

HRDA: Context-Aware High-Resolution Domain-Adaptive Semantic Segmentation

Lukas Hoyer¹, Dengxin Dai², and Luc Van Gool^{1,3}

¹ ETH Zurich, Switzerland {lhoyer,vangool}@vision.ee.ethz.ch

² MPI for Informatics, Germany ddai@mpi-inf.mpg.de

³ KU Leuven, Belgium

Abstract. Unsupervised domain adaptation (UDA) aims to adapt a model trained on the source domain (e.g. synthetic data) to the target domain (e.g. real-world data) without requiring further annotations on the target domain. This work focuses on UDA for semantic segmentation as real-world pixel-wise annotations are particularly expensive to acquire. As UDA methods for semantic segmentation are usually GPU memory intensive, most previous methods operate only on downscaled images. We question this design as low-resolution predictions often fail to preserve fine details. The alternative of training with random crops of high-resolution images alleviates this problem but falls short in capturing long-range, domain-robust context information. Therefore, we propose HRDA, a multi-resolution training approach for UDA, that combines the strengths of small high-resolution crops to preserve fine segmentation details and large low-resolution crops to capture long-range context dependencies with a learned scale attention, while maintaining a manageable GPU memory footprint. HRDA enables adapting small objects and preserving fine segmentation details. It significantly improves the state-of-the-art performance by 5.5 mIoU for GTA→Cityscapes and 4.9 mIoU for Synthia→Cityscapes, resulting in unprecedented 73.8 and 65.8 mIoU, respectively. The implementation is available at github.com/lhoyer/HRDA.

Keywords: Unsupervised Domain Adaptation; Semantic Segmentation; Multi-Resolution; High-Resolution; Attention

1 Introduction

Even though neural networks currently are the unchallenged approach to solve many computer vision problems, their training often requires a large amount of annotated data. For certain tasks, such as semantic segmentation, providing the annotations is particularly labor-intensive as pixel-wise labels of the entire image are necessary, which can take more than one hour per image [11, 51]. Therefore, several methods aim to reduce the annotation burden such as weakly-supervised learning [14, 61, 93], semi-supervised learning [17, 27, 35, 53], and unsupervised domain adaptation (UDA) [26, 28, 59, 67, 92]. In this work, we focus on UDA. To avoid the annotation effort for the target dataset, the network is trained on a source dataset with existing or cheaper annotations such as automatically labeled

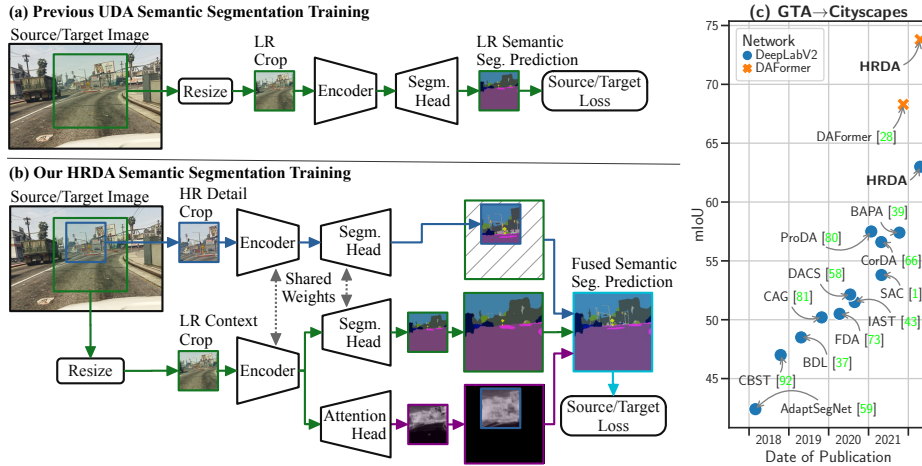


Fig. 1. (a) Most previous UDA methods were only trained with downscaled inputs to account for their high GPU memory footprint. (b) Our HRDA incorporates fine segmentation details from a random high-resolution (HR) detail crop and context information from a random low-resolution (LR) context crop. Their predictions are fused using a learned scale attention (best viewed zoomed-in). In that way, HRDA can utilize HR details and long-range context information while keeping a manageable memory footprint. (c) Compared to previous works, HRDA provides a major performance gain.

synthetic data [47, 49]. However, neural networks are usually sensitive to domain shifts. This problem is approached in UDA by adapting the network, which is trained with source data, to unlabeled target images.

UDA methods are usually more GPU memory intensive than regular supervised learning as UDA training often requires images from multiple domains, additional networks (e.g. teacher model or domain discriminator), and additional losses, which consume significant additional GPU memory. Therefore, most UDA semantic segmentation methods so far (e.g. [1, 28, 39, 58, 59, 66, 73, 90]) follow the convention of downscaling images due to GPU memory constraints (see Fig. 1 a). Taking Cityscapes as an example, current UDA methods use half the full resolution (i.e. 1024×512 pixels), while most supervised methods use the full resolution (i.e. 2048×1024 pixels). This is one of the key differences in the training setting of UDA and supervised semantic segmentation, possibly contributing to the gap between the state-of-the-art performance of UDA and supervised learning.

We question this design choice as predictions from low-resolution (LR) inputs often fail to recognize small objects such as distant traffic lights and to preserve fine segmentation details such as limbs of distant pedestrians. However, naively learning UDA with full high-resolution (HR) images is often difficult as the resolution quadratically affects the GPU memory consumption. A common remedy is training with random crops of the image. While training with HR crops allows to adapt small objects and preserve segmentation details, it limits the learned long-range context dependencies to the crop size. This is a crucial disadvantage

for UDA as context information and scene layout are often domain-robust (e.g. rider on bicycle or sidewalk at the side of road) [31, 71, 89]. Further, while HR inputs are necessary to adapt small objects, they can be disadvantageous compared to LR inputs when adapting large stuff-regions such as close sidewalks (see Sec. 5.4). At HR, these regions often contain too detailed, domain-specific features (e.g. detailed sidewalk texture), which are detrimental to UDA. LR inputs ‘hide away’ these features, while still providing sufficient details to recognize large regions across domains.

To effectively combine the strength of these two approaches while maintaining a manageable GPU memory footprint, we propose *HRDA*, a novel multi-resolution framework for UDA semantic segmentation (see Fig. 1 b). First, HRDA uses a large LR *context crop* to adapt large objects without confusion from domain-specific HR textures and to learn long-range context dependencies as we assume that HR details are not crucial for long-range dependencies. Second, HRDA uses a small HR *detail crop* from the region within the context crop to adapt small objects and to preserve segmentation details as we assume that long-range context information play only a subordinate role in learning segmentation details. In that way, the GPU memory consumption is significantly reduced while still preserving the main advantages of a large crop size and a high resolution. Given that the importance of the LR context crop vs. the HR detail crop depends on the content of the image, HRDA fuses both using an input-dependent scale attention. During UDA, the attention learns to decide how trustworthy the LR and the HR predictions are in every image region. Previous multi-resolution frameworks for supervised learning [5, 56, 72] cannot naively be applied to state-of-the-art UDA due to GPU memory constraints as they operate on full LR and HR images. To adapt HRDA to the target domain, it can be trained with pseudo-labels fused from multiple resolutions. To further increase the robustness of the detail pseudo-labels with respect to different contexts, they are generated using an overlapping sliding window mechanism.

This work makes four contributions. To the best of our knowledge, it is the first work on UDA semantic segmentation (1) systematically studying the influence of resolution and crop size, (2) exploiting HR inputs for adapting small objects and fine segmentation details, (3) applying multi-resolution fusion with a learned scale attention for object-scale-dependent adaptation, and (4) fusing a nested large LR crop to capture long-range context and small HR crop to capture details for memory-efficient UDA training. HRDA provides significant performance gains when applied to various UDA strategies [28, 58, 59, 62]. When used with the SOTA method DAFormer [28], HRDA gains +5.5 mIoU for GTA→Cityscapes and +4.9 mIoU for Synthia→Cityscapes, resulting in unprecedented 73.8 and 65.8 mIoU, respectively (see Fig. 1 c).

2 Related Work

Semantic Segmentation: Most semantic segmentation approaches are based on (convolutional) neural networks, which can be effectively trained in an end-to-

end manner to perform pixel-wise classification as first shown by Long et al. [41]. This concept was further improved in different aspects including increasing the receptive field while preserving spatial details [2, 3, 6, 48, 65, 74, 83], integrating context information [30, 75, 76, 78, 91], attention mechanisms [18, 32, 69, 84], refining boundaries [15, 36, 77], and Transformer-based architectures [40, 54, 68, 70, 85].

Several architectures [3, 4, 16, 38, 83] utilize intermediate features with different scales, which are generated from a single scale input, to aggregate context information. Furthermore, multi-scale input inference, where predictions from scaled versions of an image are combined via average or max pooling, is often used to obtain better results [4, 6, 9, 75]. However, the naive pooling is independent of the image content, which can lead to suboptimal results. Therefore, Chen et al. [5] and Yang et al. [72] segment multi-scale inputs and learn an attention-weighted sum of the predictions. Tao et al. [56] further propose a hierarchical attention that is agnostic to the number of scales during inference.

Unsupervised Domain Adaptation (UDA): To adapt a semantic segmentation network to the target domain, several strategies have been proposed, while most can be grouped into adversarial training and self-training. In adversarial training [19, 22], a domain discriminator is trained in order to provide supervision to align the domain distributions based on style-transferred inputs [21, 25, 37, 45] or network features/outputs [26, 59, 60, 62, 64]. In self-training, the network is adapted to the target domain using high-confidence pseudo-labels. In order to regularize the training and to avoid pseudo-label drift, approaches such as confidence thresholding [43, 92], pseudo-label prototypes [39, 80, 81], and consistency regularization [50, 52, 57] based on data augmentation [1, 10, 44, 46], different context [35, 89], domain-mixup [20, 29, 39, 58, 89], or multiple models [79, 87, 88] have been used. Several works also combine self-training and adversarial training [34, 37, 64, 86], minimize the entropy [7, 62, 63], refine boundaries [39], use a curriculum [12, 13, 82], or learn auxiliary tasks [8, 29, 63, 66]. The use of semantic segmentation networks with multi-scale *features* is quite common in UDA as many methods evaluate their approach with DeepLabV2 [3]. However, these features are generated from a single-scale input. While some works apply multi-scale average pooling for inference [1, 64] or enforce scale consistency [23, 33, 55] of low-resolution inputs, they fall short in learning an input-adaptive multi-scale fusion. To the best of our knowledge, HRDA is the first work to learn a multi-resolution *input* fusion for UDA semantic segmentation. For that purpose, we newly extend scale attention [5, 56] to UDA and reveal its significance for UDA by improving the adaptation process across different object scales. Further, we propose fusing nested crops with different scales and sizes, which successfully overcomes the pressing issue of limited GPU memory for multi-resolution UDA.

3 Preliminary

In UDA, a neural network f_θ is trained using source domain images $\mathcal{X}^S = \{x_{HR}^{S,m}\}_{m=1}^{N_S}$ with $x_{HR}^{S,m} \in \mathbb{R}^{H_S \times W_S \times 3}$ and target domain images $\mathcal{X}^T = \{x_{HR}^{T,m}\}_{m=1}^{N_T}$ with $x_{HR}^{T,m} \in \mathbb{R}^{H_T \times W_T \times 3}$ to achieve a good performance on the target domain.

However, labels are only available for the source domain $\mathcal{Y}^S = \{y_{HR}^{S,m}\}_{m=1}^{N_S}$ with $y_{HR}^{S,m} \in \{0, 1\}^{H_S \times W_S \times C}$. As the following definitions refer to the same source/target sample, we will drop index m to avoid convolution. Most previous UDA methods bilinearly downsample $\zeta(\cdot, \cdot)$ the images and labels x_{HR}^S, x_{HR}^T , and y_{HR}^S by a dataset-specific factor $s_S, s_T \geq 1$ to satisfy GPU memory constraints, e.g. $x_{LR}^T = \zeta(x_{HR}^T, 1/s_T) \in \mathbb{R}^{\frac{H_T}{s_T} \times \frac{W_T}{s_T} \times 3}$. Some methods such as [28, 58, 80] additionally crop the LR image but for simplicity, we consider full LR images in this section.

As only source labels are available, the supervised categorical cross-entropy loss can only be calculated for the source predictions $\hat{y}_{LR}^S = f_\theta(x_{LR}^S)$

$$\mathcal{L}^S = \mathcal{L}_{ce}(\hat{y}_{LR}^S, y_{LR}^S, 1), \quad (1)$$

$$\mathcal{L}_{ce}(\hat{y}, y, q) = - \sum_{i=1}^{H(y)} \sum_{j=1}^{W(y)} \sum_{c=1}^C q_{ij} y_{ijc} \log \zeta(\hat{y}, \frac{H(y)}{H(\hat{y})})_{ijc}. \quad (2)$$

As the predictions are usually smaller than the input due to the output stride of the segmentation network, they are upsampled to the label size $H(y) \times W(y)$.

However, a model trained only with \mathcal{L}^S usually does not generalize well to the target domain. In order to adapt the model to the target domain, UDA methods incorporate an additional loss for the target domain \mathcal{L}^T , which is added to the overall loss $\mathcal{L} = \mathcal{L}^S + \lambda \mathcal{L}^T$. The target loss can be defined according to the used training strategies such as adversarial training [59, 60, 64] or self-training [28, 43, 58, 80, 81, 92]. In this work, we mainly evaluate HRDA with the self-training method DAFormer [28], as it is the current state-of-the-art method for UDA semantic segmentation. In self-training, the model is iteratively adapted to the target domain by training it with pseudo-labels for target images, predicted by a teacher network g_ϕ :

$$p_{LR,ijc}^T = [c = \arg \max_{c'} g_\phi(x_{LR}^T)_{ijc'}], \quad (3)$$

where $[\cdot]$ denotes the Iverson bracket. The pseudo-labels are used to additionally train the network f_θ on the target domain

$$\mathcal{L}^T = \mathcal{L}_{ce}(\hat{y}_{LR}^T, p_{LR}^T, q_{LR}^T). \quad (4)$$

As the pseudo-labels are not necessarily correct, their quality is weighted by a confidence estimate q_{LR}^T [28, 43, 58, 92]. After each training step t , the teacher model g_ϕ is updated with the exponentially moving average of the weights of f_θ , implementing a temporal ensemble to stabilize pseudo-labels, which is a common strategy in semi-supervised learning [17, 52, 57] and UDA [1, 39, 58]

$$\phi_{t+1} \leftarrow \alpha \phi_t + (1 - \alpha) \theta_t. \quad (5)$$

Further, DAFormer [28] uses consistency training [50, 52, 57], i.e. the network f_θ is trained on augmented target data following DACS [58], while the teacher model g_ϕ generates the pseudo-labels using non-augmented target images. Besides self-training, DAFormer [28] uses a domain-robust Transformer network, rare class sampling, and feature regularization based on ImageNet features.

4 Methods

In this work, we propose a multi-resolution framework for UDA as small objects and segmentation details are easier to adapt with high-resolution (HR) inputs, while large stuff regions are easier to adapt with low-resolution (LR) inputs. As UDA methods require more GPU memory than regular supervised training, we design a training strategy based on a large LR context crop to learn long-range context dependencies and a small HR detail crop to preserve segmentation details while maintaining a manageable GPU memory footprint (Sec. 4.1). The strengths of both LR context and HR detail crop are combined by fusing their predictions with a learned scale attention (Sec. 4.2). For a robust pseudo-label generation, we further utilize overlapping slide inference to fuse predictions with different contexts (Sec. 4.3). The proposed method is designed to be applicable to common network architectures and can be combined with existing UDA methods.

4.1 Context and Detail Crop

Due to GPU memory constraints, it is not feasible to train state-of-the-art UDA methods with full-sized high-/multi-resolution inputs as images from multiple domains, additional networks, and additional losses are required for UDA training. Therefore, most previous works only use LR inputs. However, HR inputs are important to recognize small objects and produce fine segmentation borders. In order to still utilize HR inputs, random cropping is a possible solution. However, random cropping restricts learning context-aware semantic segmentation, especially for long-range dependencies and scene layout, which might be critical for UDA as context relations are often domain-invariant (e.g. car on road, rider on bicycle) [31, 71, 89]. In order to both train with long-range context as well as high resolution, we propose to combine different crop sizes for different resolutions, i.e. a large LR context crop and a small HR detail crop (see Fig. 2 a). The purpose of the context crop is to provide a large crop to learn long-range context relations. The purpose of detail crop is to focus on HR to recognize small objects and produce fine segmentation details, which does not necessarily require far-away context information. In order to segment the entire image during model validation, overlapping sliding window inference is used (see Sec. 4.3).

The context crop $x_c \in \mathbb{R}^{h_c \times w_c \times 3}$ is obtained by cropping from the original HR image $x_{HR} \in \mathbb{R}^{H \times W \times 3}$ and bilinear downsampling by the factor $s \geq 1$

$$x_{c,HR} = x_{HR}[b_{c,1} : b_{c,2}, b_{c,3} : b_{c,4}], \quad x_c = \zeta(x_{c,HR}, 1/s) \quad (6)$$

The crop bounding box b_c is randomly sampled from a discrete uniform distribution within the image size while ensuring that the coordinates can be divided by $k = s \cdot o$ with $o \geq 1$ denoting the output stride of the segmentation network to ensure exact alignment in the later fusion process:

$$\begin{aligned} b_{c,1} &\sim \mathcal{U}\{0, (H - sh_c)/k\} \cdot k, & b_{c,2} &= b_{c,1} + sh_c, \\ b_{c,3} &\sim \mathcal{U}\{0, (W - sw_c)/k\} \cdot k, & b_{c,4} &= b_{c,3} + sw_c. \end{aligned} \quad (7)$$

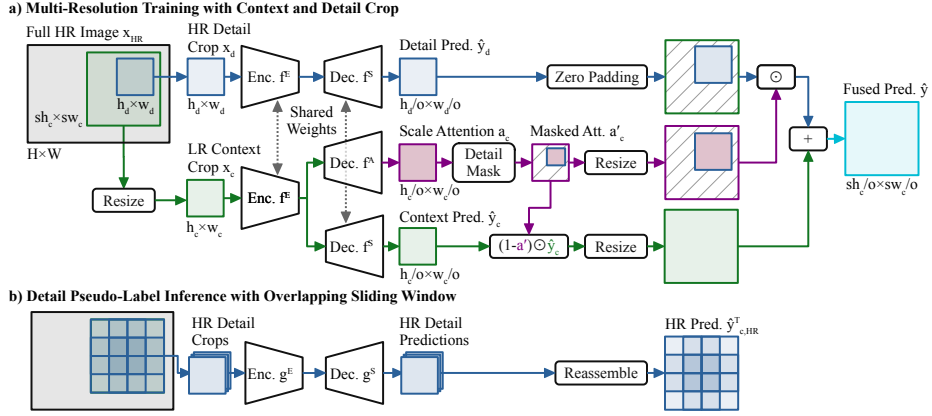


Fig. 2. (a) Multi-resolution training with low-resolution (LR) context and high-resolution (HR) detail crop. The prediction of the detail crop is fused into the context prediction within the region where it was cropped from by a learned scale attention. (b) For pseudo-label generation, multiple detail crops are generated using overlapping slide inference to cover the entire context. The pseudo-label is fused from HR pred. $\hat{y}_{c,HR}^T$ and LR pred. \hat{y}_c^T with the full attention a_c^T similar to (a) (see Sec. 4.3).

The detail crop $x_d \in \mathbb{R}^{h_d \times w_d \times 3}$ is randomly cropped from within the context crop region, which is necessary to enable the fusion of context and detail predictions in the later process:

$$x_d = x_{c,HR}[b_{d,1} : b_{d,2}, b_{d,3} : b_{d,4}], \quad (8)$$

$$b_{d,1} \sim \mathcal{U}\{0, (sh_c - h_d)/k\} \cdot k, \quad b_{d,2} = b_{d,1} + h_d, \quad (9)$$

$$b_{d,3} \sim \mathcal{U}\{0, (sw_c - w_d)/k\} \cdot k, \quad b_{d,4} = b_{d,3} + w_d.$$

In this work, we use context and detail crops of the same dimension, i.e. $h_c = h_d$ and $w_c = w_d$, to balance the required resources for both crops and provide a good trade-off between context-aware and detailed predictions. The context downscale factor is $s = 2$ following the LR design of previous UDA methods [28, 39, 58], which means that the context crop covers 4 times more content at half the resolution compared to the detail crop.

Using a feature encoder f^E and a semantic decoder f^S , the context semantic segmentation $\hat{y}_c = f^S(f^E(x_c)) \in \mathbb{R}^{\frac{h_c}{o} \times \frac{w_c}{o} \times C}$ and the detail semantic segmentation $\hat{y}_d = f^S(f^E(x_d)) \in \mathbb{R}^{\frac{h_d}{o} \times \frac{w_d}{o} \times C}$ are predicted. The networks f^E and f^S are shared for both HR and LR inputs. This not only saves memory usage but also increases the robustness of the network against different resolutions.

4.2 Multi-Resolution Fusion

While the HR detail crop is well-suited to segment small objects such as poles or distant pedestrians, it lacks the ability to capture long-range dependencies,

which is disadvantageous in segmenting large stuff regions such as large regions of sidewalk. The opposite is the case for the LR context crop. Therefore, we fuse the predictions from both crops using a learned scale attention [5] to predict in which image regions to trust predictions from context and detail crop. Additionally, the scale attention provides the advantage that it enables adapting objects at the better-suited scale. For example, small objects are easier to adapt at HR while large objects are easier to adapt at LR as the appearance of an object should have a resolution high enough to be discriminative but not too high to avoid that the network overfits to domain-specific detailed textures.

The scale attention decoder f^A learns to predict the scale attention $a_c = \sigma(f^A(f^E(x_c))) \in [0, 1]^{\frac{h_c}{o} \times \frac{w_c}{o} \times C}$ to weigh the trustworthiness of LR context and HR detail predictions. The sigmoid function σ ensures a weight in $[0, 1]$, where 1 means a focus on the HR detail crop. The attention is predicted from the context crop as it has a better grasp on the scene layout (larger context). As the predictions are smaller than the inputs due to the output stride o , the crop coordinates are scaled accordingly in the following steps. Outside of the detail crop c_d , the attention is set to 0 as there is no detail prediction available

$$a'_c \in \mathbb{R}^{\frac{h_c}{o} \times \frac{w_c}{o}}, \quad a'_c(i, j) = \begin{cases} a_c(i, j) & \text{if } \frac{b_{d,1}}{s \cdot o} \leq i < \frac{b_{d,2}}{s \cdot o} \wedge \frac{b_{d,3}}{s \cdot o} \leq j < \frac{b_{d,4}}{s \cdot o} \\ 0 & \text{otherwise} \end{cases}. \quad (10)$$

The detail crop is aligned with the (upsampled) context crop by padding it with zeros to a size of $\frac{sh_c}{o} \times \frac{sw_c}{o}$

$$\hat{y}'_d(i, j) = \begin{cases} \hat{y}_d(i - \frac{b_{d,1}}{o}, j - \frac{b_{d,3}}{o}) & \text{if } \frac{b_{d,1}}{o} \leq i < \frac{b_{d,2}}{o} \wedge \frac{b_{d,3}}{o} \leq j < \frac{b_{d,4}}{o} \\ 0 & \text{otherwise} \end{cases}. \quad (11)$$

The predictions from multiple scales are fused using the attention-weighted sum

$$\hat{y}_{c,F} = \zeta((1 - a'_c) \odot \hat{y}_c, s) + \zeta(a'_c, s) \odot \hat{y}'_d. \quad (12)$$

The encoder, segmentation head, and attention head are trained with the fused multi-scale prediction and the detail crop prediction

$$\mathcal{L}_{HRDA}^S = (1 - \lambda_d) \mathcal{L}_{ce}(\hat{y}_{c,F}^S, y_{c,HR}^S, 1) + \lambda_d \mathcal{L}_{ce}(\hat{y}'_d^S, y_d^S, 1), \quad (13)$$

where the ground truth $y_{c,HR}^S/y_d^S$ is cropped according to Eq. 6/8. Additionally supervising the detail crop is helpful to learn more robust features for HR inputs even though the attention might favor the context crop in that region. An additional loss for \hat{y}_c is not necessary as it is already directly supervised in regions without detail crop (see Eq. 10). The target loss \mathcal{L}_{HRDA}^T is adapted accordingly

$$\mathcal{L}_{HRDA}^T = (1 - \lambda_d) \mathcal{L}_{ce}(\hat{y}_{c,F}^T, p_{c,F}^T, q_{c,F}^T) + \lambda_d \mathcal{L}_{ce}(\hat{y}'_d^T, p_d^T, q_d^T). \quad (14)$$

For pseudo-label prediction, we also utilize multi-resolution fusion (see Sec. 4.3). Therefore, when predicting pseudo-labels the scale attention focuses on the better-suited resolution (e.g. HR for small objects). As the pseudo-labels are further used to train the model also with the worse-suited resolution (e.g. LR for small objects), it improves the robustness for both small and large objects.

4.3 Pseudo-Label Generation with Overlapping Sliding Window

For self-training with Eq. 14, it is necessary to generate a high-quality HRDA pseudo-label $p_{c,F}^T$ for the context crop $x_{c,HR}^T$. The underlying HRDA prediction $\hat{y}_{c,F}^T$ is fused from the LR prediction \hat{y}_c^T and HR prediction $\hat{y}_{c,HR}^T$ using the full scale attention a_c^T similar to Eq. 12

$$\hat{y}_{c,F}^T = \zeta((1 - a_c^T) \odot \hat{y}_c^T, s) + \zeta(a_c^T, s) \odot \hat{y}_{c,HR}^T. \quad (15)$$

Therefore, the HR prediction $\hat{y}_{c,HR}^T$ is necessary for the entire context crop $x_{c,HR}^T$ instead of just the detail crop x_d . Note that for pseudo-label generation g_ϕ instead of f_θ is used for predictions. Even though large HR network inputs are problematic during training, they are not an issue during pseudo-label inference as no data for the backpropagation has to be stored. However, the used DAFormer [28], as well as other Vision Transformers [70, 85], have a learned (implicit) positional embedding that works best if training and inference input size are the same. Therefore, we infer the HR prediction $\hat{y}_{c,HR}^T$ using a sliding window of size $h_d \times w_d$ over the HR context crop $x_{c,HR}^T$ (see Fig 2 b). The window is shifted with a stride of $h_d/2 \times w_d/2$ to generate overlapping predictions with different contexts, which are averaged to increase robustness. The crops of the sliding window can be processed in parallel as the images in a batch, which allows for efficient computation on the GPU.

For model validation or deployment, the full-scale HRDA semantic segmentation $\hat{y}_{F,HR}$ of the entire image x_{HR} is necessary. As the context crop is usually smaller than the entire image, $\hat{y}_{F,HR}$ is generated using an overlapping sliding window over the entire image x_{HR} with a size of $sh_c \times sw_c$ and a stride of $sh_c/2 \times sw_c/2$. Within the sliding window, the HRDA prediction is generated in the same way as $\hat{y}_{c,F}^T$ for the pseudo-label.

5 Experiments

5.1 Implementation Details

Datasets: As target data, the real-world Cityscapes dataset [11] of European street scenes with 2975 training and 500 validation images of 2048×1024 pixels is used. As source data, the synthetic datasets GTA [47] with 24,966 images of 1914×1052 pixels and Synthia [49] with 9,400 images of 1280×760 pixels are used. Previous works [28, 58, 59, 73] downsample Cityscapes to 1024×512 and GTA to 1280×720 . Instead, we maintain the full resolution for Cityscapes. To train with the same scale ratio of source and target images as previous works, we resize GTA to 2560×1440 and Synthia to 2560×1520 pixels.

Network Architecture: Our default network is based on DAFormer [28]. It consists of an MiT-B5 encoder [70] and a context-aware feature fusion decoder [28]. For the scale attention decoder, we use the lightweight SegFormer MLP decoder [70] with an embedding dimension of 256. When evaluating other

Table 1. Comparison with previous UDA methods. The results of HRDA are averaged over 3 random seeds. Further methods are shown in the supplement.

	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
GTA5 → Cityscapes																				
CBST [92]	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
DACS [58]	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
CorDA [66]	94.7	63.1	87.6	30.7	40.6	40.2	47.8	51.6	87.6	47.0	89.7	66.7	35.9	90.2	48.9	57.5	0.0	39.8	56.0	56.6
BAPA [39]	94.4	61.0	88.0	26.8	39.9	38.3	46.1	55.3	87.8	46.1	89.4	68.8	40.0	90.2	60.4	59.0	0.0	45.1	54.2	57.4
ProDA [80]	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 [28]	<u>95.7</u>	<u>70.2</u>	<u>89.4</u>	<u>53.5</u>	<u>48.1</u>	<u>49.6</u>	<u>55.8</u>	<u>59.4</u>	<u>89.9</u>	<u>47.9</u>	<u>92.5</u>	<u>72.2</u>	<u>44.7</u>	<u>92.3</u>	<u>74.5</u>	<u>78.2</u>	<u>65.1</u>	<u>55.9</u>	<u>61.8</u>	<u>68.3</u>
HRDA (Ours)	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
Synthia → Cityscapes																				
CBST [92]	68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	-	78.3	60.6	28.3	81.6	-	23.5	-	18.8	39.8	42.6
DACS [58]	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	-	<u>90.8</u>	67.6	38.3	82.9	-	38.9	-	28.5	47.6	48.3
BAPA [39]	<u>91.7</u>	<u>53.8</u>	83.9	22.4	0.8	34.9	30.5	42.8	<u>86.6</u>	-	88.2	66.0	34.1	86.6	-	51.3	-	29.4	50.5	53.3
CorDA [66]	93.3	61.6	85.3	19.6	<u>5.1</u>	37.8	36.6	42.8	84.9	-	90.4	69.7	41.8	85.6	-	38.4	-	32.6	53.9	55.0
ProDA [80]	87.8	45.7	84.6	37.1	0.6	44.0	54.6	37.0	88.1	-	84.4	<u>74.2</u>	24.3	<u>88.2</u>	-	51.1	-	40.5	45.6	55.5
DAFormer [28]	84.5	40.7	<u>88.4</u>	<u>41.5</u>	6.5	<u>50.0</u>	<u>55.0</u>	54.6	86.0	-	89.8	73.2	<u>48.2</u>	87.2	-	<u>53.2</u>	-	<u>53.9</u>	<u>61.7</u>	<u>60.9</u>
HRDA (Ours)	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

UDA methods in Tab. 2, we use a ResNet101 [24] backbone with a DeepLabV2 [3] decoder both as segmentation and scale attention head.

Training: By default, we follow the DAFormer [28] self-training strategy (see Sec. 3) and training parameters, i.e. AdamW [42] with a learning rate of 6×10^{-5} for the encoder and 6×10^{-4} for the decoder, a batch size of 2, linear learning rate warmup, $\lambda_{st}=1$, $\alpha=0.999$, and DACS [58] data augmentation. For adversarial training and entropy minimization, we use SGD with a learning rate of 0.0025 and $\lambda_{adv}=\lambda_{ent}=0.001$. The context and detail crop are generated using $h_c=w_c=h_d=w_d=512$ with $s=2$ to balance the required resources for both crops in the default case. The detail loss weight is chosen empirically $\lambda_d=0.1$. The experiments are conducted on a Titan RTX GPU with 24 GB memory.

5.2 Comparison with State-of-the-Art UDA Methods

First, we compare the proposed HRDA with previous UDA methods in Tab. 1. It can be seen that HRDA outperforms the previously best state-of-the-art method by a significant margin of +5.5 mIoU on GTA→Cityscapes and +4.9 mIoU on Synthia→Cityscapes. HRDA improves the IoU of almost all classes across both datasets. The highest performance gains are achieved for classes with fine segmentation details such as pole, traffic light, traffic sign, person, rider, motorbike, and bike. But also large classes such as truck, bus, and train benefit from HRDA. This is also reflected in the visual examples in Fig. 3.

5.3 HRDA for Different UDA Methods

HRDA is designed to be applicable to most UDA methods. In Tab. 2, we compare the performance without and with HRDA of three further representative UDA methods. It can be seen that HRDA consistently improves the performance by at least +2.4 mIoU, demonstrating the importance of high- and multi-resolution inputs for UDA in general. Also, it shows that HRDA can be applied to different network architectures. The highest improvement is achieved for self-training

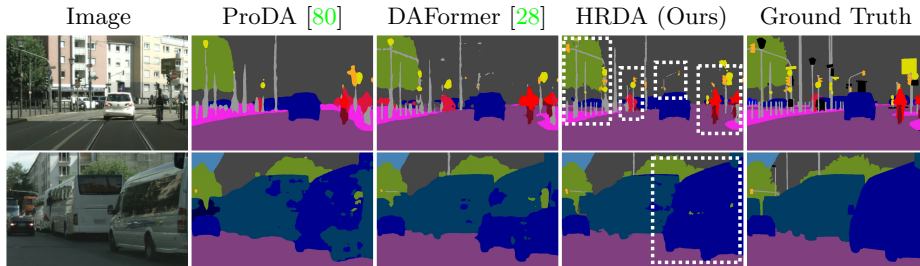


Fig. 3. Qualitative comparison of HRDA with previous methods on GTA→Cityscapes. HRDA improves the segmentation of small classes such as pole, traffic sign, traffic light, and rider as well as large and difficult classes such as bus.

Table 2. HRDA applied to different UDA methods on GTA→Cityscapes. Mean and standard deviation are provided over 3 random seeds.

	UDA Method	Network	w/o HRDA	w/ HRDA	Improvement
1	Entropy Min. [62]	DeepLabV2 [3]	44.3 ± 0.4	46.7 ± 1.2	+2.4
2	Adversarial [59]	DeepLabV2 [3]	44.2 ± 0.1	47.1 ± 1.0	+2.9
3	DACS [58]	DeepLabV2 [3]	53.9 ± 0.6	59.4 ± 1.2	+5.5
4	DAFormer [28]	DeepLabV2 [3]	56.0 ± 0.5	63.0 ± 0.4	+7.0
5	DAFormer [28]	DAFormer [28]	68.3 ± 0.5	73.8 ± 0.3	+5.5

methods (row 3-5) with +5.5 mIoU and more, which shows that the HRDA pseudo-labels positively reinforce the UDA process.

5.4 Influence of Resolution and Crop Size on UDA

In the following, we analyze the underlying principles of HRDA on GTA→Cityscapes, starting with the influence of the resolution and crop size on UDA. For the comparison, we use the relative crop size $h/\frac{H_T}{s}$, which is normalized by the image height at the corresponding resolution, to disentangle the crop size from the used image resolution. Fig. 4 shows that both an increased resolution and crop size improve the performance for both UDA and supervised learning. A large crop size is even more important for UDA than for supervised learning, i.e. a 4 times smaller LR crop reduces the performance by 39% for UDA and by 14% for supervised training. The larger crop provides more context clues and improves the performance of all classes, especially the ones that are difficult to adapt such as wall, fence, truck, bus, and train (cf. row 1 and 3 in Fig. 5), probably, as the relevant context clues are more domain-invariant [31, 71, 89]. A higher input resolution improves the UDA performance by a similar amount as it improves supervised learning. The improvement originates from a higher IoU for small classes such as pole, traffic light, traffic sign, person, motorbike, and bicycle, while some large classes such as road, sidewalk, and terrain have a decreased performance (cf. row 1 and 2 in Fig. 5). This supports that large objects are easier to adapt at LR while small objects are easier to adapt at HR, which can be exploited by the multi-resolution fusion of HRDA.

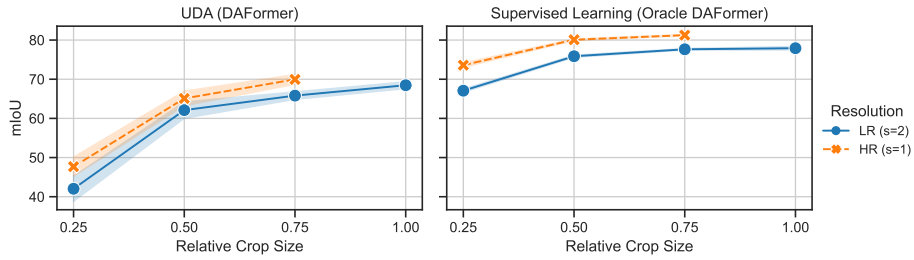


Fig. 4. Segmentation performance over the relative crop size ($h/\frac{H_T}{s}$) for different resolutions and for both the UDA method DAFormer [28] and the target-supervised oracle. There is no value for HR_{1.0} due to GPU memory constraints.

LR _{0.5}	96	71	88	46	27	44	50	56	89	45	91	70	40	91	60	66	39	50	61	62
HR _{0.5}	93	63	89	47	27	52	56	62	90	40	91	76	41	92	61	66	67	56	67	65
LR _{1.0}	96	71	89	54	49	49	56	60	90	50	92	72	46	92	69	80	67	57	62	68
LR _{1.0} +HR _{0.5}	96	74	91	62	51	57	64	69	91	48	94	79	53	94	84	86	76	64	68	74
	Road	S.walk	Build.	Wall	Fence	Pole	T.Light	T.Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU

Fig. 5. Class-wise IoU for the UDA method DAFormer [28] for different crop resolutions XR ($s_{LR}=2, s_{HR}=1$) and relative crop sizes $a=h/\frac{H_T}{s_{XR}}$. Crops are denoted as XR_a . The colors indicate the difference to the first row.

5.5 Combining Crops from Multiple Resolutions with HRDA

Multi-Resolution UDA: Next, we combine crops from LR and HR using the proposed multi-resolution training for UDA. Tab. 3 shows that training with multiple resolutions improves the performance over both LR-only and HR-only training by +3.4 mIoU (cf. row 2 and 3), which demonstrates that multi-resolution fusion with scale attention results in better domain adaptation.

Context Crop Size: Based on the observation that large crops are important for UDA (Sec. 5.4), we increase the context crop size while keeping the detail crop size fixed, which further improves the performance by +5.3 mIoU (cf. row 3 and 5), demonstrating the effectiveness of the proposed small HR detail and large LR context crops. Fig. 5 shows that the multi-resolution training combines the strength of the single-scale training with HR_{0.5} and LR_{1.0} as the multi-resolution IoU of each class is better than the best single-scale IoU (cf. row 2, 3, and 4).

Detail Crop Size: Already the combination of the context crop LR_{1.0} with an even smaller detail crop HR_{0.25} outperforms training with solely the context crop by +2.1 mIoU (cf. row 1 and 3 in Tab. 4), even though HR_{0.25} alone performs −20.8 mIoU worse (cf. row 1 and 2). This shows that the multi-resolution fusion effectively exploits the strength of the small detail crop while compensating for its lacking long-range dependencies with the context crop. Further increasing the detail crop size, results in additional performance gains (cf. row 2 and 5). This shows that even though context information is not crucial for the detail

Table 3. HRDA context size. XR_a denotes crops with resolution XR ($s_{LR}=2$, $s_{HR}=1$) and relative crop size $a=h/\frac{H_T}{s_{XR}}$.

	Context Crop	Detail Crop	mIoU
1	LR _{0.5}	–	62.1 ± 2.1
2	–	HR _{0.5}	65.1 ± 1.9
3	LR _{0.5}	HR _{0.5}	68.5 ± 0.6
4	LR _{0.75}	HR _{0.5}	71.1 ± 1.7
5	LR _{1.0}	HR _{0.5}	73.8 ± 0.3

Table 5. Comparison of HRDA with naive HR crops that have a comparable GPU memory footprint (HR_{0.75}).

	Context	Detail	Mem.	mIoU
1	–	HR _{0.75}	22.0 GB	70.0 ± 1.2
2	LR _{0.75}	HR _{0.375}	13.5 GB	71.3 ± 0.3
3	LR _{1.0}	HR _{0.5}	22.5 GB	73.8 ± 0.3

Table 4. HRDA detail size. XR_a denotes crops with resolution XR ($s_{LR}=2$, $s_{HR}=1$) and relative crop size $a=h/\frac{H_T}{s_{XR}}$.

	Context Crop	Detail Crop	mIoU
1	LR _{1.0}	–	68.5 ± 0.9
2	–	HR _{0.25}	47.7 ± 2.4
3	LR _{1.0}	HR _{0.25}	70.6 ± 0.7
4	LR _{1.0}	HR _{0.375}	71.7 ± 0.4
5	LR _{1.0}	HR _{0.5}	73.8 ± 0.3

Table 6. HRDA detail crop variants. Up-LR: LR crop upsampled to HR resolution.

	Context Crop	Detail Crop	mIoU
1	LR _{1.0}	–	68.5 ± 0.9
2	LR _{1.0}	LR _{0.5}	69.1 ± 0.4
3	LR _{1.0}	Up-LR _{0.5}	71.9 ± 1.5
4	LR _{1.0}	HR _{0.5}	73.8 ± 0.3

crop, it is still helpful to some extent while being limited due to GPU memory constraints.

Detail Crop Variants: It is crucial for HRDA to use context/detail crops with different resolutions. Using LR instead of HR for the detail crop gives only a marginal gain of +0.6 over the baseline (cf. row 1 and 2 in Tab. 6). However, using an LR crop that is bilinearly upsampled to HR as detail crop does improve the performance by +3.4 mIoU (cf. row 1 and 3 in Tab. 6) but is still -1.9 mIoU worse than using a real HR detail crop (cf. row 3 and 4 in Tab. 6). This shows that the improved performance of HRDA comes from both the additional zoomed-in context information as well as the additional details in the HR image.

Comparison with Naive HR: We compare HRDA with naive large HR crops (HR_{0.75}), which have a comparable GPU memory footprint as HRDA. This is a very strong baseline, which is already +1.7 mIoU better than DAFormer [28]. Tab. 5 shows that HRDA still outperforms HR_{0.75} crops by +3.8 mIoU (cf. row 1 and 3). Even when reducing the crop size of HRDA to match the size of HR_{0.75}, HRDA is still +1.3 mIoU better while requiring 40% less GPU memory. This demonstrates that combining LR context crop and HR detail crop performs better than naively increasing the resolution, due to HRDA’s capability of capturing large context information and multi-resolution fusion.

HRDA Component Ablations: The components of HRDA are ablated in Tab. 7. The most crucial component is the learned scale attention. While naively averaging the predictions from both scales gives no improvement over just using the context crop (cf. row 2 and 3), the learned scale attention improves the performance by +3.0 mIoU (cf. row 2 and 4). This shows that it is crucial to learn which scale is best-suited to adapt certain image regions. Generating pseudo-labels with different context views by overlapping slide detail crops results in a

Table 7. Component ablation of HRDA.

	Context	Detail	Scale Attention	Overlapping Detail	Detail Loss	mIoU
1	–	✓	–	–	–	65.1 \pm 1.9
2	✓	–	–	–	–	68.5 \pm 0.9
3	✓	✓	Average	–	–	67.5 \pm 0.8
4	✓	✓	Learned	–	–	71.5 \pm 0.5
5	✓	✓	Learned	✓	–	72.4 \pm 0.1
6	✓	✓	Learned	✓	✓	73.8 \pm 0.3

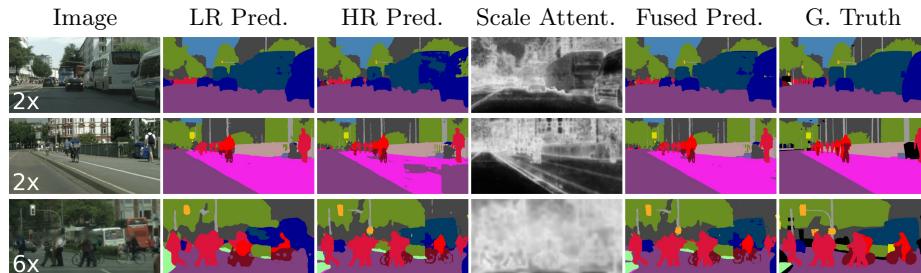


Fig. 6. Visual examples of the different predictions and the scale attention of HRDA. Large objects are better segmented from LR while small objects are better segmented from HR. The scale attention learns to utilize this pattern for fusing LR and HR predictions. The examples are zoomed in (2x or 6x) for better visibility of the details.

further gain of +0.9 mIoU (cf. row 4 and 5). Finally, additional supervision of the detail crop ($\lambda_d = 0.1$) further provides +1.4 mIoU (cf. row 5 and 6).

Qualitative Analysis: Fig. 6 provides representative visual examples demonstrating that LR predictions work better for large objects such as a bus or sidewalk (row 1/2) while HR predictions work better for small objects and fine details (row 3). The scale attention focuses on LR for large objects and on HR for small objects and segmentation borders, combining the strength of both.

Supplement: The supplement provides further parameter studies, additional baseline comparisons, a runtime analysis, and an extended qualitative analysis.

6 Conclusions

In this work, we presented HRDA, a multi-resolution approach for UDA that combines the advantages of small HR detail crops and large LR context crops using a learned scale attention, while maintaining a manageable GPU memory footprint. It can be combined with various UDA methods and achieves a consistent, significant improvement. Overall, HRDA achieves an unprecedented performance of 73.8 mIoU on GTA→Cityscapes and 65.8 mIoU on Synthia→Cityscapes, which is a respective gain of +5.5 mIoU and +4.9 mIoU over the previous SOTA.

Acknowledgements: This work is supported by the European Lighthouse on Secure and Safe AI (ELSA) and a Facebook Academic Gift on Robust Perception (INFO224).

References

1. Araslanov, N., Roth, S.: Self-supervised augmentation consistency for adapting semantic segmentation. In: CVPR. pp. 15384–15394 (2021) [2](#), [4](#), [5](#)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In: ICLR. pp. 834–848 (2015) [4](#)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. PAMI **40**(4), 834–848 (2017) [4](#), [10](#), [11](#)
4. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017) [4](#)
5. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. In: CVPR. pp. 3640–3649 (2016) [3](#), [4](#), [8](#)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 801–818 (2018) [4](#)
7. Chen, M., Xue, H., Cai, D.: Domain adaptation for semantic segmentation with maximum squares loss. In: ICCV. pp. 2090–2099 (2019) [4](#)
8. Chen, Y., Li, W., Chen, X., Gool, L.V.: Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In: CVPR. pp. 1841–1850 (2019) [4](#)
9. Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: CVPR. pp. 12475–12485 (2020) [4](#)
10. Choi, J., Kim, T., Kim, C.: Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In: ICCV. pp. 6830–6840 (2019) [4](#)
11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016), <https://www.cityscapes-dataset.com/>, dataset license: <https://www.cityscapes-dataset.com/license/> [1](#), [9](#)
12. Dai, D., Sakaridis, C., Hecker, S., Van Gool, L.: Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. IJCV **128**(5), 1182–1204 (2020) [4](#)
13. Dai, D., Van Gool, L.: Dark model adaptation: Semantic image segmentation from daytime to nighttime. In: ITSC. pp. 3819–3824 (2018) [4](#)
14. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: ICCV. pp. 1635–1643 (2015) [1](#)
15. Ding, H., Jiang, X., Liu, A.Q., Thalmann, N.M., Wang, G.: Boundary-aware feature propagation for scene segmentation. In: ICCV. pp. 6819–6829 (2019) [4](#)
16. Ding, X., Guo, Y., Ding, G., Han, J.: Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In: ICCV. pp. 1911–1920 (2019) [4](#)
17. French, G., Laine, S., Aila, T., Mackiewicz, M., Finlayson, G.: Semi-supervised semantic segmentation needs strong, varied perturbations. In: BMVC (2020) [1](#), [5](#)
18. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: CVPR. pp. 3146–3154 (2019) [4](#)
19. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR **17**(1), 2096–2030 (2016) [4](#)

20. Gao, L., Zhang, J., Zhang, L., Tao, D.: DSP: Dual Soft-Paste for Unsupervised Domain Adaptive Semantic Segmentation. In: ACMMM. pp. 2825–2833 (2021), <http://arxiv.org/abs/2107.09600> 4
21. Gong, R., Li, W., Chen, Y., Dai, D., Van Gool, L.: Dlow: Domain flow and applications. IJCV **129**(10), 2865–2888 (2021) 4
22. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS. pp. 2672–2680 (2014) 4
23. Guan, D., Huang, J., Lu, S., Xiao, A.: Scale variance minimization for unsupervised domain adaptation in image segmentation. PR **112**, 107764 (2021) 4
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 10
25. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: ICML. pp. 1989–1998 (2018) 4
26. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016) 1, 4
27. Hoyer, L., Dai, D., Chen, Y., Köring, A., Saha, S., Van Gool, L.: Three ways to improve semantic segmentation with self-supervised depth estimation. In: CVPR. pp. 11130–11140 (2021) 1
28. Hoyer, L., Dai, D., Van Gool, L.: DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: CVPR (2022) 1, 2, 3, 5, 7, 9, 10, 11, 12, 13
29. Hoyer, L., Dai, D., Wang, Q., Chen, Y., Van Gool, L.: Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. arXiv preprint arXiv:2108.12545 (2021) 4
30. Hoyer, L., Munoz, M., Katiyar, P., Khoreva, A., Fischer, V.: Grid saliency for context explanations of semantic segmentation. In: NeurIPS. pp. 6462–6473 (2019) 4
31. Huang, J., Lu, S., Guan, D., Zhang, X.: Contextual-relation consistent domain adaptation for semantic segmentation. In: ECCV. pp. 705–722 (2020) 3, 6, 11
32. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: ICCV. pp. 603–612 (2019) 4
33. Iqbal, J., Ali, M.: Msl: Multi-level self-supervised learning for domain adaptation with spatially independent and semantically consistent labeling. In: WACV. pp. 1864–1873 (2020) 4
34. Kim, M., Byun, H.: Learning texture invariant representation for domain adaptation of semantic segmentation. In: CVPR. pp. 12975–12984 (2020) 4
35. Lai, X., Tian, Z., Jiang, L., Liu, S., Zhao, H., Wang, L., Jia, J.: Semi-Supervised Semantic Segmentation With Directional Context-Aware Consistency. In: CVPR. pp. 1205–1214 (2021) 1, 4
36. Li, X., Li, X., Zhang, L., Cheng, G., Shi, J., Lin, Z., Tan, S., Tong, Y.: Improving semantic segmentation via decoupled body and edge supervision. In: ECCV. pp. 435–452. Springer (2020) 4
37. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional learning for domain adaptation of semantic segmentation. In: CVPR. pp. 6936–6945 (2019) 2, 4
38. Lin, D., Shen, D., Shen, S., Ji, Y., Lischinski, D., Cohen-Or, D., Huang, H.: Zigzag-net: Fusing top-down and bottom-up context for object segmentation. In: CVPR. pp. 7490–7499 (2019) 4

39. Liu, Y., Deng, J., Gao, X., Li, W., Duan, L.: Bapa-net: Boundary adaptation and prototype alignment for cross-domain semantic segmentation. In: ICCV. pp. 8801–8811 (2021) [2](#), [4](#), [5](#), [7](#), [10](#)
40. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–1110022 (2021) [4](#)
41. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015) [4](#)
42. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2018) [10](#)
43. Mei, K., Zhu, C., Zou, J., Zhang, S.: Instance adaptive self-training for unsupervised domain adaptation. In: ECCV. pp. 415–430 (2020) [2](#), [4](#), [5](#)
44. Melas-Kyriazi, L., Manrai, A.K.: Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In: CVPR. pp. 12435–12445 (2021) [4](#)
45. Pizzati, F., Charette, R.d., Zaccaria, M., Cerri, P.: Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation. In: WACV. pp. 2990–2998 (2020) [4](#)
46. Prabhu, V., Khare, S., Kartik, D., Hoffman, J.: Augco: Augmentation consistency-guided self-training for source-free domain adaptive semantic segmentation. arXiv preprint arXiv:2107.10140 (2021) [4](#)
47. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: ECCV. pp. 102–118 (2016), https://download.visinf.tu-darmstadt.de/data/from_games/ [2](#), [9](#)
48. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Int. Conf. Medical Image Computing and Computer-assisted Intervention. pp. 234–241 (2015) [4](#)
49. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR. pp. 3234–3243 (2016), <http://synthia-dataset.net/>, dataset license: CC BY-NC-SA 3.0 [2](#), [9](#)
50. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: NeurIPS (2016) [4](#), [5](#)
51. Sakaridis, C., Dai, D., Van Gool, L.: ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: ICCV. pp. 10765–10775 (2021) [1](#)
52. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: NeurIPS (2020) [4](#), [5](#)
53. Souly, N., Spampinato, C., Shah, M.: Semi supervised semantic segmentation using generative adversarial network. In: ICCV. pp. 5688–5696 (2017) [1](#)
54. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: ICCV. pp. 7262–7272 (2021) [4](#)
55. Subhani, M.N., Ali, M.: Learning from scale-invariant examples for domain adaptation in semantic segmentation. In: ECCV. pp. 290–306. Springer (2020) [4](#)
56. Tao, A., Sapra, K., Catanzaro, B.: Hierarchical multi-scale attention for semantic segmentation. arXiv preprint arXiv:2005.10821 (2020) [3](#), [4](#)
57. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS. pp. 1195–1204 (2017) [4](#), [5](#)

58. Tranheden, W., Olsson, V., Pinto, J., Svensson, L.: DACS: Domain Adaptation via Cross-domain Mixed Sampling. In: WACV. pp. 1379–1389 (2021) [2](#), [3](#), [4](#), [5](#), [7](#), [9](#), [10](#), [11](#)
59. Tsai, Y.H., Hung, W.C., Schuler, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR. pp. 7472–7481 (2018) [1](#), [2](#), [3](#), [4](#), [5](#), [9](#), [11](#)
60. Tsai, Y.H., Sohn, K., Schuler, S., Chandraker, M.: Domain adaptation for structured output via discriminative patch representations. In: ICCV. pp. 1456–1465 (2019) [4](#), [5](#)
61. Unal, O., Dai, D., Van Gool, L.: Scribble-supervised lidar semantic segmentation. In: CVPR. pp. 2697–2707 (2022) [1](#)
62. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR. pp. 2517–2526 (2019) [3](#), [4](#), [11](#)
63. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Dada: Depth-aware domain adaptation in semantic segmentation. In: ICCV. pp. 7364–7373 (2019) [4](#)
64. Wang, H., Shen, T., Zhang, W., Duan, L.Y., Mei, T.: Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In: ECCV. pp. 642–659 (2020) [4](#), [5](#)
65. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. PAMI **43**(10), 3349–3364 (2020) [4](#)
66. Wang, Q., Dai, D., Hoyer, L., Fink, O., Van Gool, L.: Domain adaptive semantic segmentation with self-supervised depth estimation. In: ICCV. pp. 8515–8525 (2021) [2](#), [4](#), [10](#)
67. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: CVPR. pp. 7201–7211 (2022) [1](#)
68. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: ICCV. pp. 568–578 (2021) [4](#)
69. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR. pp. 7794–7803 (2018) [4](#)
70. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In: NeurIPS (2021) [4](#), [9](#)
71. Yang, J., An, W., Yan, C., Zhao, P., Huang, J.: Context-aware domain adaptation in semantic segmentation. In: WACV. pp. 514–524 (2021) [3](#), [6](#), [11](#)
72. Yang, S., Peng, G.: Attention to refine through multi scales for semantic segmentation. In: Pacific Rim Conference on Multimedia. pp. 232–241 (2018) [3](#), [4](#)
73. Yang, Y., Soatto, S.: Fda: Fourier domain adaptation for semantic segmentation. In: CVPR. pp. 4085–4095 (2020) [2](#), [9](#)
74. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015) [4](#)
75. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: ECCV. pp. 173–190 (2020) [4](#)
76. Yuan, Y., Huang, L., Guo, J., Zhang, C., Chen, X., Wang, J.: Ocnets: Object context for semantic segmentation. IJCV pp. 1–24 (2021) [4](#)
77. Yuan, Y., Xie, J., Chen, X., Wang, J.: Segfix: Model-agnostic boundary refinement for segmentation. In: ECCV. pp. 489–506. Springer (2020) [4](#)
78. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: CVPR. pp. 7151–7160 (2018) [4](#)

79. Zhang, K., Sun, Y., Wang, R., Li, H., Hu, X.: Multiple fusion adaptation: A strong framework for unsupervised semantic segmentation adaptation. In: BMVC (2021) [4](#)
80. Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In: CVPR. pp. 12414–12424 (2021) [2](#), [4](#), [5](#), [10](#), [11](#)
81. Zhang, Q., Zhang, J., Liu, W., Tao, D.: Category anchor-guided unsupervised domain adaptation for semantic segmentation. In: NeurIPS. pp. 435–445 (2019) [2](#), [4](#), [5](#)
82. Zhang, Y., David, P., Foroosh, H., Gong, B.: A curriculum domain adaptation approach to the semantic segmentation of urban scenes. PAMI **42**(8), 1823–1841 (2019) [4](#)
83. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 2881–2890 (2017) [4](#)
84. Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C.C., Lin, D., Jia, J.: Psanet: Point-wise spatial attention network for scene parsing. In: ECCV. pp. 267–283 (2018) [4](#)
85. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR. pp. 6881–6890 (2021) [4](#), [9](#)
86. Zheng, Z., Yang, Y.: Unsupervised scene adaptation with memory regularization in vivo. In: IJCAI. pp. 1076–1082 (2020) [4](#)
87. Zheng, Z., Yang, Y.: Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. IJCV **129**(4), 1106–1120 (2021) [4](#)
88. Zhou, Q., Feng, Z., Gu, Q., Cheng, G., Lu, X., Shi, J., Ma, L.: Uncertainty-aware consistency regularization for cross-domain semantic segmentation. arXiv preprint arXiv:2004.08878 (2020) [4](#)
89. Zhou, Q., Feng, Z., Gu, Q., Pang, J., Cheng, G., Lu, X., Shi, J., Ma, L.: Context-aware mixup for domain adaptive semantic segmentation. In: WACV. pp. 514–524 (2021) [3](#), [4](#), [6](#), [11](#)
90. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR. pp. 1851–1858 (2017) [2](#)
91. Zhou, Y., Sun, X., Zha, Z.J., Zeng, W.: Context-reinforced semantic segmentation. In: CVPR. pp. 4046–4055 (2019) [4](#)
92. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV. pp. 289–305 (2018) [1](#), [2](#), [4](#), [5](#), [10](#)
93. Zou, Y., Zhang, Z., Zhang, H., Li, C.L., Bian, X., Huang, J.B., Pfister, T.: Pseudoseg: Designing pseudo labels for semantic segmentation. In: ICLR (2021) [1](#)