MoDA: Leveraging Motion Priors from Videos for Advancing Unsupervised Domain Adaptation in Semantic Segmentation

Fei Pan[1] Xu Yin[2] Seokju Lee[3] Axi Niu[2] Sungeui Yoon[2] In So Kweon[2]
[1]University of Michigan, Ann Arbor [2]KAIST [3]KENTECH
feipan@umich.edu, slee@kentech.ac.kr, {yinofsgvr,sungeui,iskweon77}@kaist.ac.kr

Abstract

*Unsupervised domain adaptation (UDA) has been a potent technique to handle the lack of annotations in the target domain, particularly in semantic segmentation task. This study introduces a different UDA scenarios where the target domain contains unlabeled video frames. Drawing upon recent advancements of self-supervised learning of the object motion from unlabeled videos with geometric constraint, we design a **Mo**tion-guided **D**omain **A**daptive semantic segmentation framework (MoDA). MoDA harnesses the self-supervised object motion cues to facilitate cross-domain alignment for segmentation task. First, we present an object discovery module to localize and segment target moving objects using object motion information. Then, we propose a semantic mining module that takes the object masks to refine the pseudo labels in the target domain. Subsequently, these high-quality pseudo labels are used in the self-training loop to bridge the cross-domain gap. On domain adaptive video and image segmentation experiments, MoDA shows the effectiveness utilizing object motion as guidance for domain alignment compared with optical flow information. Moreover, MoDA exhibits versatility as it can complement existing state-of-the-art UDA approaches.*

1. Introduction

Fully-supervised semantic segmentation [4, 20] is a data-hungry task that requires all pixels of training images to be assigned with a semantic label. However, providing pixel-wise human annotations for semantic segmentation is expensive and time-consuming [5]. Recently, unsupervised domain adaptation (UDA) has become an effective technique to alleviate the necessity of pixel-wise data labeling. Existing UDA for semantic segmentation methods utilizes adversarial learning for feature or output-level domain alignment [25, 31, 32] or self-training techniques that refine target pseudo labels in an iterative process [1, 30, 47, 48]. While these methods show notable improvements, particularly for background categories (*e.g.*, tree, road, sky), they are limited to the real-world dynamic scenes containing multiple moving objects, as an example shown in Fig. 1.

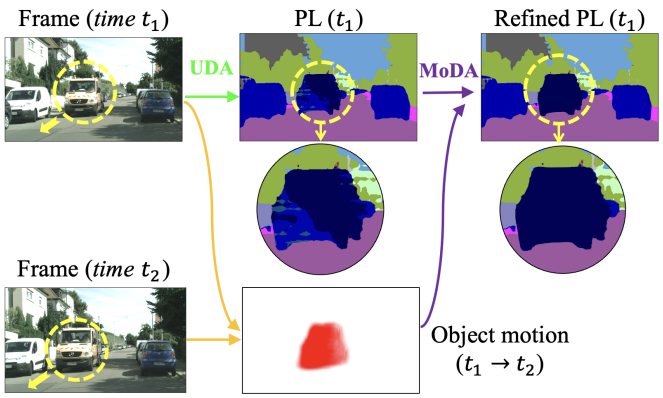


Figure 1. Current UDA methods show notable performance for background categories (*e.g.*, tree, road, sky), but they are limited to the real-world dynamic scenes containing multiple moving objects (*e.g.*, buses). We propose MoDA that uses object motion from unlabeled videos as complementary guidance to refine pseudo labels (PL) in the target domain. Note that the object motion is learned using using *self-supervised geometric constraints* from sequential video frames (t_1 and t_2), *without requiring any annotations*.

The recent trend in dynamic scene understanding involves learning the object motion and the camera's ego-motion from unlabeled sequential image pairs. A bird's eye view in Fig. 2(a) shows an example of the *object motion* and the *camera's ego-motion*. Existing works [8, 17, 18] suggest learning a motion network to separate the local object motion from the global camera ego-motion in the static scenes with *self-supervised geometrical constraints*. This separation helps to isolate the object motion from the camera's ego-motion. As an example in Fig. 1, the object motion is capable of *segmenting* the moving yellow bus out from static scenes.

We propose that the object motion learned from geometric constraints can be used as *complementary* guidance for

domain adaptation to the target domain. Specifically, the segmentation network suffers from the cross-domain gap due to the lack of semantic labels in the target domain. However, the motion network is trained separately from the segmentation network. Moreover, the motion network is trained only using target frames in the *self-supervised* manner with geometric constraints, which *does not require* any semantic labels. So the object motion from the motion network is not affected by the cross-domain gaps caused by the labeling issue.

In this work, we present a motion-guided unsupervised domain adaption (MoDA) method for segmentation task, which utilizes the *self-supervised object motion from geometrical constraints as cues for domain alignment*. First, we present an object discovery module that takes the object motion information learned from the unlabeled videos as input and localizes and segments the moving object in the target domain. Then, we propose a semantic mining module to refine target pseudo labels by utilizing the object masks from the target domain. Subsequently, these high-quality target pseudo labels are sent as input in the self-training loop to optimize the target segmentation model. On domain adaptive video and image segmentation benchmarks, MoDA shows the effectiveness utilizing object motion as guidance for domain alignment compared with optical flow information. Moreover, it is versatile to complement existing state-of-the-art UDA approaches.

Contribution:

- We are the first to utilize self-supervised object motion information from unlabeled videos to facilitate cross-domain alignment for semantic segmentation task, without requiring any target annotations.
- MoDA contains the object discovery module and the semantic mining module to refine target pseudo labels. These two modules are directly used without the need for training.
- MoDA shows superior performance on the benchmarks for the domain adaptive video and image segmentation compared with optical flow baselines. MoDA is also versatile to complement existing state-of-the-art UDA approaches.

2. Related Works

Domain adaptive image segmentation. The goal of domain adaptation is to align the domain shift between the labeled source and target domains [7, 15]. For domain adaptive image segmentation, adversarial learning [19, 25, 31–33] and self-training [1, 16, 21, 26, 29, 30, 37, 41, 45, 47, 48] approaches are widely adopted, and demonstrate compelling adaption results. [25] designs two discriminator networks to implement the inter-domain and intra-domain alignment. [44] proposes to average the predictions from the source and the target domain to stabilize the self-training process and further incorporate the uncertainty [45] to minimize the prediction variance. Existing works consider the domain alignment on the category level [34] and instance level [40] to learn domain-invariant features. [43] proposes to handle more diverse data in the target domain. adopt image-level annotations from the target domain to bridge the domain gap. [10] propose combined learning of depth and segmentation for domain adaptation with self-supervision from geometric constraints. Different from existing approaches, we consider using motion priors as guidance for domain alignment in this work.

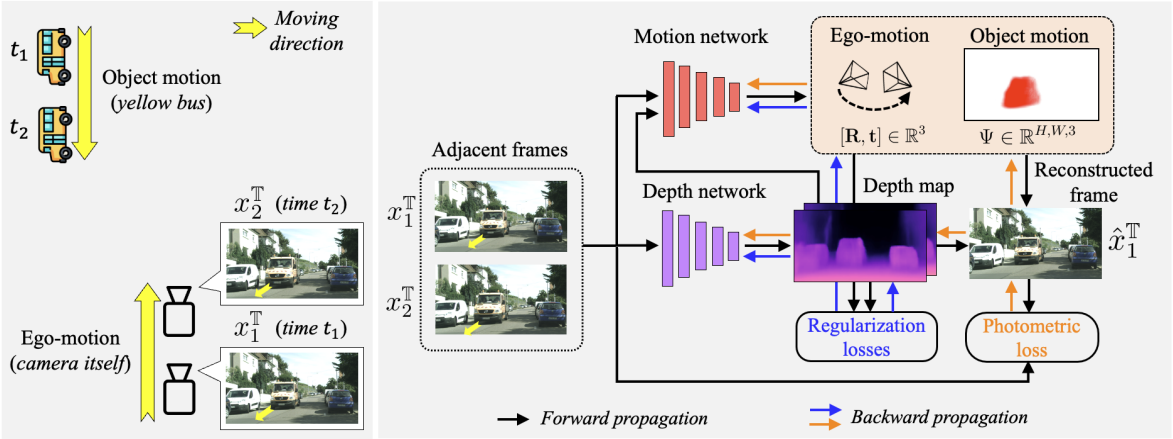
Self-training for domain alignment. In self-training, the network is trained with pseudo-labels from the target domain, which can be pre-computed offline or calculated online during training [1, 12, 13, 37, 41]. [41] proposes to estimate category-level prototypes on the fly and refine the pseudo labels iteratively, to enhance the adaptation effect. [1] utilizes data augmentation and momentum updates to regulate cross-domain consistency. [46] proposes to align the cross-domain gaps on structural affinity space for the segmentation task. In this work, we introduce the motion masks that provide complementary object geometric information as prior. Specifically, we develop the motion-guided self-training and the moving object label mining module to refine the target pseudo labels and thus improve the adaptation performance.

Domain adaptive video segmentation. Exploiting motion information like optical flow [24] to separate the objects in videos to regulate the segmentation training is well explored in the video segmentation field. In UDA, there are several attempts that introduce temporal supervision signals to enforce the domain alignment. [9] regulates the cross-domain temporal consistency with adversarial training to minimize the distribution discrepancy. [39] proposes to capture the spatiotemporal consistency in the source domain by data augmentation across frames. In this work, we propose to utilize the 3D object motion [18] of the target sequential image pairs, which provides rich information for localizing and segmenting the moving objects.

3. Preliminary

We train the motion network G_m to learn object motion prediction in the target domain illustrated in Sec. 3.1. Then, we conduct motion mask preprocessing to obtain the instance-level motion masks in the target domain shown in Sec. 3.2.

Notations. We have a set of source images $X^{\mathbb{S}} = \{x_n^{\mathbb{S}}\}_{n=1}^{N^{\mathbb{S}}}$ with the corresponding segmentation annotations $Y^{\mathbb{S}} = \{y^{\mathbb{S}}\}_{n=1}^{N^{\mathbb{S}}}$, where $N^{\mathbb{S}}$ is the number of the source images. We also have a set of unlabeled sequential frames $X^{\mathbb{T}} = \{(x_{n,1}^{\mathbb{T}}, x_{n,2}^{\mathbb{T}}, \ldots, x_{n,k}^{\mathbb{T}})\}_{n=1}^{N^{\mathbb{T}}}$ in the target domain, where $x_{n,2}^{\mathbb{T}}$ is an adjacent frame of $x_{n,1}^{\mathbb{T}}$, and $N^{\mathbb{T}}$ is the number of the target video sequences. Note that $x^{\mathbb{T}} \in \mathbb{R}^{H,W,3}$,



(a) A bird's eye view of the scene (b) The geometric learning and motion mask preprocessing in target domain

Figure 2. The object motion information is learned by self-supervised geometric constraints from unlabeled target video frames, without any annotations. (a) The visualization of a bird's eye view of the dynamic scene, where the yellow bus is moving toward the camera. We indicate the *object motion* of the yellow bus and the *ego motion* from the camera itself. (b) The diagram for geometric training to learn the object motion from a pair of target adjacent video frames. The motion network and depth network are trained by the *self-supervised* losses (photometric loss and regularization losses) following geometric constraints.

$x^{\mathbb{S}} \in \mathbb{R}^{\hat{H},\hat{W},3}$, $y^{\mathbb{S}} \in \mathbb{B}^{\hat{H},\hat{W},C}$ as pixel-wise one-hot vectors, and C is the number of all the categories $\mathbb{C}$. For the geometric part, our motion network and the depth network are represented by G_m and G_p. For the semantic part, our segmentation network is indicated by G_e. Our objective is to adapt the segmentation model G_e to the unlabeled target domain.

3.1. Target Domain Geometric Learning

The joint acquisition of knowledge concerning the moving objects and the motion of the ego camera within a local static scene has obtained considerable attention within the area of dynamic scene understanding [3, 8, 17, 18]. Recent investigations have introduced a method to disentangle the local objects' independent motion (called *object motion*) from the global camera's motion (called *ego-motion*) in a self-supervised manner with geometric constraints [17, 18]. We use the object motion to segment the moving objects out from the static scene.

Given an unlabeled target frame and its adjacent frame $\{x_1^{\mathbb{T}}, x_2^{\mathbb{T}}\}$, our depth network G_p is trained to estimate their depth maps $\{d_1^{\mathbb{T}}, d_2^{\mathbb{T}}\} \in \mathbb{R}^{H,W}$. Then, these frames with their corresponding depth maps are concatenated as input $\{x_1^{\mathbb{T}}, d_1^{\mathbb{T}}, x_2^{\mathbb{T}}, d_2^{\mathbb{T}}\}$ and sent into the motion net G_m. Then G_m is trained generates the camera's ego-motion $[R, t]$ (in 6 DoF) and the object motion $\Psi \in \mathbb{R}^{H,W,3}$ (in 3D space: x, y, and z-axis), where $R \in \mathbb{R}^3$ is the camera's ego rotation and $t \in \mathbb{R}^3$ is the camera's ego translation. On this basis, we use the camera's ego-motion $[R, t]$, the object motion Ψ, and the adjacent frame $x_2^{\mathbb{T}}$ to reconstruct the original image $x_1^{\mathbb{T}}$ with an inverse warping operation

$$\hat{x}_1^{\mathbb{T}} = \mathcal{F}(x_2^{\mathbb{T}}, d_1^{\mathbb{T}}, R, t, \Psi, K), \tag{1}$$

where $\hat{x}_1^{\mathbb{T}}$ is the reconstructed image by warping the adjacent image $x_2^{\mathbb{T}}$ (as reference), $\mathcal{F}$ is the projection operation using camera geometry [17], and $K \in \mathbb{R}^{3\times3}$ is the camera's intrinsic parameters. We adopt the photometric loss [8] and the regularization losses [17, 18] to optimize the motion network G_m and the depth network G_p together shown in Fig. 2 (b).

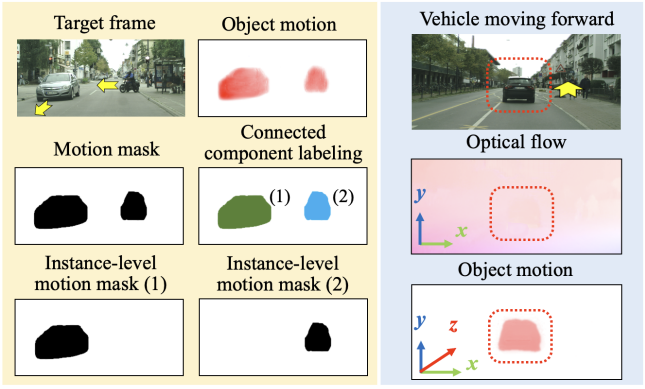
3.2. Motion Mask Preprocessing

Our motion mask preprocessing (MMP) aims at localizing the moving instances based on the object motion information in the target domain. An exemplary procedure of MMP is shown in Fig. 3 (a). Given the motion network G_m optimized by the photometric loss and the regularization losses, we first predict the object motion map $\Psi \in \mathbb{R}^{H,W,3}$ from the adjacent frames $x_1^{\mathbb{T}}, x_2^{\mathbb{T}} \in \mathbb{R}^{H,W,3}$ as input. On this basis, we extract a binary motion mask $\Psi^{\mathbb{M}} \in \mathbb{B}^{H,W}$ from Ψ via

$$\Psi^{\mathbb{M}}_{(i)} = \begin{cases} 1, & \text{if } |\Psi_{(i,d)}| > 0, \forall d \in \{1,2,3\} \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where i indicate the pixel coordinate, $\Psi^{\mathbb{M}}_{(i)}$ is the mask value on the image pixel $x^{\mathbb{T}}_{(i)}$. Note that the object motion Ψ predicted G_m is relative to the scene, *e.g.*, it is separated from the camera's ego-motion. Therefore, we use $\Psi^{\mathbb{M}}$ to localize and segment all the moving objects in the scene.

The binary motion mask $\Psi^{\mathbb{M}}$ could potentially include multiple moving instances, as it is common for multiple cars



(a) Motion mask preprocessing (b) Optical flow *vs.* object motion

Figure 3. (a) The motion mask extracted from the object motion map include multiple moving instances. Therefore, we adopt connect component labeling to identify each moving instance. (b) The object motion map is in 3D space (x, y, z-axis). Object motion is capable to capture the motion pattern at z-axis (moving forward/backward) such as this vehicle. In contrast, optical flow which lies in 2D space fails to capture the motion of this vehicle.

and motorcycles to move together within the same scenes. To differentiate various moving instances, we utilize connected component labeling [36] which is to identify each moving instance from $\Psi^{\mathbb{M}}$ with a unique label. We first run connected component labeling on $\Psi^{\mathbb{M}}$ to get a "new" map $\tilde{\Psi}^{\mathbb{M}}$ with unique labels

$$\tilde{\Psi}^{\mathbb{M}} = \Gamma(\Psi^{\mathbb{M}}), \tag{3}$$

where Γ is the connected component labeling process. For each unique label m in $\tilde{\Psi}^{\mathbb{M}}$, we extract a binary motion mask $\psi^m \in \mathbb{B}^{H,W}$ for the m^{th} moving instance. In this regard, we generate a set of binary instance-level motion masks $\{\psi^m\}_{m=1}^{M}$ via

$$\{\psi^m\}_{m=1}^{M} = \Delta(\tilde{\Psi}^{\mathbb{M}}), \tag{4}$$

where Δ denotes the instance-level motion mask extraction mentioned above (an example is shown in Fig. 3 (a)), and M is the number of the moving instance masks in $\Psi^{\mathbb{M}}$.

Difference with Optical Flow. The object motion differs from the optical flow in two aspects. 1) Optical flow is not accurate for detecting the motion pattern on the front-to-back axis (z-axis) which is a common motion pattern in the real world, as it is a motion representation in *2D space*. However, this issue doesn't exist in object motion which lies in *3D space*. 2) Optical flow is a motion representation of all the pixels with respect to the camera's movement. Therefore, optical flow is a mixed motion representation of the object motion and ego-motion. In contrast, Object motion [17, 18] represents the independent movement of the objects, which is disentangled from the camera's ego-motion through the learning process in Sec. 3.1. Provided a car moving on the z-axis, we visualize the optical flow and the object motion in Fig. 3 (b), where the object (the vehicle's) motion and camera's ego-motion are similar (toward the z-axis with similar velocity). The object motion successfully detects the the vehicle's motion at the z-axis. However, the optical flow cannot easily detect it because the the vehicle's motion is similar to the camera's ego-motion.

4. Methodology

MoDA consists of two modules: object discovery and semantic mining. The object discovery module takes the instance-level motion masks and extracts the moving object masks (in Sec. 4.2). Subsequently, the semantic mining module utilizes the object masks to refine target pseudo labels (in Sec. 4.3).

4.1. Segmentation Warm-up

We first conduct warm-up training for segmentation network following DACS [30]. Given a labeled source frame $x^{\mathbb{S}}$ and its ground-truth map $y^{\mathbb{S}}$, we first train the segmentation network G_e with the supervised segmentation loss

$$\mathcal{L}_{ce}^{\mathbb{S}} = -\sum_{i=1}^{\hat{H},\hat{W}}\sum_{c=1}^{C} y_{(i,c)}^{\mathbb{S}} \log(p_{(i,c)}^{\mathbb{S}}), \tag{5}$$

where the source prediction $p^{\mathbb{S}} = G_e(x^{\mathbb{S}}) \in \mathbb{R}^{\hat{H},\hat{W},C}$, $p_{(i,c)}^{\mathbb{S}}$ denotes the predicted softmax possibility on c^{th} category on the pixel $x_{(i)}^{\mathbb{S}}$, G_e is the segmentation network, and C is the total number of categories. The segmentation network trained solely on the source domain lacks generalization when applied to the target domain. To bridge the domain gap, we optimize the cross-entropy loss with the target pseudo labels. For simplicity, let the target pseudo label at the t^{th} iteration denote as $\tilde{y}^{\mathbb{T}} \in \mathbb{B}^{H,W,C}$. The cross-entropy loss is defined by

$$\mathcal{L}_{ce}^{\mathbb{T}} = -\sum_{i=1}^{H,W}\sum_{c=1}^{C} \tilde{y}_{(i,c)}^{\mathbb{T}} \log(p_{(i,c)}^{\mathbb{T}}). \tag{6}$$

During warm-up stage, we also follow the *mixing augmentation between source and target samples* used in DACS.

4.2. Target Object Discovery

Directly applying the instance-level motion masks $\{\psi^m\}_{m=1}^{M}$ for boosting the quality of target pseudo labels encounters two points of challenge. 1) The instance-level motion masks provide a *coarse* segmentation of the moving objects due to the side effects of the motion regularization loss. Therefore, directly using the instance-level motion masks leads to noisy pseudo labels which might affect the performance of domain alignment. 2) There are some special cases where some moving instances might

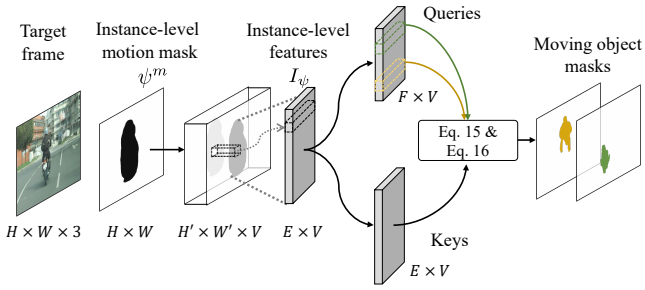


Figure 4. The instance-level motion mask might contain multiple moving objects bound together such as the *rider* and the *motorcycle*. The object discovery module takes an instance-level motion mask as input and predicts accurate object masks. Specifically, given a target image and its instance-level motion masks, we compute an objectness score map by computing the similarity of each query with all the keys in Eq. 8 and the processing in Eq. 9.

contain *multiple* objects. For example, a motorcycle and its rider (two objects) are bounded into one moving instance (presented in Fig. 4).

To tackle these two issues, we propose a self-supervised object discovery module (ODM) to learn the accurate moving object masks from the instance-level motion masks $\{\psi^m\}_{m=1}^M$, as shown in Fig. 4. First, given the target frame $x^{\mathbb{T}}$, a dense feature map $I \in \mathbb{R}^{H',W',V}$ is extracted from the segmentation network pre-trained on the source domain (shown in Sec. 4.1). Then we adopt a self-supervised design to promote the objectness in the features' attention. Given an instance-level motion mask $\psi^m \in \mathbb{B}^{H,W}$ of $x^{\mathbb{T}}$ (generated by Eq. 4), we bilinearly downsample ψ to the spatial size of I and select all the instance-level features that are covered by ψ^m, denoted by $I_\psi \in \mathbb{R}^{E,K}$, which is computed by

$$I_\psi = \Upsilon\big(I \odot \texttt{repmat}(\texttt{bd}(\psi^m))\big), \tag{7}$$

where bd represents the binear downsampling, repmat indicates the repeating operation that makes the shape of ψ^m to be same as I, $\odot$ is the element-wise production, and Υ is to select all non-zero feature vectors.

To generate the binary moving object masks, we construct the queries and the keys from I_ψ. Our queries $Q \in \mathbb{R}^{F,V}$ are generated by a bilinear downsampling of I_ψ where F is the downsampled size, and our keys $K \in \mathbb{R}^{E,V}$ are from I_ψ itself. Given a query $Q_e \in \mathbb{R}^V$ in Q, we calculate its cosine similarity with all keys in K. Thus, we produce an *objectness* score map $S \in \mathbb{R}^{E,F}$ by

$$S_{e,f} = \texttt{cosim}(Q_e, K_f), \tag{8}$$

where $K_f \in \mathbb{R}^K$ is the f^{th} key of K, and cosim represents the cosine similarity which is the dot product of two vectors with $\mathcal{L}_2$ normalization. Next, the *objectness* score map is transformed by a normalization into a soft map where the scores are adjusted into the range of $[0, 1]$. To extract

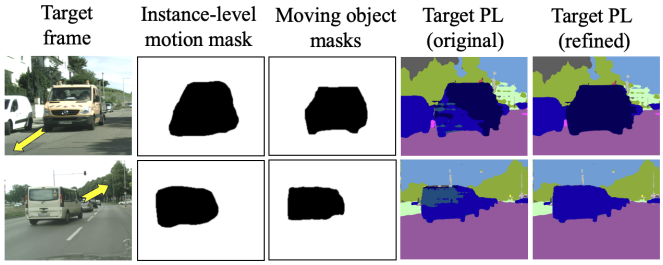


Figure 5. Directly utilizing instance-level motion masks might be sub-optimal as they are coarse masks for moving objects. The object discovery module is proposed to extract accurate object masks from coarse instance-level motion masks. The semantic mining module takes the moving object masks as guidance to refine the target pseudo labels.

the binary moving object masks, a threshold value τ is applied to the soft map. The resulting binary moving object masks are ranked based on their objectness scores, and any redundant masks are eliminated through non-maximum-suppression (NMS). The entire procedure to get the moving object masks $\{o^j\}_{j=1}^J$ is represented by

$$\{o^j\}_{j=1}^J = \texttt{NMS}\big(\texttt{rank}(\texttt{norm}(S))\big), \tag{9}$$

where $o^j \in \mathbb{B}^{H,W}$ and J is the number of the object masks predicted from ψ^m.

4.3. Target Semantic Mining

Our semantic mining module (SMM) takes the moving object masks as guidance to refine the noisy target pseudo labels. SMM is based on the assumption of *rigidity of the moving objects*, *e.g.*, vehicles, and motorbikes on traffic roads. For example, if a vehicle is moving, all parts of the vehicle are moving together. Based on the rigidity of the moving objects, all the pixels covered by a moving object mask in an image must have the same categorical label. Subsequently, we have the following remark:

Remark 1. *If a moving object mask is present, then the image pixels that it covers should have a semantic label that corresponds to the same moving categorical label.*

Given a target pseudo label $\tilde{y}'^{\mathbb{T}}$, we choose a dominant category c^* in $\tilde{y}'^{\mathbb{T}}$ that are covered by the moving object mask o^j. Concretely, c^* is determined by the moving category with the highest occurrence. Based on **Remark 1**, we introduce a semantic mining weight $w \in \mathbb{R}^{H,W,C}$ to update the target pseudo label via

$$w_{(i,c)} = \lambda o^j_{(i)} \mathbb{1}(c = c^*), \tag{10}$$

where $\mathbb{1}(\cdot)$ is the indicator function, $o^j_{(i)}$ is the object mask value on the pixel $x^{\mathbb{T}}_{(i)}$, and λ is a non-negative hyperparam-

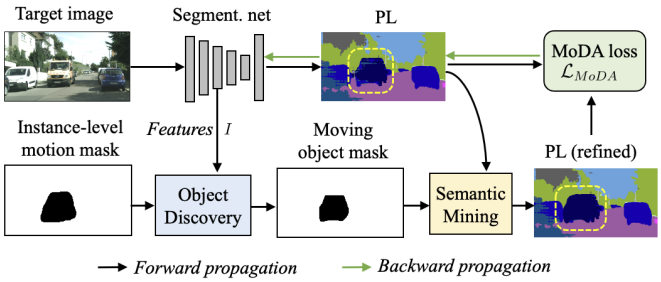


Figure 6. MoDA utilizes object motion information (instance-level motion masks) as cues to refine target pseudo labels. The key components of MoDA are the object discovery and the semantic mining module, which do not require any training. The object discovery module takes instance-level motion masks and extracts the moving object masks. These moving object masks are then sent as input to the semantic mining module to refine the target pseudo labels (PL). Note that the pseudo labels (PL) are generated by the segmentation network pre-trained at the warm-up stage.

eter for weighting. Then, we update the target pseudo label using the following equation

$$\tilde{y}'^{\mathbb{T}} = \Gamma(\texttt{softmax}((w+1) \odot p^{\mathbb{T}})), \tag{11}$$

where $\tilde{y}'^{\mathbb{T}}$ is the updated target pseudo label by our target semantic mining, and Γ is the process of taking the most probable category from $p^{\mathbb{T}}$. We provide an illustration of using object motion masks to update noisy target pseudo labels shown in Fig. 5. We use the updated target pseudo label to update the segmentation network G_e by $\mathcal{L}_{SMM}$ which is defined by

$$\mathcal{L}_{MoDA} = \sum_{i=1}^{H,W} \sum_{c=1}^{C} \tilde{y}'^{\mathbb{T}}_{(i,c)} \log\left(G_e(x^{\mathbb{T}})\right). \tag{12}$$

The overview of MoDA is shown in Fig. 6.

Optical Flow Regularization (OFR). Our baseline is to use the optical flow information from the adjacent target frames. Specifically, we can propagate the prediction of the previous frame to the current frame using optical flow estimates between the frames and subsequently ensure the consistency between the prediction of the current frame and the propagated prediction from the previous frame. Given two adjacent frames $\{x_1^{\mathbb{T}}, x_2^{\mathbb{T}}\}$, we forward them as input to get the predictions $\{p_1^{\mathbb{T}}, p_2^{\mathbb{T}}\}$. Moreover, we use FlowNet [14] to estimate the optical flow $f_{1\to2}^{\mathbb{T}}$ from $x_1^{\mathbb{T}}$ to $x_2^{\mathbb{T}}$. Then we warp the prediction $p_1^{\mathbb{T}}$ to the propagated prediction $\bar{p}_2^{\mathbb{T}}$. Then the optical flow regularization loss $\mathcal{L}_{OFR}$ is formulated as

$$\mathcal{L}_{OFR} = \|p_1^{\mathbb{T}} - \bar{p}_2^{\mathbb{T}}\|_2. \tag{13}$$

We conduct experiments and compare MoDA with the baseline method using optical flow regularization $\mathcal{L}_{OFR}$ that considers the temporal consistency of the target frames in Sec. 5.

5. Experiments

The datasets and implementation details are in Sec. 5.1. The evaluation is conducted on domain adaptive video segmentation and image segmentation task in Sec. 5.2. The ablation study and hyperparameter analysis are presented in Sec. 5.3.

5.1. Experiment Setup

Datasets. For *domain adaptive image segmentation*, we have GTA5 [27] as the source domains. GTA5 contains 24,966 training images with a resolution $2,048 \times 1,024$ and 19 categories. For *domain adaptive video segmentation*, we adopt VIPER [28] as the source domain. VIPER [28] contains $133,670$ synthetic frames with the corresponding pixel-wise annotations from 77 videos generated from game engines. For the target domain, we use Cityscapes-Seq [9, 39] which contains 500 video sequences from the Cityscapes dataset [5], and each sequence consists of 30 frames. We also include 500 validation images from the Cityscapes dataset for evaluation.

Implementation details. We conduct experiments on two types of architectures: CNN-based architecture and Transformer-based architecture. For CNN-based architectures, our warm-up stage follows DACS [30]. We first adopt DeepLab-V2 [4] with ReseNet-101 [11] for the segmentation network, pre-trained on ImageNet [6]. For Transformer-based architecture, we adopt MiT-B5 [38] as the encoder and incorporate our MoDA with existing state-of-the-art approaches [12, 13] by using the pre-trained weights from them. Our batch size is 16 with 8 source and 8 target images with the resolution $1,024 \times 512$. Threshold τ is set with 0.5. The optimizer for segmentation is SGD [2] with learning rate of $2.5e^{-4}$, momentum 0.9, and weight decay of 5×10^{-4}. We optimize the discriminator using Adam with the initial learning rate of 10^{-4}. For the momentum network, we set $\lambda = 0.99$. We implement MoDA with PyTorch and the training process is running on two Titan RTX A6000 GPUs.

5.2. Evaluation Results

Domain Adaptive Video Segmentation. We evaluate MoDA in the setting of domain adaptive video segmentation: *VIPER→Cityscapes-Seq*. Our baseline approaches are DA-VSN [9] and TPS [39] which have included optical flow regularization. We also include domain adaptive image segmentation baselines: AdvEnt [32], CBST [47], IntraDA [25], CRST [48], PixMatch [23], and DACS [30]. To make a fair comparison, all these domain adaptive image segmentation baselines are optimized with optical flow regularization ($\mathcal{L}_{OFR}$ in Eq. 13). For example, DACS + OFR represents the training results of DACS [30] with optical flow regularization. The experimental results

Table 1. The comparison with baseline methods on domain adaptive *video* segmentation benchmark VIPER→Cityscapes-Seq and domain adaptive *image* segmentation benchmark GTA5→Cityscapes-Seq. Current baseline methods are optimized with optical flow regularization (represented by *+OFR*) on the target frames. MoDA utilizes object motion as cues to guide domain adaptation and demonstrates superior performance against current baseline methods using optical flow regularization. This is to demonstrate that *object motion serves as stronger guidance for domain adaptation compared to optical flow*. We also show MoDA is *versatile* as it complements existing state-of-the-art domain adaptive image segmentation approach (*+MoDA*).

VIPER → Cityscapes-Seq

Method	Road	Side.	Buil.	Fence	TL	TS	Vege.	Terr.	Sky	Pers.	Car	Truck	Bus	Motor	Bike	mIoU
Backbone: ResNet-101																
AdvEnt + OFR [32]	78.6	30.0	79.9	23.9	27.3	28.1	82.2	13.0	81.2	59.5	62.3	6.4	40.3	4.8	2.7	41.3
CBST + OFR [47]	48.0	20.9	**85.6**	12.4	19.2	21.0	82.6	19.5	83.2	60.0	71.8	3.9	39.5	23.1	38.5	41.9
IntraDA + OFR [25]	80.4	35.1	80.7	**24.7**	28.2	24.7	82.5	14.0	79.8	60.0	63.1	6.0	41.8	6.4	4.2	42.1
CRST + OFR [48]	55.5	22.0	82.4	12.8	19.9	16.0	85.8	19.3	83.1	62.6	72.4	5.0	39.8	**29.6**	34.9	42.7
DACS + OFR [30]	71.2	26.8	83.5	20.7	23.3	24.1	82.9	24.9	81.8	58.4	76.3	28.8	41.3	24.7	22.8	46.1
PixMatch + OFR [23]	78.8	28.4	82.2	18.2	**30.8**	25.3	84.6	**31.4**	**83.3**	59.1	75.2	34.3	43.7	15.9	13.1	46.9
DA-VSN [9]	**87.1**	**38.3**	82.2	23.7	29.8	28.4	**85.9**	26.6	80.8	60.3	78.7	21.7	46.9	22.0	10.5	48.2
TPS [39]	83.4	35.8	78.9	9.6	25.7	**29.5**	77.9	28.4	81.6	60.2	81.1	40.7	39.8	27.7	31.4	48.7
MoDA	72.2	25.9	80.9	18.3	24.6	21.1	79.1	23.2	78.3	**68.7**	**84.1**	**43.2**	**49.5**	28.8	**38.6**	**49.1**

GTA5 → Cityscapes-Seq

Method	Road	Side.	Buil.	Wall	Fence	Pole	TL	TS	Vege.	Terr.	Sky	Pers.	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
Backbone: ResNet-101																				
AdatpSegNet + OFR [31]	86.3	36.2	79.8	24.1	23.9	24.1	36.0	15.2	82.6	31.9	74.4	58.7	26.8	75.1	33.9	35.8	4.1	29.7	28.8	42.5
AdvEnt + OFR [32]	91.6	54.0	79.8	32.4	21.5	33.6	29.1	21.7	84.2	35.4	81.7	52.9	23.8	81.9	31.0	35.3	16.8	26.2	**43.7**	46.1
IntraDA + OFR [25]	91.6	38.1	81.7	33.0	20.4	28.6	31.9	22.7	85.6	41.1	78.9	59.2	31.8	86.1	31.9	48.3	0.2	30.9	35.7	46.2
CRST+ OFR [48]	90.0	56.1	83.3	32.8	24.5	36.6	33.7	25.4	84.9	35.3	81.0	58.3	25.6	84.0	28.7	31.8	27.4	25.8	42.9	47.8
SIM+ OFR [35]	89.8	46.1	85.9	33.5	27.8	36.7	35.0	37.6	84.7	**45.4**	83.2	56.4	31.5	82.7	43.2	49.9	2.1	33.4	38.7	49.7
CAG-UDA + OFR [42]	91.2	50.3	84.1	37.5	26.9	37.6	26.2	49.3	86.2	37.9	77.4	55.8	37.9	86.2	20.1	43.2	**44.1**	28.6	36.5	50.4
IAST + OFR [22]	**92.5**	**59.3**	84.2	**40.7**	25.8	27.3	42.8	36.5	82.7	31.6	89.5	61.7	30.3	87.6	41.4	46.7	28.9	33.0	42.4	51.8
DACS + OFR [30]	90.4	40.1	**86.6**	31.4	**40.8**	39.2	**45.8**	53.4	88.6	42.7	89.5	66.1	35.7	85.6	46.7	51.3	0.2	28.7	34.6	52.5
MoDA	91.7	43.2	85.3	31.7	39.6	**40.5**	44.3	**54.8**	**89.7**	41.8	**90.4**	**68.5**	**41.4**	**87.9**	**48.1**	**56.7**	11.8	**35.3**	39.6	**54.9**
Backbone: Transformer																				
HRDA [12]	97.1	**74.5**	90.8	**62.2**	51.0	58.2	63.4	**70.5**	**92.0**	48.3	94.6	78.8	53.1	94.5	**83.6**	84.9	76.8	64.3	65.7	73.9
HRDA + **MoDA**	**97.3**	74.4	**91.4**	61.4	**51.4**	**59.2**	**64.1**	70.0	91.0	**49.5**	**95.9**	**80.1**	**57.1**	**95.1**	83.1	**89.4**	**77.5**	**72.0**	**68.8**	**75.2**

Table 2. Ablation study on the components of MoDA: object discovery module (ODM) and semantic mining module (SMM).

Configuration	Warm-up	SMM	ODM	mIoU	Gap
Only Warm-up	✓			45.9	-3.2
w/o Semantic Mining	✓		✓	45.9	-3.2
w/o Object Discovery	✓	✓		46.7	-2.4
Full Framework (**MoDA**)	✓	✓	✓	49.1	-

are shown in Table 1. MoDA outperforms the baseline DACS+OFR, which shows that *object motion is stronger guidance for domain adaptation compared with optical flow information*. Specifically, MoDA outperforms DA-VSN and TPS, indicating that *self-supervised object motion from unlabeled video sequences are important information to be considered in video domain adaptation setting*. We provide qualitative examples of MoDA and the baseline DACS+OFR in Fig. 7. We also visualize examples of object motion maps learned from the motion network as cues to refine target pseudo labels in Fig. 8.

Domain Adaptive Image Segmentation. We include existing ResNet-101 based approaches for comparison: AdaptSegNet [31], AdvEnt [32], IntraDA [25], SIM [35], CRST [48], CAG-UDA [42], IAST [22], DACS [30], and HRDA [12]. We evaluate the performance of MoDA in Table 1 (GTA5 → Cityscapes-Seq). To make a fair comparison, all the domain adaptive image segmentation baselines are optimized with optical flow regularization ($\mathcal{L}_{OFR}$ in Eq. 13). Experimental results on Table 1 demonstrate that MoDA outperforms existing domain adaptive image segmentation methods with optical flow regularization. This is to show that *the object motion information is a strong guidance for domain adaptation compared with optical flow information*. We also combine MoDA with current state-of-the-art approaches and the results are showing with **+MoDA**. For example, the results of HRDA**+MoDA** are obtained by using HRDA [12] as the warm-up stage to generate target pseudo labels and refine these pseudo labels with MoDA. This is to show that MoDA *complements existing state-of-the-art UDA approach*.

5.3. Ablation Study

Optical Flow Regularization. We provide an ablation study using object motion as guidance in comparison with the optical flow regularization. We show the evaluation of VIPER→Cityscapes-Seq in Table 3. DACS+OFR achieves 46.1% of mIoU by using optical flow regularization (OFR) which is 0.2% higher than the DACS model [30]. MoDA produces 49.1% of mIoU which is higher than DACS+OFR. This is to show that object motion functions as a stronger guidance compared with optical flow.

Different Components in MoDA. We conduct an ablation study on the effectiveness of the object discovery module

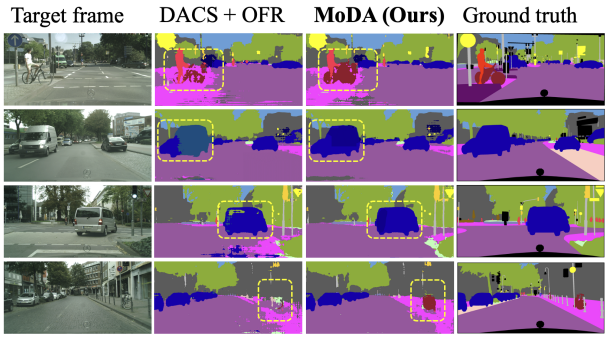


Figure 7. Comparison of the qualitative results generated from MoDA and the baseline model on VIPER→Cityscapes-Seq benchmark.

Table 3. We compare object motion with optical flow regularization after the same warm-up on VIPER→Cityscapes-Seq benchmark. MoDA adopts DACS as the warm-up step and utilizes object motion as cues to refine target pseudo labels. DACS+OFR uses DACS as the warm-up step and applies optical flow regularization. This is to show that object motion functions as a stronger guidance compared with optical flow.

Configuration	$\mathcal{L}_{ce}^{\mathbb{S}}$	$\mathcal{L}_{ce}^{\mathbb{T}}$	$\mathcal{L}_{MoDA}$	$\mathcal{L}_{OFR}$	mIoU	*Gain*
DACS [30]	✓	✓			45.9	-
DACS + OFR	✓	✓		✓	46.1	+0.2
MoDA	✓	✓	✓		49.1	+3.2

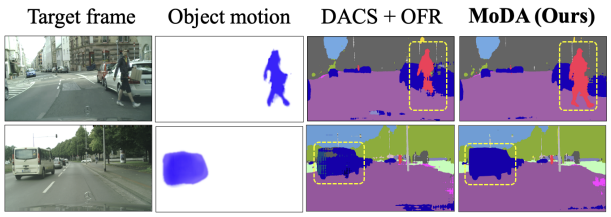


Figure 8. MoDA utilizes the object motion as cues to guide domain adaptation for segmentation task. MoDA outperforms the baseline model DACS + OFR [30] which adopts the optical flow regularization. This is to show that the object motion information is a strong cue to guide adaptation compared with optical flow information.

(ODM) and semantic mining module (SMM) in Table 2. On VIPER→Cityscapes-Seq benchmark, by only using the warm-up step (without using ODM or SMM), we get the score 45.9% of mIoU. On the other hand, utilizing the warm-up along with SMM (without using ODM), MoDA experienced a significant performance drop to 46.7% of mIoU. It is caused by two reasons: *1) The instance-level motion masks are not accurate to be used as the object masks in the semantic mining module; 2) Some instance-level motion masks contain multiple moving objects bounded together such as a motorcycle with a rider.*

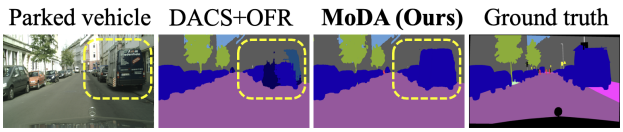


Figure 9. MoDA utilizes object motion information as guidance for domain adaptation. It is effective both for the moving and the static objects, such as the *parked* vehicle in the yellow box.

Table 4. The ablation study on the hyperparameter λ for weighting in semantic mining module.

λ	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4
mIoU	45.9	52.9	53.4	54.1	49.1	49.1	49.1	49.1

What happens to the potentially movable, but static objects (*e.g.*, parked cars, standing persons)? First of all, the overall training pipeline of our approach does not harm static objects like parked cars or standing pedestrians during the domain transfer. Since MoDA uses motion-guided object masks to update the noisy predictions of the pseudo labels, the performance on static objects will also be upgraded by updating the segmentation net with these new pseudo labels. As an example in Fig. 9, MoDA generates more accurate predictions on the parked vehicle that is not moving in comparison with the baseline DACS+OFR [30].

Hyperparameter λ. We conduct an ablation study on the hyperparameter λ in semantic mining module(Eq. 10). We present different values of λ for the final performance in GTA5→Cityscapes-Seq in Table 4. The bigger value of λ puts more weight on semantic mining to update target pseudo labels. Our ablation results indicate that MoDA reaches the best performance when λ reaches 0.8. Additionally, it shows that MoDA's performance is insensitive to the hyperparameter λ value when it is greater than 0.8.

6. Conclusion

This paper proposed a novel motion-guide domain adaption method, namely MoDA for the semantic segmentation task. MoDA utilizes the self-supervised object motion as cues to guide cross-domain alignment for segmentation task. MoDA consists of two modules namely, object discovery and semantic mining, to refine target pseudo labels. These refined pseudo labels are used in the self-training loop to bridge the cross-domain gap. On domain adaptive image and video segmentation experiments MoDA shows the effectiveness utilizing object motion as guidance for domain alignment compared with optical flow information. Moreover, MoDA is versatile as it complements existing state-of-the-art UDA approaches.

References

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15384–15394, Virtual/Online, 2021. 1, 2

[2] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Int. Conf. on Comput. Statis.*, pages 177–186, Paris, France, 2010. Springer. 6

[3] Zhe Cao, Abhishek Kar, Christian Hane, and Jitendra Malik. Learning independent object motion from unlabelled stereoscopic videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3

[4] 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. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017. 1, 6

[5] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3213–3223, Las Vegas, Nevada, 2016. 1, 6

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, Miami, USA, 2009. Ieee. 6

[7] Yuefang Gao, Yiteng Cai, Xuanming Bi, Bizheng Li, Shunpeng Li, and Weiping Zheng. Cross-domain facial expression recognition through reliable global–local representation learning and dynamic label weighting. *Electronics*, 12(21): 4553, 2023. 2

[8] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Int. Conf. Comput. Vis.*, pages 8977–8986, Seoul, South Korea, 2019. 1, 3

[9] Dayan Guan, Jiaxing Huang, Aoran Xiao, and Shijian Lu. Domain adaptive video segmentation via temporal consistency regularization. In *Int. Conf. Comput. Vis.*, pages 8053–8064, Virtual/Online, 2021. 2, 6, 7

[10] Vitor Guizilini, Jie Li, Rareș Ambruș, and Adrien Gaidon. Geometric unsupervised domain adaptation for semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 8537–8547, Virtual/Online, 2021. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, Las Vegas, Nevada, 2016. 6

[12] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Eur. Conf. Comput. Vis.*, pages 372–391, Tel Aviv, Israel, 2022. Springer. 2, 6, 7

[13] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9924–9935, New Orleans, US, 2022. 2, 6

[14] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2462–2470, Hawaii, US, 2017. 6

[15] Longfei Jia, Xianlong Tian, Yuguo Hu, Mengmeng Jing, Lin Zuo, and Wen Li. Style-guided adversarial teacher for cross-domain object detection. *Electronics*, 13(5):862, 2024. 2

[16] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12975–12984, Seattle, US, 2020. 2

[17] Seokju Lee, Francois Rameau, Fei Pan, and In So Kweon. Attentive and contrastive learning for joint depth and motion field estimation. In *Int. Conf. Comput. Vis.*, pages 4862–4871, Virtual/Online, 2021. 1, 3, 4

[18] Hanhan Li, Ariel Gordon, Hang Zhao, Vincent Casser, and Anelia Angelova. Unsupervised monocular depth learning in dynamic scenes. In *Conf. on Robot Learn.*, pages 1908–1917, London, UK, 2021. PMLR. 1, 2, 3, 4

[19] Rui Li, Wenming Cao, Si Wu, and Hau-San Wong. Generating target image-label pairs for unsupervised domain adaptation. *IEEE Trans. Image Process.*, 29:7997–8011, 2020. 2

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3431–3440, Boston, US, 2015. 1

[21] Zhihe Lu, Da Li, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Uncertainty-aware source-free domain adaptive semantic segmentation. *IEEE Trans. Image Process.*, pages 1–1, 2023. 2

[22] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *Eur. Conf. Comput. Vis.*, pages 415–430, Glasgow, UK, 2020. Springer. 7

[23] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12435–12445, Virtual/Online, 2021. 6, 7

[24] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 122–132, Seattle, US, 2020. 2

[25] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3764–3773, Seattle, US, 2020. 1, 2, 6, 7

[26] Fei Pan, Sungsu Hur, Seokju Lee, Junsik Kim, and In So Kweon. Ml-bpm: Multi-teacher learning with bidirectional photometric mixing for open compound domain adaptation in semantic segmentation. In *Eur. Conf. Comput. Vis.*, pages 236–251, Tel Aviv, Israel, 2022. Springer. 2

[27] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer

games. In *Eur. Conf. Comput. Vis.*, pages 102–118, Amsterdam, US, 2016. Springer. 6

[28] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *Int. Conf. Comput. Vis.*, 2017. 6

[29] Inkyu Shin, Sanghyun Woo, Fei Pan, and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In *Eur. Conf. Comput. Vis.*, pages 532–548, Glasgow, United Kingdom, 2020. Springer. 2

[30] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proc. of IEEE/CVF Wint. Conf. on Appli. of Comput. Visi.*, pages 1379–1389, Hawaii, USA., 2021. 1, 2, 4, 6, 7, 8

[31] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7472–7481, Salt Lake City, US, 2018. 1, 2, 7

[32] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2517–2526, Long Beach, US, 2019. 1, 6, 7

[33] Qi Wang, Junyu Gao, and Xuelong Li. Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes. *IEEE Trans. Image Process.*, 28(9):4376–4386, 2019. 2

[34] Shuang Wang, Dong Zhao, Chi Zhang, Yuwei Guo, Qi Zang, Yu Gu, Yi Li, and Licheng Jiao. Cluster alignment with target knowledge mining for unsupervised domain adaptation semantic segmentation. *IEEE Trans. Image Process.*, 31: 7403–7418, 2022. 2

[35] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12635–12644, Seattle, US, 2020. 7

[36] Kesheng Wu, Ekow Otoo, and Arie Shoshani. Optimizing connected component labeling algorithms. In *Medic. Imagi.: Image Processing*. SPIE, 2005. 4

[37] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023. 2

[38] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inform. Process. Syst.*, 34:12077–12090, 2021. 6

[39] Yun Xing, Dayan Guan, Jiaxing Huang, and Shijian Lu. Domain adaptive video segmentation via temporal pseudo supervision. In *Eur. Conf. Comput. Vis.*, pages 621–639, Tel Aviv, Israel, 2022. Springer. 2, 6, 7

[40] Bo Yuan, Danpei Zhao, Shuai Shao, Zehuan Yuan, and Changhu Wang. Birds of a feather flock together: Category-divergence guidance for domain adaptive segmentation. *IEEE Trans. Image Process.*, 31:2878–2892, 2022. 2

[41] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12414–12424, Virtual/Online, 2021. 2

[42] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *Adv. Neural Inform. Process. Syst.*, 32, 2019. 7

[43] Yuyang Zhao, Zhun Zhong, Zhiming Luo, Gim Hee Lee, and Nicu Sebe. Source-free open compound domain adaptation in semantic segmentation. *IEEE Trans. Circuit Syst. Video Technol.*, 32(10):7019–7032, 2022. 2

[44] Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. 2019. 2

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

[46] Wei Zhou, Yukang Wang, Jiajia Chu, Jiehua Yang, Xiang Bai, and Yongchao Xu. Affinity space adaptation for semantic segmentation across domains. *IEEE Trans. Image Process.*, 30:2549–2561, 2021. 2

[47] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Eur. Conf. Comput. Vis.*, pages 289–305, Munich, Germany, 2018. 1, 2, 6, 7

[48] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Int. Conf. Comput. Vis.*, pages 5982–5991, Seoul, South Korea, 2019. 1, 2, 6, 7