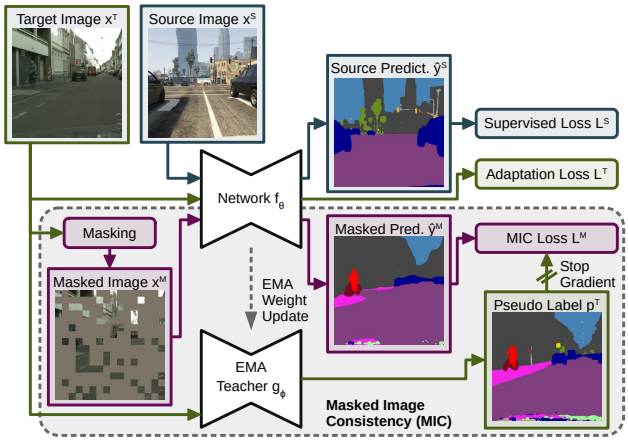



Figure 3. UDA with the proposed Masked Image Consistency (MIC). In UDA, a network is typically trained with a supervised loss on the source domain (blue) and an unsupervised adaptation loss on the target domain (green). MIC enforces the consistency between predictions of masked target images (purple) and pseudo-labels that are generated based on the complete image by an exponential moving average (EMA) teacher. To minimize the MIC loss, the network has to learn to infer the predictions of the masked regions from their context.

where $H{=}W{=}1$ in case of classification. For object detection, a box regression and a box classification loss are commonly utilized [58].

However, a model trained on the source domain usually experiences a performance drop when applied to another domain. Therefore, unsupervised domain adaptation (UDA) methods use unlabeled images from the target domain $\mathcal{X}^T = \{x_k^T\}_{k=1}^{N_T}$ to adapt the network. For that purpose, an additional unsupervised loss for the target domain $\mathcal{L}^T$ is added to the optimization problem with a weight $\lambda^T$

$$\min_\theta \frac{1}{N_S}\sum_{k=1}^{N_S}\mathcal{L}_k^S + \frac{1}{N_T}\sum_{k=1}^{N_T}\lambda^T\mathcal{L}_k^T\,. \tag{3}$$

The target loss $\mathcal{L}^T$ is defined according to the UDA strategy such as adversarial training [8, 19, 57, 73, 74, 78] or self-training [30, 53, 72, 89, 90, 97].

### 3.2. Masked Image Consistency (MIC)

To recognize an object (or stuff region), a model can utilize clues from different parts of the image. This can be local information, which originates from the same image patch as the corresponding cell in the feature map, or context information, which comes from surrounding image patches that can belong to different parts of the object or its environment [33]. Many network architectures [16, 26] have the capability to integrate both local and context information in their features. While the learning of context clues can be guided by ground truth in supervised learning, there

is no ground truth supervision available for the target domain in UDA. Current unsupervised losses are not powerful enough to enable effective learning of context clues as empirically observed such as in Fig. 1 a). Therefore, we propose to specifically encourage the learning of context relations on the target domain to provide additional clues for robust recognition of classes with similar local appearances.

In order to facilitate the learning of context relations on the target domain, we introduce a Masked Image Consistency (MIC) module, which can be easily plugged into various existing UDA methods. The domain adaptation process with MIC is illustrated in Fig. 3 and explained below.

MIC withholds local information by randomly masking out patches of the target image. For that purpose, a patch mask $\mathcal{M}$ is randomly sampled from a uniform distribution

$$\mathcal{M}_{\substack{mb+1:(m+1)b, \\ nb+1:(n+1)b}} = [v > r] \quad \text{with } v \sim \mathcal{U}(0,1)\,, \tag{4}$$

where $[\cdot]$ denotes the Iverson bracket, $b$ the patch size, $r$ the mask ratio, and $m \in [0\,..\,W/b-1]$, $n \in [0\,..\,W/b-1]$ the patch indices. The masked target image $x^M$ (see Fig. 3) is obtained by element-wise multiplication of mask and image

$$x^M = \mathcal{M} \odot x^T\,. \tag{5}$$

The masked target prediction $\hat{y}^M$ can only use the limited information of the unmasked image regions

$$\hat{y}^M = f_\theta(x^M)\,, \tag{6}$$

making the prediction more difficult. This is also reflected in Fig. 3, where the prediction misses a part of the sidewalk. In order to train the network to use the remaining context clues to still reconstruct the correct label without access to the entire image, the MIC loss $\mathcal{L}^M$ is introduced

$$\mathcal{L}^M = q^T \mathcal{H}(\hat{y}^M, p^T)\,, \tag{7}$$

where $p^T$ denotes a pseudo-label and $q^T$ its quality weight. MIC uses pseudo-labels as there is no ground truth available for the target domain. The pseudo-label is the prediction of a teacher network $g_\phi$ of the complete target image $x^T$. For image classification and semantic segmentation,

$$p_{ij}^{T,cls/seg} = [c = \arg\max_{c'} g_\phi(x^T)_{ijc'}]\,. \tag{8}$$

For object detection pseudo-labels, box predictions from $g_\phi(x^T)$ are filtered with a confidence threshold $\delta$ and non-maximum suppression [58].

The teacher network $g_\phi$ is implemented as an EMA teacher [71]. Its weights are the exponential moving average of the weights of $f_\theta$ with smoothing factor $\alpha$

$$\phi_{t+1} \leftarrow \alpha\phi_t + (1-\alpha)\theta_t\,, \tag{9}$$

where $t$ denotes a training step. The EMA teacher realizes a temporal ensemble of previous student models $f_\theta$ [71], which increases the robustness and temporal stability of pseudo-labels. It is a common strategy used in semi-supervised learning [17,29,71] and UDA [1,30,31,72]. As the teacher is updated based on the student $f_\theta$, it will gradually obtain the enhanced context learning capability from $f_\theta$. In contrast to the student $f_\theta$, the teacher $g_\phi$ has privileged access to the original image $x^T$ (see Eq. 8), so that it can use both the context and the intact local appearance information to generate pseudo labels of higher quality.

As the pseudo-labels are potentially wrong (especially at the beginning of the training), the loss is weighted by the quality estimate $q^T$. For image classification, we use the maximum softmax probability as certainty estimate [91]

$$q^{T,cls} = \max_{c'} g_\phi(x^T)_{c'}\,. \tag{10}$$

For semantic segmentation, we follow [30,31,72] and utilize the ratio of pixels exceeding a threshold $\tau$ of the maximum softmax probability

$$q^{T,seg} = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}[\max_{c'} g_\phi(x^T)_{(ijc')} > \tau]}{H \cdot W}\,. \tag{11}$$

And for object detection, we apply the quality estimate from Eq. 10 to each bounding box in the classification branches.

The MIC consistency training can be easily integrated into the UDA optimization problem

$$\min_\theta \frac{1}{N_S}\sum_{k=1}^{N_S}\mathcal{L}_k^S + \frac{1}{N_T}\sum_{k=1}^{N_T}(\lambda^T\mathcal{L}_k^T + \lambda^M\mathcal{L}_k^M)\,. \tag{12}$$

# 4. Experiments

## 4.1. Implementation Details

**Semantic Segmentation:** We study synthetic-to-real, clear-to-adverse-weather, and day-to-nighttime adaptation of street scenes. As synthetic datasets, we use GTA [59] containing 24,966 images and Synthia [60] with 9,400 images. As real-world datasets, we use Cityscapes (CS) [10] consisting of 2,975 training and 500 validation images for clear weather, DarkZurich [65] with 2,416 training and 151 test images for nighttime, and ACDC [66] containing 1,600 training, 406 validation, and 2,000 test images for adverse weather (fog, night, rain, and snow). The training resolution follows the used UDA methods (e.g. half resolution for DAFormer [30] or full resolution for HRDA [31]).

We evaluate MIC based on a DAFormer network [30] with a MiT-B5 encoder [83], and a DeepLabV2 [5] with a ResNet-101 [26] backbone. All backbones are initialized with ImageNet pretraining. In the default UDA setting, we follow the HRDA [31] multi-resolution self-training strategy and training parameters, i.e. AdamW [51] with a learning

rate of $6\times10^{-5}$ for the encoder and $6\times10^{-4}$ for the decoder, 40k training iterations, a batch size of 2, linear learning rate warmup, a loss weight $\lambda_{st}^{T}$=1, an EMA factor $\alpha$=0.999, DACS [72] data augmentation, Rare Class Sampling [30], and ImageNet Feature Distance [30]. For adversarial training and entropy minimization, SGD with a learning rate of 0.0025 and $\lambda_{adv}^{T}=\lambda_{ent}^{T}$=0.001 is used.

**Image Classification:** We evaluate MIC on the VisDA-2017 dataset [56], which consists of 280,000 synthetic and real images of 12 classes, as well as the Office-Home dataset [75], which contains 15,500 images from 65 classes for the domains art (A), clipart (C), product (P) and real-world (R). We conduct the experiments with ResNet-101 [26] and ViT-B/16 [16]. For UDA training, we follow SDAT [57], which utilizes CDAN [48] with MCC [36] and a smoothness enhancing loss [57]. We use the same training parameters, i.e. SGD with a learning rate of 0.002, a batch size of 32, and a smoothness parameter of 0.02.

**Object Detection:** For object detection UDA, we evaluate MIC on CS [10] to Foggy CS [64]. The experiments are performed based on Faster R-CNN [58] with ResNet-50 [26] and FPN [46]. For UDA, we adopt SADA [8], which utilizes adversarial training on image and instance level. The same parameters as in [8] are used, i.e. 0.0025 initial learning rate, 60k training iterations, $\lambda_{adv}^{T}$=0.1, and a batch size of 2. Following previous works [8, 61], we report the results in mean Average Precision (mAP) with a 0.5 IoU threshold.

**MIC Parameters:** MIC uses a patch size $b$=64, a mask ratio $r$=0.7, a loss weight $\lambda^{M}$=1, an EMA weight $\alpha$=0.999 following [30, 31], and color augmentation (brightness, contrast, saturation, hue, and blur) following the parameters of [30, 31, 72]. We set the pseudo-label box threshold $\delta$=0.8 following [13, 44] and the quality threshold $\tau$=0.968 following [30, 31, 72]. If a UDA method trains with half resolution [8, 30, 72, 73, 76], the patch size is divided by 2. For image classification and object detection, we use $\alpha$=0.9. For object detection, we reduce the mask ratio $r$=0.5 as the objects of interest are more sparse and a high $r$ increases the risk that they are completely masked out. For target domains with nighttime images (DarkZurich and ACDC), we forgo color augmentation as it can corrupt the content of dark nighttime images due to the locally already low brightness and contrast. The experiments are conducted on an RTX 2080 Ti or a Titan RTX depending on the required memory.

### 4.2. MIC for Semantic Segmentation

First, we combine MIC with different UDA methods and network architectures for semantic segmentation on GTA→CS. Tab. 1 shows that MIC achieves consistent and significant improvements across various UDA methods with different network architectures, ranging from +1.2 up to +4.7 mIoU. Specifically, MIC does not only benefit powerful Transformers such as DAFormer [30] but also CNNs such as DeepLabV2 [5]. Across UDA methods, the performance improvement decreases with a higher UDA performance as expected due to performance saturation.

Table 1. Segmentation performance (mIoU in %) of MIC with different UDA methods on GTA→CS.

| Network | UDA Method | w/o MIC | w/ MIC | Diff. |
|---|---|---|---|---|
| DeepLabV2 [5] | Adversarial [73] | 44.2 | 48.2 | +4.0 |
| DeepLabV2 [5] | Entropy Min. [76] | 44.3 | 49.0 | +4.7 |
| DeepLabV2 [5] | DACS [72] | 53.9 | 56.0 | +2.1 |
| DeepLabV2 [5] | DAFormer [30] | 56.0 | 59.4 | +3.4 |
| DeepLabV2 [5] | HRDA [31] | 63.0 | 64.2 | +1.2 |
| DAFormer [30] | DAFormer [30] | 68.3 | 70.6 | +2.3 |
| DAFormer [30] | HRDA [31] | 73.8 | 75.9 | +2.1 |

Second, we evaluate the performance of MIC combined with the best-performing UDA method HRDA [31] for further domain adaptation scenarios: synthetic-to-real (GTA→CS and Synthia→CS), day-to-nighttime (CS→DarkZurich), and clear-to-adverse-weather (CS→ACDC). Tab. 2 shows clear performance improvements on each benchmark. Specifically, MIC improves the state-of-the-art performance by +2.1 mIoU on GTA→CS, by +1.5 mIoU on Synthia→CS, by +4.3 mIoU on CS→DarkZurich, and by +2.4 mIoU on CS→ACDC. Considering the class-wise IoU in Tab. 2, MIC achieves consistent improvements for most classes when compared to the previous state-of-the-art method HRDA. Classes that most profit from MIC are *sidewalk*, *fence*, *pole*, *traffic sign*, *terrain*, and *rider*. These classes have a comparably low UDA performance, meaning that they are difficult to adapt. Here, context clues appear to play an important role in successful adaptation. For some classes such as *building* or *vegetation* on synthetic-to-real adaptation, MIC increases the performance by a smaller margin, probably because the target context clues play a smaller role for them. In a few particular cases, the performance of single classes decreases for MIC such as *truck* on CS→DarkZurich. These are rare classes, which are underrepresented in the data, which might cause MIC to pick up misleading context biases. The observations from Tab. 2 are also reflected in the example predictions in Fig. 4. While previous methods often recognize only parts of ambiguous regions, MIC fixes these issues by using correctly detected parts as context. For instance, the grille of the bus in Fig. 4 resembles a traffic cabinet (*building*). However, a cabinet between two vehicles is unlikely. Probably using this context prior, MIC can resolve the ambiguity.

### 4.3. MIC for Image Classification

For image classification UDA, we combine MIC with the state-of-the-art method SDAT [57]. On VisDA-2017 (Tab. 3), MIC significantly improves the UDA performance by +2.5 and +3.0 percent points when used with a ResNet and ViT network, respectively The improvement is consistent over almost all classes, where difficult classes generally benefit the most. On Office-Home (Tab. 4), MIC clearly improves

Table 2. Semantic segmentation performance (IoU in %) on four different UDA benchmarks.

| Method | Road | S.walk | Build. | Wall | Fence | Pole | Tr.Light | Sign | Veget. | Terrain | Sky | Person | Rider | Car | Truck | Bus | Train | M.bike | Bike | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Synthetic-to-Real: GTA→Cityscapes (Val.)** | | | | | | | | | | | | | | | | | | | | |
| ADVENT [76] | 89.4 | 33.1 | 81.0 | 26.6 | 26.8 | 27.2 | 33.5 | 24.7 | 83.9 | 36.7 | 78.8 | 58.7 | 30.5 | 84.8 | 38.5 | 44.5 | 1.7 | 31.6 | 32.4 | 45.5 |
| DACS [72] | 89.9 | 39.7 | 87.9 | 30.7 | 39.5 | 38.5 | 46.4 | 52.8 | 88.0 | 44.0 | 88.8 | 67.2 | 35.8 | 84.5 | 45.7 | 50.2 | 0.0 | 27.3 | 34.0 | 52.1 |
| ProDA [89] | 87.8 | 56.0 | 79.7 | 46.3 | 44.8 | 45.6 | 53.5 | 53.5 | 88.6 | 45.2 | 82.1 | 70.7 | 39.2 | 88.8 | 45.5 | 59.4 | 1.0 | 48.9 | 56.4 | 57.5 |
| DAFormer [30] | 95.7 | 70.2 | 89.4 | 53.5 | 48.1 | 49.6 | 55.8 | 59.4 | 89.9 | 47.9 | 92.5 | 72.2 | 44.7 | 92.3 | 74.5 | 78.2 | 65.1 | 55.9 | 61.8 | 68.3 |
| HRDA [31] | 96.4 | 74.4 | 91.0 | **61.6** | 51.5 | 57.1 | 63.9 | 69.3 | 91.3 | 48.4 | 94.2 | 79.0 | 52.9 | 93.9 | 84.1 | 85.7 | 75.9 | 63.9 | 67.5 | 73.8 |
| MIC (HRDA) | **97.4** | **80.1** | **91.7** | 61.2 | **56.9** | **59.7** | **66.0** | **71.3** | **91.7** | **51.4** | **94.3** | **79.8** | **56.1** | **94.6** | **85.4** | **90.3** | **80.4** | **64.5** | **68.5** | **75.9** |
| **Synthetic-to-Real: Synthia→Cityscapes (Val.)** | | | | | | | | | | | | | | | | | | | | |
| ADVENT [76] | 85.6 | 42.2 | 79.7 | 8.7 | 0.4 | 25.9 | 5.4 | 8.1 | 80.4 | – | 84.1 | 57.9 | 23.8 | 73.3 | – | 36.4 | – | 14.2 | 33.0 | 41.2 |
| DACS [72] | 80.6 | 25.1 | 81.9 | 21.5 | 2.9 | 37.2 | 22.7 | 24.0 | 83.7 | – | 90.8 | 67.6 | 38.3 | 82.9 | – | 38.9 | – | 28.5 | 47.6 | 48.3 |
| ProDA [89] | **87.8** | 45.7 | 84.6 | 37.1 | 0.6 | 44.0 | 54.6 | 37.0 | **88.1** | – | 84.4 | 74.2 | 24.3 | 88.2 | – | 51.1 | – | 40.5 | 45.6 | 55.5 |
| DAFormer [30] | 84.5 | 40.7 | 88.4 | 41.5 | 6.5 | 50.0 | 55.0 | 54.6 | 86.0 | – | 89.8 | 73.2 | 48.2 | 87.2 | – | 53.2 | – | 53.9 | 61.7 | 60.9 |
| HRDA [31] | 85.2 | 47.7 | 88.8 | **49.5** | 4.8 | 57.2 | 65.7 | 60.9 | 85.3 | – | 92.9 | 79.4 | 52.8 | 89.0 | – | **64.7** | – | 63.9 | **64.9** | 65.8 |
| MIC (HRDA) | 86.6 | **50.5** | **89.3** | 47.9 | **7.8** | **59.4** | **66.7** | **63.4** | 87.1 | – | **94.6** | **81.0** | **58.9** | **90.1** | – | 61.9 | – | **67.1** | 64.3 | **67.3** |
| **Day-to-Nighttime: Cityscapes→DarkZurich (Test)** | | | | | | | | | | | | | | | | | | | | |
| ADVENT [76] | 85.8 | 37.9 | 55.5 | 27.7 | 14.5 | 23.1 | 14.0 | 21.1 | 32.1 | 8.7 | 2.0 | 39.9 | 16.6 | 64.0 | 13.8 | 0.0 | 58.8 | 28.5 | 20.7 | 29.7 |
| MGCDA† [65] | 80.3 | 49.3 | 66.2 | 7.8 | 11.0 | 41.4 | 38.9 | 39.0 | 64.1 | 18.0 | 55.8 | 52.1 | 53.5 | 74.7 | 66.0 | 0.0 | 37.5 | 29.1 | 22.7 | 42.5 |
| DANNet† [81] | 90.0 | 54.0 | 74.8 | 41.0 | 21.1 | 25.0 | 26.8 | 30.2 | **72.0** | 26.2 | **84.0** | 47.0 | 33.9 | 68.2 | 19.0 | 0.3 | 66.4 | 38.3 | 23.6 | 44.3 |
| DAFormer [30] | 93.5 | 65.5 | 73.3 | 39.4 | 19.2 | 53.3 | 44.1 | 44.0 | 59.5 | 34.5 | 66.6 | 53.4 | 52.7 | 82.1 | 52.7 | 9.5 | 89.3 | 50.5 | 38.5 | 53.8 |
| HRDA [31] | 90.4 | 56.3 | 72.0 | 39.5 | 19.5 | 57.8 | **52.7** | 43.1 | 59.3 | 29.1 | 70.5 | 60.0 | 58.6 | **84.0** | **75.5** | 11.2 | 90.5 | 51.6 | 40.9 | 55.9 |
| MIC (HRDA) | **94.8** | **75.0** | **84.0** | **55.1** | **28.4** | **62.0** | 35.5 | **52.6** | 59.2 | **46.8** | 70.0 | **65.2** | **61.7** | 82.1 | 64.2 | **18.5** | **91.3** | **52.6** | **44.0** | **60.2** |
| **Clear-to-Adverse-Weather: Cityscapes→ACDC (Test)** | | | | | | | | | | | | | | | | | | | | |
| ADVENT [76] | 72.9 | 14.3 | 40.5 | 16.6 | 21.2 | 9.3 | 17.4 | 21.2 | 63.8 | 23.8 | 18.3 | 32.6 | 19.5 | 69.5 | 36.2 | 34.5 | 46.2 | 26.9 | 36.1 | 32.7 |
| MGCDA† [65] | 73.4 | 28.7 | 69.9 | 19.3 | 26.3 | 36.8 | 53.0 | 53.3 | 75.4 | 32.0 | 84.6 | 51.0 | 26.1 | 77.6 | 43.2 | 45.9 | 53.9 | 32.7 | 41.5 | 48.7 |
| DANNet† [81] | 84.3 | 54.2 | 77.6 | 38.0 | 30.0 | 18.9 | 41.6 | 35.2 | 71.3 | 39.4 | 86.6 | 48.7 | 29.2 | 76.2 | 41.6 | 43.0 | 58.6 | 32.6 | 43.9 | 50.0 |
| DAFormer [30] | 58.4 | 51.3 | 84.0 | 42.7 | 35.1 | 50.7 | 30.0 | 57.0 | 74.8 | 52.8 | 51.3 | 58.3 | 32.6 | 82.7 | 58.3 | 54.9 | 82.4 | 44.1 | 50.7 | 55.4 |
| HRDA [31] | 88.3 | 57.9 | 88.1 | **55.2** | 36.7 | 56.3 | **62.9** | 65.3 | 74.2 | 57.7 | 85.9 | 68.8 | 45.7 | 88.5 | **76.4** | 82.4 | 87.7 | 52.7 | 60.4 | 68.0 |
| MIC (HRDA) | **90.8** | **67.1** | **89.2** | 54.5 | **40.5** | **57.2** | 62.0 | **68.4** | **76.3** | **61.8** | **87.0** | **71.3** | **49.4** | **89.7** | 75.7 | **86.8** | **89.1** | **56.9** | **63.0** | **70.4** |

† Method uses additional daytime/clear-weather geographically-aligned reference images.

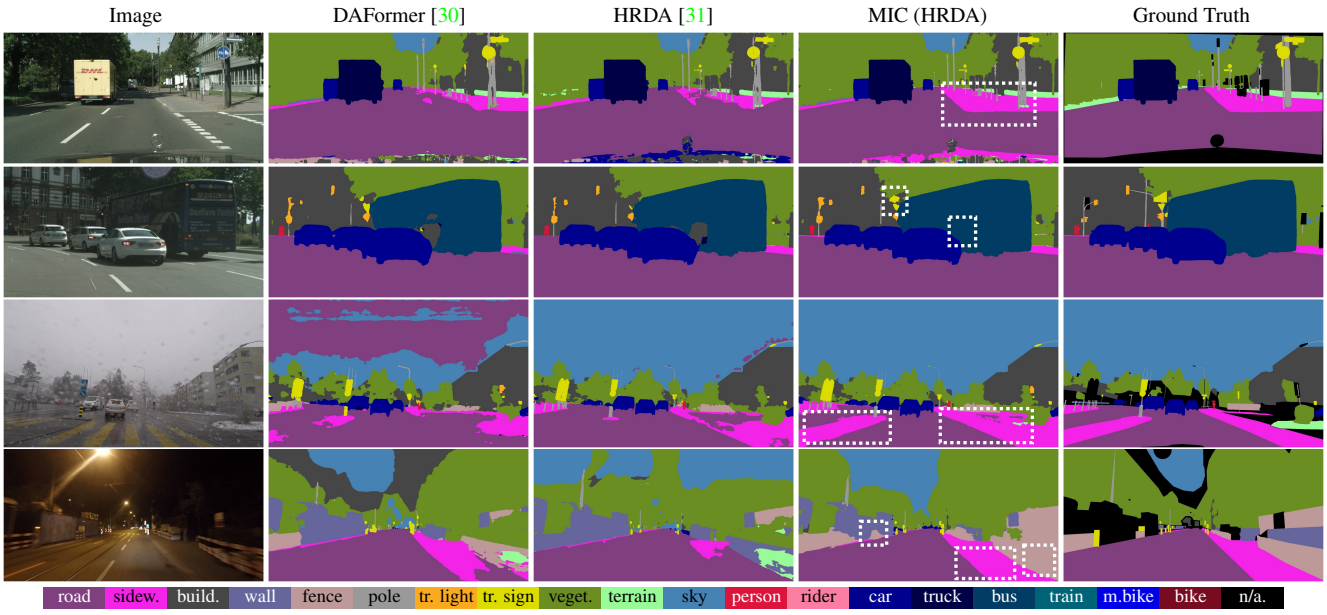



Figure 4. Qualitative comparison of MIC with previous methods on GTA→CS (row 1 and 2), CS→ACDC (row 3), and CS→DarkZurich (row 4). MIC better segments difficult classes such as *sidewalk*, *fence*, *traffic sign*, and *bus*. Further examples are shown in the supplement.

the UDA performance by +1.9. Domains that are difficult to adapt such as A→C or P→C benefit most. Tab. 3 further provides a baseline of SDAT with MAE [25] pretraining, which includes masked image modeling (MIM) and ImageNet supervision. Compared to regular SDAT, additional MIM reduces the performance by -1.4. This demonstrates that naive MIM as additional pretraining is not sufficient to capture the relevant target context dependencies, probably as the learned context is specific to ImageNet and does not transfer well to the target domain.

Table 3. Image classification accuracy in % on VisDA-2017 for UDA. The last column contains the mean across classes.

| Method | | Plane | Bcycl | Bus | Car | Horse | Knife | Mcyle | Persn | Plant | Sktb | Train | Truck | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CDAN [48] | ResNet | 85.2 | 66.9 | 83.0 | 50.8 | 84.2 | 74.9 | 88.1 | 74.5 | 83.4 | 76.0 | 81.9 | 38.0 | 73.9 |
| MCC [36] | | 88.1 | 80.3 | 80.5 | 71.5 | 90.1 | 93.2 | 85.0 | 71.6 | 89.4 | 73.8 | 85.0 | 36.9 | 78.8 |
| SDAT [57] | | 95.8 | 85.5 | 76.9 | 69.0 | 93.5 | **97.4** | 88.5 | 78.2 | 93.1 | 91.6 | 86.3 | 55.3 | 84.3 |
| MIC (SDAT) | | **96.7** | **88.5** | **84.2** | **74.3** | **96.0** | 96.3 | **90.2** | **81.2** | **94.3** | **95.4** | **88.9** | **56.6** | **86.9** |
| TVT [87] | ViT | 92.9 | 85.6 | 77.5 | 60.5 | 93.6 | 98.2 | 89.3 | 76.4 | 93.6 | 92.0 | 91.7 | 55.7 | 83.9 |
| CDTrans [85] | | 97.1 | 90.5 | 82.4 | 77.5 | 96.6 | 96.1 | 93.6 | 88.6 | 97.9 | 86.9 | 90.3 | 62.8 | 88.4 |
| SDAT [57] | | 98.4 | 90.9 | 85.4 | 82.1 | 98.5 | 97.6 | 96.3 | 86.1 | 96.2 | 96.7 | 92.9 | 56.8 | 89.8 |
| SDAT w/ MAE [25] | | 97.1 | 88.4 | 80.9 | 75.3 | 95.4 | 97.9 | 94.3 | 85.5 | 95.8 | 91.0 | 93.0 | 65.4 | 88.4 |
| MIC (SDAT) | | **99.0** | **93.3** | **86.5** | **87.6** | **98.9** | **99.0** | **97.2** | **89.8** | **98.9** | **98.9** | **96.5** | **68.0** | **92.8** |

Table 4. Image classification acc. in % on Office-Home for UDA.

| Method | A→C | A→P | A→R | C→A | C→P | C→R | P→A | P→C | P→R | R→A | R→C | R→P | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CDTrans [85] | 68.8 | 85.0 | 86.9 | 81.5 | 87.1 | 87.3 | 79.6 | 63.3 | 88.2 | 82.0 | 66.0 | 90.6 | 80.5 |
| TVT [87] | 74.9 | 86.8 | 89.5 | 82.8 | 87.9 | 88.3 | 79.8 | 71.9 | 90.1 | 85.5 | 74.6 | 90.6 | 83.6 |
| SDAT [57] | 70.8 | 87.0 | 90.5 | 85.2 | 87.3 | 89.7 | **84.1** | 70.7 | 90.6 | 88.3 | 75.5 | **92.1** | 84.3 |
| MIC (SDAT) | **80.2** | **87.3** | **91.1** | **87.2** | **90.0** | **90.1** | 83.4 | **75.6** | **91.2** | **88.6** | **78.7** | 91.4 | **86.2** |

Table 5. Object detection AP in % on CS→Foggy CS.

| Method | Bus | Bcycl | Car | Mcycle | Persn | Rider | Train | Truck | mAP |
|---|---|---|---|---|---|---|---|---|---|
| DAFaster [7] | 29.2 | 40.4 | 43.4 | 19.7 | 38.3 | 28.5 | 23.7 | 32.7 | 32.0 |
| SW-DA [61] | 31.8 | 44.3 | 48.9 | 21.0 | 43.8 | 28.0 | 28.9 | **35.8** | 35.3 |
| SC-DA [96] | 33.8 | 42.1 | 52.1 | 26.8 | 42.5 | 26.5 | 29.2 | 34.5 | 35.9 |
| MTOR [4] | 38.6 | 35.6 | 44.0 | 28.3 | 30.6 | 41.4 | 40.6 | 21.9 | 35.1 |
| SIGMA [42] | 50.4 | 40.6 | 60.3 | 31.7 | 44.0 | 43.9 | **51.5** | 31.6 | 44.2 |
| SADA [8] | 50.3 | 45.4 | 62.1 | 32.4 | 48.5 | 52.6 | 31.5 | 29.5 | 44.0 |
| MIC (SADA) | **52.4** | **47.5** | **67.0** | **40.6** | **50.9** | **55.3** | 33.7 | 33.9 | **47.6** |

### 4.4. MIC for Object Detection

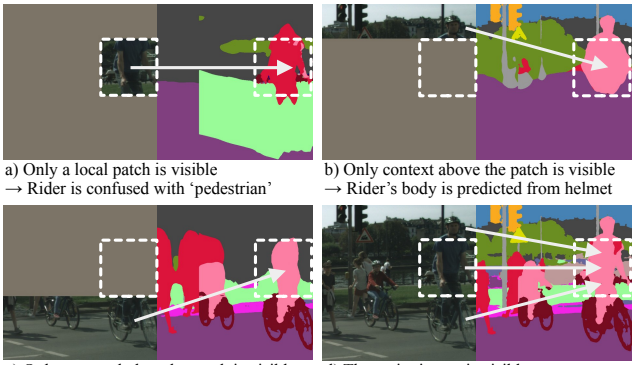
For object detection UDA, we combine MIC with the state-of-the-art Scale-aware Domain Adaptive Faster-RCNN (SADA) [8]. On CS→Foggy CS (Tab. 5), MIC obtains consistent improvements over all categories and achieves +3.6 mAP gain compared to the baseline SADA. The classes *car*, *motorcycle*, and *rider* benefit the most. MIC also shows a clear advantage for most categories compared to more recent methods such as SIGMA [42].

### 4.5. In-Depth Analysis of MIC

**Context Utilization:** To verify that MIC has learned context priors on the target domain, we mask out an image patch, let the trained model predict the semantics of the patch from the visible context, and calculate the mIoU for the patch. As the patch is masked out, the model can only utilize context information to infer its semantics. We repeat this process for all non-overlapping patches of the size 256×256 in the CS val. set. MIC(HRDA) achieves a strong context performance of 52.5 mIoU on GTA→CS while HRDA only reaches 22.8 mIoU, showing that MIC indeed enhances context learning.

To further illustrate the learned comprehensive context relations, we visualize predictions of masked images in Fig. 5. It shows the learned context priors of helmet and bicycle, which MIC internally exploits to predict the rider's body.

**Where to apply MIC?** Tab. 6 shows the performance of MIC with HRDA using images from different domains as masked input: (1) source, (2) target, and (3) both source and target. We observe that: for (1) the performance is -2.7 mIoU worse than HRDA for GTA→CS but it increases by +1.2 mIoU for CS→ACDC, for (2) the performance increases by +2.1 for GTA→CS and +1.6 mIoU for CS→ACDC, and for (3) the mIoU increases by +0.7 for GTA→CS and +2.7 for CS→ACDC. Both benchmarks differ in the domain gap of context relations. While the distributions of context relations can vary between synthetic (GTA) and real data (CS), the context relations of CS and ACDC are very similar as both datasets were recorded in the real world and partly even in the same city. If the context domain gap is large, context relations learned on source images do not transfer well to the target domain and can even hamper the adaptation. However, if the context gap is small, source context relations transfer well to the target domain and can boost the adaptation performance. Therefore, we also apply MIC to the source domain, in addition to the default target domain, for clear-to-adverse-weather and day-to-nighttime adaptation.



a) Only a local patch is visible → Rider is confused with 'pedestrian'
b) Only context above the patch is visible → Rider's body is predicted from helmet
c) Only context below the patch is visible → Rider's body is predicted from bicycle
d) The entire image is visible → All local and context clues can be used

Figure 5. Predictions of MIC for masked variants of the same image demonstrating the learned context priors of MIC.

Table 6. MIC with HRDA [31] for images from different domain.

| MIC Domain | $\text{mIoU}_{GTA \to CS}$ | $\text{mIoU}_{CS \to ACDC(Val)}$ |
|---|---|---|
| – | 73.8 | 65.3 |
| Source | 71.1 | 66.5 |
| Target | 75.9 | 66.9 |
| Source+Target | 74.5 | 68.0 |

Table 7. MIC ablation study with DAFormer [30] on GTA→CS.

| | Masked Img. | Color Aug. | EMA Teacher | Pseudo Lbl. Weight | mIoU |
|---|---|---|---|---|---|
| 1 | – | – | – | – | 68.3 |
| 2 | ✓ | ✓ | ✓ | ✓ | 70.6 |
| 3 | – | ✓ | ✓ | ✓ | 50.6 |
| 4 | ✓ | – | ✓ | ✓ | 70.3 |
| 5 | ✓ | ✓ | – | ✓ | 69.9 |
| 6 | ✓ | ✓ | ✓ | – | 69.0 |

Table 8. Parameter study of the patch size $b$ and the mask ratio $r$ of MIC with DAFormer [30] on GTA→CS. The color indicates the difference to the DAFormer performance of 68.3 mIoU.

| Patch Size b \ Mask Ratio r | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|
| 32 | 69.3 | 69.9 | 69.7 | 69.3 |
| 64 | 69.2 | 70.3 | 70.6 | 69.7 |
| 128 | 68.7 | 70.5 | 70.4 | 68.2 |
| 256 | 66.2 | 69.5 | 69.8 | 68.0 |

Table 9. Comparison of UDA on GTA→CS and supervised training on CS. "Rel." indicates $\text{mIoU}_{UDA}/\text{mIoU}_{Superv.}$.

| | $\text{mIoU}_{UDA}$ | $\text{mIoU}_{Superv.}$ | Rel. |
|---|---|---|---|
| DAFormer [30] | 68.3 | 77.6 | 88.0% |
| MIC (DAFormer) | 70.6 | 77.9 | 90.6% |
| Improvement | +2.3 | +0.3 | +2.6% |

**Component Ablation:** To gain further insights, we ablate the components of MIC and evaluate the performance with DAFormer [30] (due to the faster training) on GTA→CS in Tab. 7. The complete MIC achieves 70.6 mIoU (row 2), which is +2.3 mIoU better than DAFormer (row 1). First, the masking of the image is ablated, meaning that the consistency training is done with unmasked but still augmented target images (see "MIC Parameters" in Sec. 4.1). Without masking out image patches, the performance heavily decreases by -20.0 mIoU (cf. rows 2 and 3). On the other side, color augmentation is not essential for MIC as its ablation only reduces the performance by -0.3 mIoU (cf. rows 2 and 4). This demonstrates the importance of context learning with masked images. Replacing the EMA predictions with the regular model predictions decreases the performance of MIC by -0.7 mIoU (cf. rows 2 and 5). Without the pseudo-label confidence loss weight, the mIoU drops by -1.6 (cf. rows 2 and 6) showing that it is important to reduce the weight of uncertain samples for MIC training.

**Patch Size and Mask Ratio:** Tab. 8 shows the influence of the mask patch size $b$ and mask ratio $r$. Compared to DAFormer, MIC achieves significant improvements in a range of $b$ between 64 and 128 and $r$ between 0.5 and 0.7. The best performance is achieved for $b$=64 and $r$=0.7. Only for a very large $b$ of 256, which is a quarter of the image height, MIC decreases the performance. Note that $b$ is internally divided by 2 as DAFormer uses half resolution.

**MIC for Supervised Training:** We compare the UDA and the supervised performance of DAFormer with and without MIC in Tab. 9. Also for supervised training, MIC achieves a slight improvement of +0.3 mIoU. However, the improvement for UDA is much more significant with +2.3 mIoU, showing that MIC is particularly useful for UDA. Therefore, MIC is able to increase the relative UDA performance (column "Rel") by +2.6 percent points, so that UDA with MIC achieves remarkable 90.6% of the performance of a network trained with full supervision on the target domain.

Table 10. Runtime and memory consumption during training and inference on an RTX 2080 Ti (row 1-4) or Titan RTX (row 5-6).

| | Training | | Inference | |
|---|---|---|---|---|
| | Throughput | GPU Memory | Throughput | GPU Memory |
| Adversarial [73] | 1.40 it/s | 5.38 GB | 11.2 img/s | 0.5 GB |
| MIC (Adversarial) | 0.81 it/s | 5.55 GB | 11.2 img/s | 0.5 GB |
| DAFormer [30] | 0.71 it/s | 9.64 GB | 8.6 img/s | 1.0 GB |
| MIC (DAFormer) | 0.57 it/s | 9.74 GB | 8.6 img/s | 1.0 GB |
| HRDA [31] | 0.36 it/s | 22.46 GB | 0.8 img/s | 9.4 GB |
| MIC (HRDA) | 0.29 it/s | 22.55 GB | 0.8 img/s | 9.4 GB |

**Runtime/Memory:** Tab. 10 shows the runtime and GPU memory footprint of representative UDA methods with and without MIC. For methods without an EMA teacher such as adversarial training, MIC reduces training speed by 75% due to the additional calculations for MIC and increases the GPU memory consumption by 3% due to the EMA teacher. The memory increase is small as the loss terms $\mathcal{L}^S$, $\mathcal{L}^T$, and $\mathcal{L}^M$ are backpropagated separately. For UDA methods that already use an EMA teacher such as DAFormer or HRDA, the teacher and its predictions can be re-used, so that the training speed only increases by 24% and the memory footprint by 1%. Importantly, MIC is only used during training and does not increase the inference time at all.

**Supplement:** The supplement provides results on CS→FoggyZurich, further results of MIC with DAFormer and $\text{HRDA}_{\text{DeepLabV2}}$, additional parameter and behavior studies, an extended qualitative analysis, and further discussions.

## 5. Conclusions

In this paper, we presented Masked Image Consistency (MIC), a UDA module to improve the learning of target domain context relations. By enforcing consistency of predictions from partly masked and complete images, the network is trained to utilize robust context clues. MIC can be utilized for UDA across various visual recognition tasks such as image classification, semantic segmentation, and object detection as well as multiple domain adaptation scenarios such as synthetic-to-real, clear-to-adverse-weather, and day-to-nighttime. In a comprehensive evaluation, we have shown that MIC achieves significant performance improvements in all of these UDA tasks. For instance, MIC respectively improves the state-of-the-art performance by +2.1 and +3.0 on GTA→CS and VisDA-2017. We hope that, due to its simplicity, MIC can be used as part of future UDA methods to narrow the gap between UDA and supervised learning.

## References


[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *CVPR*, pages 15384–15394, 2021. 2, 4

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2021. 3

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 3

[4] 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 *CVPR*, pages 11457–11466, 2019. 2, 7

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 40(4):834–848, 2017. 4, 5

[6] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*, pages 1841–1850, 2019. 2

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 2, 7

[8] Yuhua Chen, Haoran Wang, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Scale-aware domain adaptive faster r-cnn. *IJCV*, 129(7):2223–2243, 2021. 2, 3, 5, 7

[9] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *ICCV*, pages 6830–6840, 2019. 2

[10] 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 *CVPR*, pages 3213–3223, 2016. Dataset URL: https://www.cityscapes-dataset.com/, Dataset License: https://www.cityscapes-dataset.com/license/. 1, 4, 5

[11] Dengxin Dai, Christos Sakaridis, Simon Hecker, and Luc Van Gool. Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *IJCV*, 128(5):1182–1204, 2020. 2

[12] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *ITSC*, pages 3819–3824, 2018. 2

[13] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021. 5

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[15] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 3

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 3, 5

[17] Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *BMVC*, 2020. 4

[18] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *ICLR*, 2018. 2

[19] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016. 2, 3

[20] Rui Gong, Wen Li, Yuhua Chen, Dengxin Dai, and Luc Van Gool. Dlow: Domain flow and applications. *IJCV*, 129(10):2865–2888, 2021. 1

[21] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *CVPR*, pages 2477–2486, 2019. 2

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. 2

[23] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *NeurIPS*, 17, 2004. 2

[24] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *NeurIPS*, 19, 2006. 2

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 3, 6, 7

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 4, 5

[27] 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 *ICML*, pages 1989–1998, 2018. 1, 2

[28] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 2

[29] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Köring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. In *CVPR*, pages 11130–11140, 2021. 4

[30] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 8

[31] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *ECCV*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[32] Lukas Hoyer, Dengxin Dai, Qin Wang, Yuhua Chen, and Luc Van Gool. Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. *arXiv preprint arXiv:2108.12545*, 2021. 2

[33] Lukas Hoyer, Mauricio Munoz, Prateek Katiyar, Anna Khoreva, and Volker Fischer. Grid saliency for context explanations of semantic segmentation. In *NeurIPS*, pages 6462–6473, 2019. 3

[34] Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu, and Ling Shao. Category contrast for unsupervised domain adaptation in visual tasks. In *CVPR*, pages 1203–1214, 2022. 2

[35] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In *ECCV*, pages 705–722, 2020. 3

[36] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *ECCV*, pages 464–480, 2020. 5, 7

[37] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. *arXiv preprint arXiv:2203.12719*, 2022. 3

[38] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-Supervised Semantic Segmentation With Directional Context-Aware Consistency. In *CVPR*, pages 1205–1214, 2021. 2

[39] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, 2013. 2

[40] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *ECCV*, pages 481–497, 2020. 3

[41] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan. Scan: Cross domain object detection with semantic conditioned adaptation. In *AAAI*, volume 6, page 7, 2022. 2

[42] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, pages 5291–5300, 2022. 2, 7

[43] Xiaotong Li, Yixiao Ge, Kun Yi, Zixuan Hu, Ying Shan, and Ling-Yu Duan. mc-beit: Multi-choice discretization for image bert pre-training. *arXiv preprint arXiv:2203.15371*, 2022. 3

[44] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *CVPR*, pages 7581–7590, 2022. 2, 5

[45] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer for visual representation. *NeurIPS*, 34:13165–13176, 2021. 3

[46] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 5

[47] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015. 2

[48] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *NeurIPS*, 31, 2018. 2, 5, 7

[49] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *NeurIPS*, 2016. 2

[50] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, pages 2208–2217, 2017. 2

[51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 4

[52] Yawei Luo, Ping Liu, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Category-level adversarial adaptation for semantic segmentation using purified features. *PAMI*, 2021. 2

[53] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*, pages 415–430, 2020. 2, 3

[54] Luke Melas-Kyriazi and Arjun K. Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *CVPR*, pages 12435–12445, 2021. 2

[55] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *CVPR*, pages 2239–2247, 2019. 2

[56] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. Dataset URL: https://github.com/VisionLearningGroup/taskcv-2017-public/tree/master/classification. 5

[57] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, Arihant Jain, and Venkatesh Babu Radhakrishnan. A closer look at smoothness in domain adversarial training. In *ICML*, pages 18378–18399, 2022. 1, 3, 5, 7

[58] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 3, 4, 5

[59] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, pages 102–118, 2016. Dataset URL: https://download.visinf.tu-darmstadt.de/data/from_games/. 4

[60] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, pages 3234–3243, 2016. Dataset URL: http://synthia-dataset.net/, Dataset License: CC BY-NC-SA 3.0. 4

[61] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 5, 7

[62] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsuper-

vised domain adaptation. In *CVPR*, pages 3723–3732, 2018. 2

[63] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NeurIPS*, 2016. 2

[64] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, 2018. Dataset URL: https://www.cityscapes-dataset.com/. 5

[65] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *PAMI*, 2020. Dataset URL: https://www.trace.ethz.ch/publications/2019/GCMA_UIoU/. 4, 6

[66] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, pages 10765–10775, 2021. Dataset URL: https://acdc.vision.ee.ethz.ch/, Dataset License: https://acdc.vision.ee.ethz.ch/license. 1, 4

[67] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 2

[68] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016. 2

[69] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, pages 443–450, 2016. 2

[70] Tao Sun, Cheng Lu, Tianshuo Zhang, and Haibin Ling. Safe self-refinement for transformer-based domain adaptation. In *CVPR*, pages 7191–7200, 2022. 2

[71] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pages 1195–1204, 2017. 2, 4

[72] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. DACS: Domain Adaptation via Cross-domain Mixed Sampling. In *WACV*, pages 1379–1389, 2021. 1, 2, 3, 4, 5, 6

[73] 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 *CVPR*, pages 7472–7481, 2018. 1, 2, 3, 5, 8

[74] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *ICCV*, pages 1456–1465, 2019. 3

[75] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017. Dataset URL: https://www.hemanthdv.org/officeHomeDataset.html. 5

[76] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, pages 2517–2526, 2019. 2, 5, 6

[77] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *ICCV*, pages 7364–7373, 2019. 2

[78] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *ECCV*, pages 642–659, 2020. 3

[79] Qin Wang, Dengxin Dai, Lukas Hoyer, Olga Fink, and Luc Van Gool. Domain adaptive semantic segmentation with self-supervised depth estimation. In *ICCV*, pages 8515–8525, 2021. 1, 2

[80] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, pages 14668–14678, 2022. 3

[81] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *CVPR*, pages 15769–15778, 2021. 6

[82] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *arXiv preprint arXiv:2204.08808*, 2022. 2

[83] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *NeurIPS*, 2021. 4

[84] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663, 2022. 3

[85] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021. 2, 7

[86] Jinyu Yang, Weizhi An, Chaochao Yan, Peilin Zhao, and Junzhou Huang. Context-aware domain adaptation in semantic segmentation. In *WACV*, pages 514–524, 2021. 3

[87] Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2108.05988*, 2021. 7

[88] Kai Zhang, Yifan Sun, Rui Wang, Haichang Li, and Xiaohui Hu. Multiple fusion adaptation: A strong framework for unsupervised semantic segmentation adaptation. In *BMVC*, 2021. 2

[89] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*, pages 12414–12424, 2021. 2, 3, 6

[90] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. In *NeurIPS*, pages 435–445, 2019. 2, 3

[91] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *CVPR*, pages 3801–3809, 2018. 2, 4

[92] Yang Zhang, Philip David, Hassan Foroosh, and Boqing Gong. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *PAMI*, 42(8):1823–1841, 2019. 2

[93] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV*, 129(4):1106–1120, 2021. 2

[94] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Uncertainty-aware consistency regularization for cross-domain semantic segmentation. *arXiv preprint arXiv:2004.08878*, 2020. 2

[95] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Jiangmiao Pang, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Context-aware mixup for domain adaptive semantic segmentation. In *WACV*, pages 514–524, 2021. 2

[96] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *CVPR*, pages 687–696, 2019. 7

[97] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, pages 289–305, 2018. 1, 2, 3