Semantically Consistent Image-to-Image Translation for Unsupervised Domain Adaptation

Stephan Brehm, Sebastian Scherer and Rainer Lienhart
Department of Computer Science, University of Augsburg, Universitätsstr. 6a, Augsburg, Germany

Keywords: Image Translation, Semi-supervised Learning, Unsupervised Learning, Domain Adaptation, Semantic Segmentation, Synthetic Data, Semantic Consistency, Generative Adversarial Networks.

Abstract: Unsupervised Domain Adaptation (UDA) aims to adapt models trained on a source domain to a new target domain where no labelled data is available. In this work, we investigate the problem of UDA from a synthetic computer-generated domain to a similar but real-world domain for learning semantic segmentation. We propose a semantically consistent image-to-image translation method in combination with a consistency regularisation method for UDA. We overcome previous limitations on transferring synthetic images to real looking images. We leverage pseudo-labels in order to learn a generative image-to-image translation model that receives additional feedback from semantic labels on both domains. Our method outperforms state-of-the-art methods that combine image-to-image translation and semi-supervised learning on relevant domain adaptation benchmarks, i.e., on GTA5 to Cityscapes and SYNTHIA to Cityscapes.

1 INTRODUCTION

The problem of domain adaptation from a synthetic source domain to the real target domain is mainly motivated by the cheap and almost endless possibilities of automated creation of synthetic data. In contrast, data from the real domain is often hard to acquire. This is especially true for most types of labelled data. A common problem in computer vision, for which acquiring real labelled data is exceptionally hard, is semantic segmentation. Semantic segmentation requires annotations at the pixel level. These annotations commonly need to be created manually in a time-consuming process that also demands rigorous and continuous focus from human annotators. Due to this, a reasonable approach is to learn as much as possible from synthetic data and then transfer the knowledge to the real domain. However, convolutional neural networks (CNNs), in general, learn features from the domain on which they are trained on. This causes networks to perform poorly on unseen domains due to the visual gap between these domains. Unsupervised domain adaptation (UDA) aims to bridge this domain gap in order to learn models that perform well in the target domain without ever using labels from that domain.

In this work, we aim to improve the quality and usefulness of synthetic data by transforming synthetic images such that they look more like real images. However, such an Image-to-Image Translation (I2I) approach cannot change fundamental differences in image content because it needs to keep the transformed images consistent to the semantic labels of synthetic images. Because of this, we cannot bridge the gap between synthetic and real domains solely by an I2I approach.

The remaining differences are in image content and include object shapes, object frequency, and differences in viewpoint. Following recent work, we utilise a Semi-Supervised Learning (SSL) framework which allows us to include unlabelled data from the real target domain into the training of a semantic segmentation network. The main contributions of our work are summarised as follows:

1. We propose a semantically consistent I2I method that combines an adversarial approach with a segmentation objective that is optimised jointly by both generator and discriminator.
2. We improve both adversarial and segmentation objectives with self-supervised techniques that include the unlabelled data from the real target domain into the training. Our method outperforms state-of-the-art combinations of I2I and SSL on challenging benchmarks for UDA. We provide extensive experiments and analyses and show which components are essential to our approach.

Brehm, S., Scherer, S. and Lienhart, R.
Semantically Consistent Image-to-Image Translation for Unsupervised Domain Adaptation.
DOI: 10.5220/0010786000003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 2*, pages 131-141
ISBN: 978-989-758-547-0; ISSN: 2184-433X

2 RELATED WORK

Semantic Segmentation: is the task of labelling every pixel of an image according to the object class it belongs to. In 2015, Long *et al*. proposed fully convolution neural networks (FCNs) (Long et al., 2015) improving massively upon classical methods for semantic segmentation. Since then, many new methods based on deep convolutional neural networks were proposed (Chen et al., 2018; Wang et al., 2021; Yu and Koltun, 2016). However, these methods are directly learned on the real target domain in a supervised fashion. In contrast, our method is able to learn about the real domain without the necessity of annotated real data.

Image-to-Image Translation (I2I): methods (Choi et al., 2019; Pizzati et al., 2020) have been widely used to bridge the gap between synthetic and real data. For image data, this is commonly achieved by learning a deep convolutional neural network that receives a synthetic image from the source domain as input and manipulates it in such a way that it looks more realistic. Basically, these methods try to make the synthetic data look more real. For our task, it is important that the manipulated versions of the synthetic data can be used for supervised training of a segmentation method subsequently. In order to learn useful manipulations for such a task, these systems need mechanisms that keep the overall image consistent with the synthetic labels.

Unsupervised Domain Adaptation (UDA): for semantic segmentation has been extensively studied in the last years. Adversarial training is often used in UDA methods to adapt either input space, feature space or output space of a semantic segmentation network (Toldo et al., 2020). On the feature level, a discriminator is trained to distinguish between feature maps from different domains (Hoffman et al., 2016; Chen et al., 2017; Hong et al., 2018). Popular input space adaptation techniques utilise frameworks based on cycle consistency (Zhu et al., 2017) for UDA (Hoffman et al., 2018; Sankaranarayanan et al., 2018; Chen et al., 2019; Murez et al., 2018). However, cycle consistency allows almost arbitrary transformations as long as they can be reversed. In contrast to these methods, we perform input space adaptation in our I2I method without the need for cycle consistency. Our method simply uses the annotations of the synthetic data to directly enforce consistency.

Semi-Supervised Learning (SSL): aims to include unlabelled data into the training which allows to use data from the real domain. The dominant approaches for SSL are pseudo-labelling and consistency regularisation. An extensive overview can be found in this survey (van Engelen and Hoos, 2020). Pseudo-Labeling was first proposed in (Lee, 2013). Xie *et al*. (Xie et al., 2020) recently showed that pseudo-labels can indeed improve overall performance on image classification tasks. They train a network on labelled data and reuse high-confidence predictions on unlabelled data as pseudo-labels. Pseudo-labels are then included in the full dataset on which a new model is subsequently trained from scratch.

Many recent SSL methods (Ke et al., 2019; Tarvainen and Valpola, 2017) include consistency regularisation. They employ unlabelled data to produce consistent predictions under different perturbations. Possible perturbations can be data augmentation, dropout or simple noise on the input data. The trained model should be robust against such perturbations. These approaches leverage the idea that a classifier should output the same distribution for different augmented versions of an unlabelled sample. This is typically achieved by minimising the difference between the prediction of a model on different perturbed versions of the same input. We utilise such an approach to boost our performance even further. Our I2I translation module also leverages pseudo-labels created by such a consistency regularised method.

3 METHOD

In this work, we assume that image data is available from both the synthetic *source* domain as well as the real *target* domain. However, annotations are only available in the synthetic *source* domain. We aim to bridge the gap between the *source* and the *target* data. Our image-sets are sampled from their respective domains. The first step is to bring the two domains closer together visually which means transforming the synthetic data such that it looks more real. By keeping consistency with the labels of the synthetic source data, we are able to generate a dataset of images with corresponding labels that closely resembles the real target data in terms of colour, texture and lighting. Basically, we aim to improve domain adaptation by transferring the style of the real target domain onto images of the synthetic source domain. However, fundamental differences in image composition cannot be learned with such an I2I method. For this reason, we use the translated synthetic data in conjunction with unlabelled data from the real *target* domain to train our segmentation model. We propose a training process that consists of

three phases.

- (a) In the *warm-up phase* the SSL method is used to train an initial segmentation model M_0 as the initial pseudo label generator.
- (b) In the *I2I training phase*, we use the pseudo labels from model M_0 and train our generator G to produce real looking images from synthetic images.
- (c) In the *segmentation training phase*, we combine the SSL method with the translated images by generator G and train our new segmentation network M_1.

3.1 Notation

In the following, we will denote $\mathcal{S}$ as the *source* domain and $\mathcal{T}$ as the *target* domain. $X_\mathcal{S}$ and $X_\mathcal{T}$ are sets of images sampled from $\mathcal{S}$ and $\mathcal{T}$, i.e., synthetic images and real images, respectively. $\mathbf{x}^s \in X_\mathcal{S}$ and $\mathbf{x}^t \in X_\mathcal{T}$ are images sampled from their respective image sets $X_\mathcal{S}$ and $X_\mathcal{T}$.

Both domains $\mathcal{S}$ and $\mathcal{T}$ share a common set of categories C. We denote the label of an image $\mathbf{x}^d$ of domain $d \in \{\mathcal{S}, \mathcal{T}\}$ as $\mathbf{y}^d$. In this work $\mathbf{y}^s \in Y_\mathcal{S}$ generally is a synthetic segmentation mask and $\hat{\mathbf{y}}^t \in \hat{Y}_\mathcal{T}$ is a segmentation mask that contains pseudo-labels. For the purpose of this work, the real segmentation labels $\mathbf{y}^t \in Y_\mathcal{T}$ are not available for training. $\mathcal{L}^{W,n}_{i,j,c}(\mathbf{x}, \mathbf{y})$ denotes a loss function based on image $\mathbf{x}$ and label $\mathbf{y}$ that is used to update trainable parameters of a network W. n is an identifier/name for the loss function. Note that we often omit the arguments $(\mathbf{x}, \mathbf{y})$ in order to simplify our notation. Also note that W only indicates which network is updated by $\mathcal{L}^n$ and thus may also be omitted in more general statements and definitions. i, j, c are indices that we use to refer to specific dimensions and or individual values in the calculated loss if necessary. In general, we assume multidimensional arrays to be in the order of $height \times width \times channels$.

In this work, we aim to train a segmentation network $F_\mathcal{S}$ on $X_\mathcal{S}$ (and possibly $X_\mathcal{T}$) that estimates $Y_\mathcal{T}$ without the need for any $\mathbf{y}^t \in Y_\mathcal{T}$ during training, i.e., we aim to generalise from domain $\mathcal{S}$ to domain $\mathcal{T}$ without ever using any labels from $\mathcal{T}$.

3.2 Image-to-Image Translation

In order to generalise to a target domain $\mathcal{T}$ we need to bring our source images $X_\mathcal{S}$ closer to $\mathcal{T}$. In this section, we detail our approach to this goal. Figure 1 illustrates the employed architecture. We aim to transfer images from a synthetic domain to the real domain, because the synthetic domain allows to easily create images and corresponding labels by simulating an environment that is similar to the real environment. The transformed images are supposed to deliver maximum performance on real data when learning a fully convolutional method for semantic segmentation. In order to achieve this, such a transformation needs to meet two major requirements.

1. Transformed images need to look as realistic as possible
2. Transformed images need to maintain consistency to the segmentation labels

We propose an I2I method that consists of two networks in an adversarial setting. Our learning task is designed to enforce the requirements given above. Requirement 1. is tackled by an adversarial objective. For Requirement 2., we extend the adversarial setting with an additional cooperative segmentation objective that is jointly optimised by both adversaries. Following, we give a brief overview of the employed generator and discriminator architectures.

Generator. We use a simple encoder-decoder architecture with strided convolutions for down-sampling the input image by a factor of 8. These down-sampling operations are followed by a single residual block. The decoder is simply a stack of deconvolution layers. Note that we do not include any skip/residual connections between encoder and decoder. In every layer, we append a learned scaling factor as well as a learned bias term. We use leaky-ReLU activations for all layers except the output layer. We also use deconvolution kernels, PixelNorm, Equalized Learning Rate, Adaptive Instance Normalization and Stochastic Variation as proposed in (Karras et al., 2018). Similar to Taigman *et al.* (Taigman et al., 2017), we incorporate images from the target domain in the training of the generator by adding an identity objective. Thus, the generator learns to encode and reconstruct real images in addition to its main I2I task. The additional feedback is a way to directly learn about the structure and texture of real images.

Discriminator. Instead of discriminating between generated images and real images directly, we propose to use prior knowledge for discrimination. We use image-features of an ImageNet-pretrained (Deng et al., 2009) VGG16 network (Simonyan and Zisserman, 2015) after the *block3_conv3* layer. We build additional 3 residual blocks (He et al., 2016) on top of these features. Note that we do not fine-tune the pre-trained VGG16 part of the discriminator. Our discriminator features two distinct outputs which are

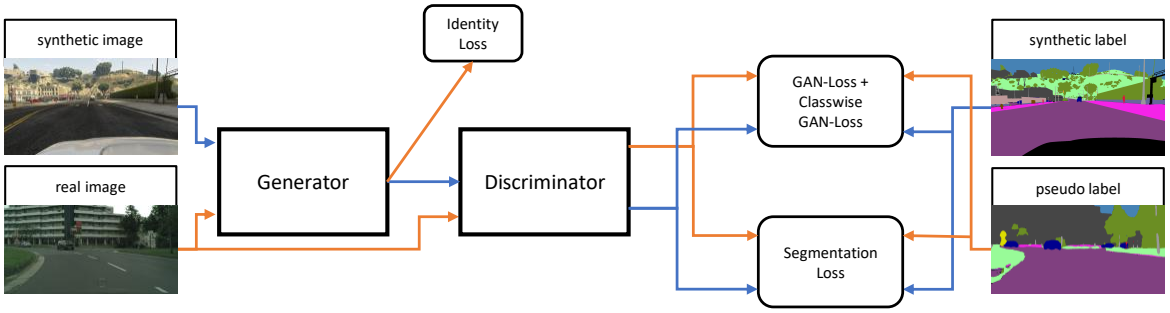

Figure 1: Schematic illustration of our proposed I2I method. The colored arrows visualize data-flow.

each computed on top of a separate decoder, as illustrated in Figure 1. We refer to these decoders as the GAN-head (Goodfellow et al., 2014) and the auxiliary classifier head (AC-head) (Odena et al., 2017). Both heads output results at image resolution. The task of the AC-head is to learn semantic segmentation in the target domain. Thus we feed the translated images in conjunction with the labels of the synthetic data as well as real images with corresponding pseudo-labels. Note that these pseudo-labels can be created online during training. However, using pseudo-labels created by our proposed segmentation system that is detailed in section 3.3 generally improved results. The AC-head is important in many ways. First, the segmentation feedback generated in the AC-head forces previous layers to learn segmentation-specific features which in turn helps the GAN-head to discriminate between images that are translated from the synthetic domain and images from the real domain. Second, it provides consistency feedback for the generator. We achieve this by propagating gradients based on the segmentation loss of the AC-head into the weights of the discriminator as well as into the weights of the generator. This effectively optimises a joint segmentation objective in both generator and discriminator. Thus, the generator is punished if the translated images are not consistent with the label of the synthetic input data. Third, it can be used to generate pseudo-labels. The GAN-Head is responsible for the discrimination between real examples and fake examples. Note that GAN-Heads are commonly learned with a single label only (*real* or *fake*). We extend this formulation to include class information. Thus, allowing the discriminator to directly learn to compare images on a class specific level. We utilize segmentation labels in the GAN-head to create this class-specific feedback. In GAN-terminology this means that we treat pseudo-labels as real and the synthetic labels as fake. For each class, we produce an additional output feature map. On this feature map, feedback is only applied at positions that belong to the given object class as defined by either synthetic labels or pseudo-labels.

Optimisation. Given the supervised segmentation loss $\mathcal{L}^{seg}(D(G(\mathbf{x}^s)),\mathbf{y}^s)$ (softmax cross-entropy) based on synthetic labels and the supervised segmentation loss $\mathcal{L}^{seg}(D(\mathbf{x}^t),\hat{\mathbf{y}}^t)$ based on the pseudo-labels, we compute the total segmentation loss $\mathcal{L}^{seg}$ as given in Equation 1. We use the symmetric cross-entropy loss (Wang et al., 2019) for $\mathcal{L}^{seg}(D(\mathbf{x}^t),\hat{\mathbf{y}}^t)$ because it increases robustness when using noisy pseudo-labels $\hat{\mathbf{y}}^t$.

$$\mathcal{L}^{seg} = \mathcal{L}^{seg}(D(G(\mathbf{x}^s)),\mathbf{y}^s) + \lambda^{pl}\mathcal{L}^{seg}(D(\mathbf{x}^t),\hat{\mathbf{y}}^t) \quad (1)$$

where λ^{pl} is a scaling factor that we use to control the relative impact of the pseudo-labels on the overall error of the model.

Like the segmentation loss, we calculate the GAN-Loss $\mathcal{L}^{gan}$ on a pixel-wise basis, i.e., the GAN-loss is the average of all losses computed over all pixels $I_{i,j}$ of an input image I. Let $m(a,b)$ be a derivable distance metric between a and b. We give a general error function $\mathcal{L}^{D,dgan}$ for a discriminator D as well as a general error function $\mathcal{L}^{G,dgan}$ for a generator G in Equation 2 and Equation 3 respectively.

$$\mathcal{L}^{D,dgan}_{i,j} = m(D(G(\mathbf{x}^s))_{i,j},0) + m(D(\mathbf{x}^t)_{i,j},1) \quad (2)$$

$$\mathcal{L}^{G,dgan}_{i,j} = m(D(G(\mathbf{x}^s))_{i,j},1) \quad (3)$$

We use the mean-squared error for the distance metric m.

Now let, $\mathbf{y}^s$ be the one-hot encoded segmentation label for a given image $\mathbf{x}^s$, i.e., $\mathbf{y}^s_{i,j,c}$ equals one if $I_{i,j}$ belongs to class c. Otherwise $\mathbf{y}^s_{i,j,c}$ is zero. The same applies to the pseudo-label $\hat{\mathbf{y}}^t$. We define a class-wise and pixel-wise error $\mathcal{L}^{D,cgan}_{i,j,c}$ for discriminator D in Equation 4 and a class-wise and pixel-wise error $\mathcal{L}^{G,cgan}_{i,j,c}$ for generator G in Equation 5.

$$\mathcal{L}^{D,cgan}_{i,j,c} = m(D(G(\mathbf{x}^s))_{i,j,c},0)\cdot\mathbf{y}^s_{i,j,c} + m(D(\mathbf{x}^t)_{i,j,c},1)\cdot\hat{\mathbf{y}}^t_{i,j,c} \quad (4)$$

$$\mathcal{L}^{G,cgan}_{i,j,c} = m(D(G(\mathbf{x}^s))_{i,j,c},1)\cdot\mathbf{y}^s_{i,j,c} \quad (5)$$

Hence, the full adversarial loss for D is:

$$\mathcal{L}^{D,gan} = \sum_{i=0}^{H}\sum_{j=0}^{W}\left(\mathcal{L}_{i,j}^{D,dgan} + \frac{\lambda^{cgan}}{|C|}\sum_{c\in C}\mathcal{L}_{i,j,c}^{D,cgan}\right). \quad (6)$$

where λ^{cgan} is a scaling factor that we use to control the relative impact of the class-wise loss on the overall error of the model, H is the height and W is the width of both images and labels. $\mathcal{L}^{G,gan}$ can be computed analogous.

The total loss of the discriminator $\mathcal{L}^{D}$ is a combination of segmentation and GAN losses as given in Equation 7.

$$\mathcal{L}^{D} = \frac{1}{HW}\left(\mathcal{L}^{D,seg} + \mathcal{L}^{D,gan}\right) \quad (7)$$

With $\mathcal{L}^{G,id}$ being the identity reconstruction error on a real image $\mathbf{x}^t$ from the target domain, we compute the generator loss $\mathcal{L}^{G}$ as given in Equation 9.

$$\mathcal{L}^{G,id} = \sum_{i=0}^{H}\sum_{j=0}^{W}\|G(\mathbf{x}^t)_{i,j} - \mathbf{x}^t_{i,j}\|_1 \quad (8)$$

$$\mathcal{L}^{G} = \frac{1}{HW}\left(\mathcal{L}^{G,seg} + \mathcal{L}^{G,gan} + \mathcal{L}^{G,id}\right) \quad (9)$$

Note that the second term of $\mathcal{L}^{seg}$ from Equation 1 is not affected by G and thus $\mathcal{L}^{G,seg}$ is reduced to the first term of $\mathcal{L}^{seg}$ when back-propagating. This means that G should generate images that match the synthetic segmentation labels. On the other hand, $\mathcal{L}^{G,gan}$ incentivises more realistic images. By using $\mathcal{L}^{G,id}$ we are able to show real images from the target domain to G. This combination of various objectives enables us to learn an I2I model that satisfies the requirement of realistic transformation while simultaneously keeping consistency with the synthetic labels. We use the resulting images to subsequently learn a better segmentation model.

3.3 Segmentation Training

At this point, we are able to transform synthetic images into real images while keeping the semantic content of the image unchanged. Thus, we can use the transformed images with the original synthetic annotations. However, some issues remain to be solved. An I2I method cannot bridge certain differences between the domains. The remaining differences are in image content and include object shapes, object frequency, and differences in viewpoint which can not be learned by our I2I module. As a solution, we incorporate images from the real domain directly into the training of the segmentation model via a SSL framework.

Similar to Tarvainen *et al*. (Tarvainen and Valpola, 2017), we make use of two networks: a student network F_S and a teacher network F_T. The architecture of the teacher network is identical to the one of the student network. The weights of the teacher model are an Exponential Moving Average (EMA) of the student's weights. Given an image from the target domain, the teacher's prediction serves as a label for the student, forcing consistency in the prediction of both models under different perturbations.

The overall objective $\mathcal{L}^{F_S}$ is a combination of a supervised segmentation loss $\mathcal{L}^{F_S,seg}$ and the self-supervised consistency loss $\mathcal{L}^{F_S,con}$ as detailed in Equation 10.

$$\mathcal{L}^{F_S} = \mathcal{L}^{F_S,seg}(\mathbf{x}^s,\mathbf{y}^s) + \lambda^{con}\mathcal{L}^{F_S,con}(\mathbf{x}^t), \quad (10)$$

where λ^{con} is a trade-off parameter.

For the supervised training, we incorporate synthetic images $\mathbf{x}^s \in X_S$ and a transformed version obtained from the generator $G(\mathbf{x}^s)$ of our proposed I2I method. We calculate a combined loss $\mathcal{L}^{F_S,seg}$ as given in Equation 11

$$\mathcal{L}^{F_S,seg} = \mathcal{L}^{seg}((F_S(\mathbf{x}^s)),\mathbf{y}^s) + \mathcal{L}^{seg}((F_S(G(\mathbf{x}^s))),\mathbf{y}^s), \quad (11)$$

where $\mathcal{L}^{seg}$ again is the softmax cross-entropy error.

For the semi-supervised training, we incorporate images from the real domain $\mathbf{x}^t \in X_T$. We calculate the consistency loss $\mathcal{L}^{F_S,con}(\mathbf{x}^t)$ as given in Equation 12.

$$\mathcal{L}^{F_S,con} = ||\sigma(F_T(\mathbf{x}^t)) - \sigma(F_S(\mathbf{P}(\mathbf{x}^t)))||_2^2, \quad (12)$$

where σ is the softmax activation function and $\mathbf{P}(\mathbf{x}^t)$ is a strongly perturbed version of the input image $\mathbf{x}^t$. Similar to (Zhou et al., 2020), we utilise color jittering, Gaussian blurring and noise as perturbations on $\mathbf{x}^t$.

As our I2I model can also leverage pseudo-labels, we perform an identical training with a supervised loss and a consistency loss, but without transformed images $G(\mathbf{x}^s)$ in the *warm-up phase*. We use the resulting model to generate pseudo-labels that we use to train G in the *I2I training phase*. We then use G to translate synthetic images. These translated images in conjunction with the original synthetic images constitute the training data for the *segmentation training phase*.

4 EXPERIMENTS AND RESULTS

In this section, we detail experiments that we conducted in order to show the performance of our

Table 1: Results of domain adaptation from GTA5 to Cityscapes using a VGG backbone.

	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorbike	bicycle	**mIoU**^{19}
Source only (ours)	79.2	30.6	76.0	22.7	11.1	19.2	11.0	2.2	80.3	30.4	73.5	39.3	0.5	75.3	17.3	9.5	0.0	1.4	0.0	30.5
CyCADA (Hoffman et al., 2018)	85.2	37.2	76.5	21.8	15.0	23.8	22.9	21.5	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	**4.5**	9.8	0.0	35.4
CBST-SP (Zou et al., 2018)	90.4	50.8	72.0	18.3	9.5	27.2	28.6	14.1	82.4	25.1	70.8	42.6	14.5	76.9	5.9	12.5	1.2	14.0	28.6	36.1
DCAN (Wu et al., 2018)	82.3	26.7	77.4	23.7	20.5	20.4	30.3	15.9	80.9	25.4	69.5	52.6	11.1	79.6	24.9	21.2	1.3	17.0	6.7	36.2
SWD (Lee et al., 2019)	91.0	35.7	78.0	21.6	21.7	31.8	30.2	25.2	80.2	23.9	74.1	53.1	15.8	79.3	22.1	26.5	1.5	17.2	**30.4**	39.9
SIM (Wang et al., 2020b)	88.1	35.8	83.1	25.8	23.9	29.2	28.8	28.6	83.0	36.7	82.3	53.7	22.8	82.3	26.4	38.6	0.0	19.6	17.1	42.4
TGCF-DA + SE (Choi et al., 2019)	90.2	51.5	81.1	15.0	10.7	**37.5**	35.2	28.9	84.1	32.7	75.9	62.7	19.9	82.6	22.9	28.3	0.0	23.0	25.4	42.5
FADA (Wang et al., 2020a)	92.3	51.1	83.7	33.1	**29.1**	28.5	28.0	21.0	82.6	32.6	85.3	55.2	28.8	83.5	24.4	37.4	0.0	21.1	15.2	43.8
PCEDA (Yang et al., 2020)	90.2	44.7	82.0	28.4	28.4	24.4	33.7	35.6	83.7	40.5	75.1	54.4	28.2	80.3	23.8	39.4	0.0	22.8	30.8	44.6
Zhou *et al*. (Zhou et al., 2020)	**95.1**	**66.5**	84.7	35.1	19.8	31.2	35.0	32.1	86.2	**43.4**	82.5	61.0	25.1	87.1	**35.3**	**46.1**	0.0	24.6	17.5	47.8
Ours	94.4	65.3	**85.9**	**39.0**	22.2	35.4	**39.1**	**37.3**	**86.7**	42.3	**88.1**	**62.7**	**36.2**	**87.6**	33.8	45.0	0.0	**26.5**	24.2	**50.1**
Target only (ours)	96.9	79.7	89.6	46.0	47.6	47.4	55.9	64.5	90.0	60.7	91.7	72.4	49.1	92.4	56.7	75.4	54.0	51.0	69.4	68.0

Table 2: Results of domain adaptation from SYNTHIA to Cityscapes using a VGG backbone. **mIoU**^{16} and **mIoU**^{13} are computed on 16 and 13 classes respectively. * means that classes are included in **mIoU**^{16} but excluded in **mIoU**^{13}.

	road	sidewalk	building	wall*	fence*	pole*	traffic light	traffic sign	vegetation	sky	person	rider	car	bus	motorbike	bicycle	**mIoU**^{16}	**mIoU**^{13}
Source only (ours)	42.0	19.6	60.4	6.3	0.1	28.3	2.1	10.3	76.2	76.0	44.3	7.2	62.5	14.8	3.2	10.6	29.0	33.0
DCAN (Wu et al., 2018)	79.9	30.4	70.8	1.6	**0.6**	22.3	6.7	23.0	76.9	73.9	41.9	16.7	61.7	11.5	10.3	38.6	35.4	-
CBST (Zou et al., 2018)	69.6	28.7	69.5	**12.1**	0.1	25.4	11.9	13.6	82.0	81.9	49.1	14.5	66.0	6.6	3.7	32.4	35.4	40.7
SWD (Lee et al., 2019)	83.3	35.4	**82.1**	-	-	-	**12.2**	12.6	**83.8**	76.5	47.4	12.0	71.5	17.9	1.6	29.7	-	43.5
FADA (Wang et al., 2020a)	80.4	35.9	80.9	2.5	0.3	**30.4**	7.9	22.3	81.8	**83.6**	48.9	16.8	77.7	**31.1**	13.5	17.9	39.5	46.0
TGCF-DA + SE (Choi et al., 2019)	90.1	48.6	80.7	2.2	0.2	27.2	3.2	14.3	82.1	78.4	**54.4**	16.4	82.5	12.3	1.7	21.8	38.5	46.6
PCEDA (Yang et al., 2020)	79.7	35.2	78.7	1.4	**0.6**	23.1	10.0	**28.9**	79.6	81.2	51.2	**25.1**	72.2	24.1	**16.7**	**50.4**	41.1	48.7
Zhou *et al*. (Zhou et al., 2020)	93.1	53.2	81.1	2.6	**0.6**	29.1	7.8	15.7	81.7	81.6	53.6	20.1	82.7	22.9	7.7	31.3	41.5	48.6
Ours	**94.8**	**67.2**	81.9	6.1	0.1	29.6	0.1	19.7	82.2	81.1	50.2	17.0	**84.6**	30.8	12.4	25.1	**42.7**	**49.8**
Target only (ours)	96.9	79.7	89.6	46.0	47.6	47.4	55.9	64.5	90.0	91.7	72.4	49.1	92.4	75.4	51.0	69.4	70.0	75.4

proposed methods. In Section 4.1 we shortly introduce the datasets that we used. In Section 4.2 we give important details on the training and validation protocols. In Section 4.3 we compare our methods to previous state-of-the art methods on two public benchmarks. In Section 4.4 we conduct an ablation study to show the impact of individual components on our results.

4.1 Datasets

Following common practice for unsupervised domain adaptation in semantic segmentation, we use the GTA5 (Richter et al., 2016) and SYNTHIA (Ros et al., 2016) datasets as our synthetic domain and the Cityscapes dataset (Cordts et al., 2016) as our real domain. The datasets are detailed below.

Cityscapes: dataset (Cordts et al., 2016) contains images of urban street scenes collected around Germany and neighbouring countries. It consists of a training set with 2975 images and a validation set of 500 images. We report **mIoU**^{19} for the 19 object classes that are annotated. We also use the available 89250 unlabeled images for learning our I2I method.

GTA5: dataset (Richter et al., 2016) contains images rendered by the game Grand Theft Auto 5. It consists of 24966 images with corresponding pixel-level semantic segmentation annotations and a set of object classes that is compatible to the annotations of the Cityscapes dataset.

SYNTHIA: dataset (Ros et al., 2016) consists of a collection of images rendered from a virtual city. We use the SYNTHIA-RAND-CITYSCAPES subset, which consists of 9400 images with pixel-wise annotations. Note that the *terrain*, *truck*, *train* classes are not annotated in the SYNTHIA dataset. Thus, we use the remaining 16 classes that are common with the Cityscapes dataset. We evaluate **mIoU**^{16} on these 16 classes and **mIoU**^{13} on a subset of 13 classes. **mIoU**^{13} is a common metric for evaluation on the SYNTHIA dataset that excludes certain classes that are especially hard and/or underrepresented in the data.

4.2 Implementation Details

For a fair comparison to previous work (Choi et al., 2019; Zhou et al., 2020), we adopt the VGG16 backbone (Simonyan and Zisserman, 2015) pre-trained

on the ImageNet dataset (Deng et al., 2009). Following Deeplab-v2 (Chen et al., 2018), we incorporate Atrous Spatial Pyramid Pooling (ASPP) as the decoder and use an bi-linear up-sampling to get the segmentation output at image resolution. We use this model for all our experiments. We use the Adam optimiser (Kingma and Ba, 2015) for the segmentation training with an initial learning rate of 1×10^{-5} and exponential weight decay. The generator and discriminator are trained with a constant learning rate of 1×10^{-5} for 1 million iterations. For validation purposes, we keep exponential moving averages of the generator weights. We use a 50/50 split of synthetic and transformed data to learn the segmentation model. All networks are trained with gradient clipping at a global norm of 5. We set all EMA decay values to 0.999. During the first 10.000 training steps of F_S, we keep λ^{con} to zero. We report results of the teacher network which is a smoothed version of the trained segmentation network. We fade λ^{pseudo} and λ^{cgan} linearly from zero at iteration 20,000 to 0.3 at iteration 100,000 when learning with pseudo-labels created online. When learning with pre-computed pseudo-labels from our proposed segmentation model, we keep λ_{pseudo} and λ_{cgan} at 0.3 at all times. All images in all experiments are re-scaled such that the longer side is 1024 pixels. For training, we crop 384×384 regions from these images.

4.3 Comparisons with the State-of-the-Art

We compare the results of our method to state-of-the-art methods that combine I2I and SSL on the two standard benchmarks: "GTA5 $\rightarrow$ Cityscapes" and "SYNTHIA $\rightarrow$ Cityscapes" in Table 1 and Table 2, respectively. Our method improves upon current state-of-the art in both benchmarks. Like many other methods, we suffer from fundamental differences in the definition of certain objects, e.g., the *train* class in the Cityscapes dataset which includes objects like street-cars is actually very similar to the *bus* class in the GTA5 dataset (Zhang et al., 2017). Such a difference obviously impacts performance on both classes. Indeed, due to the similarity of buses in the source data to trains in the target data, we would actually expect the converted buses to incorporate certain features of the target trains. During segmentation training, we then actively enforce predictions for the bus class which ultimately results in poor performance. We identified an additional issue that affects the *car* and *truck* annotations in the GTA5 dataset. Here, one of the synthetic car models is consistently annotated with the *truck* label. We think that our method that directly enforces label consistency is overly prone to such differences and errors in the annotations. Also note that domain adaptation from SYNTHIA to Cityscapes, in general, is inferior to the variant trained on the GTA5 dataset. The domain gap between the SYNTHIA dataset and the Cityscapes dataset is simply wider than the gap between GTA5 and Cityscapes. This is mainly due to the fact that the viewpoint in the scenes in the SYNTHIA dataset varies a lot while Cityscapes and GTA5 mainly show scenes from the viewpoint of a pedestrian or the viewpoint of a driver. The SYNTHIA dataset also consists of less annotated images. Common practice excludes many of the hard classes which results in $\mathbf{mIoU}^{13}$ scores being very similar to the $\mathbf{mIoU}^{19}$ scores that we obtain when training with the GTA5 dataset. The above-mentioned issues apply to most methods. However, our method outperforms other methods. Thus, in the next section, we conduct an ablation study to identify the contribution of the individual components of our method to this outperformance.

Table 3: Ablation study for the proposed image-to-image translation method from the synthetic GTA5 dataset to the Cityscapes dataset. We report mean $\mathbf{mIoU}^{19}$ on the Cityscapes validation set.

Component	$\mathbf{mIoU}^{19}$
LSGAN (Mao et al., 2017)	**43.9**
SGAN (Goodfellow et al., 2014)	42.9
w/o $\mathcal{L}^{D,G,clsgan}$	42.7
w/o $\mathcal{L}^{G,seg}$	38.1
w/o pre-computed pseudo-labels	40.0

4.4 Ablation Study

In this section, we conduct an ablation study on various components of our proposed method. In addition, we compare different components directly to their counterparts in similar methods (Choi et al., 2019; Zhou et al., 2020). We also analyse the remaining domain gap to get a better understanding of our results.

Image-to-Image Translation. The results of the ablation study are shown in Table 3. For simplicity, we omit the SSL part of the third stage of our method, i.e., we use the images from our I2I method to learn a segmentation model in a purely supervised fashion. We train on 50% source data and 50% translated data. As expected, our performance is highly dependent on the semantic consistency loss $\mathcal{L}^{G,seg}$ from Equation 9 that we use to supervise the generator. We lose $\approx$

Table 4: Ablation study and comparison to existing similar work. We evaluate and compare the image translation and constancy regularization. We report results from GTA5 to Cityscapes. $\mathcal{L}^{Fs,seg}$ is the segmentation loss defined in Equation 11. SSL refers to the consistency loss $\mathcal{L}^{Fs,con}$ as defined in Equation 12. *I2I* means that we use translated images from our proposed I2I method.

Method	Component	mIoU
Source only (ours)	$\mathcal{L}^{Fs,seg}$	30.5
(Choi et al., 2019)	$\mathcal{L}^{Fs,seg} + SSL$	32.6
(Zhou et al., 2020)	$\mathcal{L}^{Fs,seg} + SSL$	35.6
Ours	$\mathcal{L}^{Fs,seg} + SSL$	**39.2**
(Choi et al., 2019)	$\mathcal{L}^{Fs,seg} + I2I$	35.4
(Zhou et al., 2020)	$\mathcal{L}^{Fs,seg} + I2I$	35.1
Ours	$\mathcal{L}^{Fs,seg} + I2I$	**43.9**
(Choi et al., 2019)	$\mathcal{L}^{Fs,seg} + SSL + I2I$	42.5
(Zhou et al., 2020)	$\mathcal{L}^{Fs,seg} + SSL + I2I$	47.8
Ours	$\mathcal{L}^{Fs,seg} + SSL + I2I$	**50.1**

5.7% **mIoU**^{19} absolute performance by removing this component which demonstrates the effectiveness of our approach. Note that this ablation evaluates the impact of $\mathcal{L}^{G,seg}$. $\mathcal{L}^{D,seg}$ is still applied during training, i.e., the discriminator has access to the segmentation information but is not able to properly transfer this information to the generator. This can be improved by applying $\mathcal{L}^{G,seg}$ during training which explicitly enforces the transfer of segmentation knowledge. Note that our method is still able to deliver 40% **mIoU**^{19} when using simple pseudo-labels created online during training from the output of the AC-head, i.e., if we train without pseudo-labels supplied by an external method. Nevertheless, using higher quality pseudo-labels from the proposed segmentation method increases performance by an absolute 3.8% **mIoU**^{19}. Removing the class-wise GAN feedback reduces performance by around 1.2% **mIoU**^{19}. Swapping the LSGAN target to a standard GAN (SGAN) target reduces performance by $\approx 1\%$ **mIoU**^{19}.

Figure 2 shows examples of transformed synthetic images. We observe that the textures of roads and trees look much more realistic. Also, the sky is generally more cloudy which is very characteristic for the Cityscapes dataset.

Comparison with Similar Methods. In Table 4, we compare our two main components to the main components of similar work by Choi *et al*. (Choi et al., 2019) and Zhou *et al*. (Zhou et al., 2020). We compare the components in isolation and in combination. Both components improve performance substantially upon previous work. We can clearly see that the improvement of the I2I method is predominantly achieved through the semantic consistency framework that we described in subsection 3.2. However, in Table 3, we can also clearly see, that the performance of our I2I method is heavily impacted by the quality of the pseudo-labels. This shows that both components benefit each other.

Table 5: Domain gap evaluation. Our method closes the domain gap between GTA5 and Cityscapes by 68.3%.

	mIoU^{19}	domain gap
Cityscapes Model	68.0	0.0%
Source only	39.6	100.0%
Ours	59.0	31.7%

4.5 Domain Gap Analysis

In order to estimate the remaining domain gap we retrain the linear classification layer of the segmentation model on real labels from the Cityscapes dataset. We compare to a model that is trained on Cityscapes only. The results are summarised in Table 5. We argue that retraining the classification layer is necessary for a proper comparison because it allows to overcome fundamental differences in class annotations between the synthetic source and the real target data. In essence, this means that we evaluate the quality of the learned features and their applicability to data from the target domain. In this context, feature quality refers to the linear separability of the object classes from the Cityscapes dataset in the features. This linear separability can be assessed by training a linear classifier on top of these features. Such an approach is commonly referred to as a *linear probe* (Alain and Bengio, 2017). We argue that a supervised model that is trained on annotated data from the target domain gives a reasonable upper bound on the achievable segmentation performance. The total domain gap between the target domain $\mathcal{T}$ and the source domain $\mathcal{S}$ then is the difference between the performance of this model and a model trained on source data only. Again, we retrain the linear classification layer of this source model with data from the target domain. This source model achieves 39.6% **mIoU**^{19} compared to the upper bound of 68.0% **mIoU**^{19}. Thus, we can conclude that the total domain gap between $\mathcal{S}$ and $\mathcal{T}$ is equal to 28.4% **mIoU**^{19} points. In comparison, our method achieves 59.0% **mIoU**^{19} when retraining the classification layer. This reduces our estimate of the remaining domain gap to $\approx 9\%$. This equals a reduction of the domain gap by 68.3% when compared to the model that is trained on source data only.

Figure 2: Examples of translated images. The left column shows synthetic images from the GTA5 dataset. The right column shows the corresponding translated images.

5 CONCLUSION

In this work, we have investigated the problem of unsupervised domain adaptation for semantic segmentation. We proposed two complementary approaches in order to reduce the gap between the data of a synthetic source domain and the real-world target domain. More specifically, we have shown that an adversarial image-to-image translation model that is trained with an auxiliary segmentation task on images of both domains yields significantly better results. We show that pseudo-labels can be leveraged to improve this process. The combination of the proposed methods outperforms previous state-of-the-art combinations of image-to-image translation and semi-supervised learning for domain adaptation on relevant benchmarks by a considerable margin.

REFERENCES

Alain, G. and Bengio, Y. (2017). Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net. 8

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848. 2, 7

Chen, Y., Chen, W., Chen, Y., Tsai, B., Wang, Y. F., and Sun, M. (2017). No more discrimination: Cross city adaptation of road scene segmenters. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2011–2020. IEEE Computer Society. 2

Chen, Y., Lin, Y., Yang, M., and Huang, J. (2019). Crdoco: Pixel-level domain transfer with cross-domain consistency. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1791–1800. Computer Vision Foundation / IEEE. 2

Choi, J., Kim, T., and Kim, C. (2019). Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6829–6839. IEEE. 2, 6, 7, 8

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3213–3223. IEEE Computer Society. 6

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference*

on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pages 248–255. IEEE Computer Society. 3, 7

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680. 4, 7

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society. 3

Hoffman, J., Tzeng, E., Park, T., Zhu, J., Isola, P., Saenko, K., Efros, A. A., and Darrell, T. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1994–2003. PMLR. 2, 6

Hoffman, J., Wang, D., Yu, F., and Darrell, T. (2016). Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *CoRR*, abs/1612.02649. 2

Hong, W., Wang, Z., Yang, M., and Yuan, J. (2018). Conditional generative adversarial network for structured domain adaptation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1335–1344. Computer Vision Foundation / IEEE Computer Society. 2

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. 3

Ke, Z., Wang, D., Yan, Q., Ren, J. S. J., and Lau, R. W. H. (2019). Dual student: Breaking the limits of the teacher in semi-supervised learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6727–6735. IEEE. 2

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. 7

Lee, C., Batra, T., Baig, M. H., and Ulbricht, D. (2019). Sliced wasserstein discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10285–10295. Computer Vision Foundation / IEEE. 6

Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. 2

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3431–3440. IEEE Computer Society. 2

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., and Smolley, S. P. (2017). Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2813–2821. IEEE Computer Society. 7

Murez, Z., Kolouri, S., Kriegman, D. J., Ramamoorthi, R., and Kim, K. (2018). Image to image translation for domain adaptation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4500–4509. Computer Vision Foundation / IEEE Computer Society. 2

Odena, A., Olah, C., and Shlens, J. (2017). Conditional image synthesis with auxiliary classifier gans. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2642–2651. PMLR. 4

Pizzati, F., de Charette, R., Zaccaria, M., and Cerri, P. (2020). Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 2979–2987. IEEE. 2

Richter, S. R., Vineet, V., Roth, S., and Koltun, V. (2016). Playing for data: Ground truth from computer games. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pages 102–118. Springer. 6

Ros, G., Sellart, L., Materzynska, J., Vázquez, D., and López, A. M. (2016). The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3234–3243. IEEE Computer Society. 6

Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S., and Chellappa, R. (2018). Learning from synthetic data: Addressing domain shift for semantic segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3752–3761. Computer Vision Foundation / IEEE Computer Society. 2

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015,*

San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 3, 6

Taigman, Y., Polyak, A., and Wolf, L. (2017). Unsupervised cross-domain image generation. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. 3

Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1195–1204. 2, 5

Toldo, M., Maracani, A., Michieli, U., and Zanuttigh, P. (2020). Unsupervised domain adaptation in semantic segmentation: a review. *CoRR*, abs/2005.10876. 2

van Engelen, J. E. and Hoos, H. H. (2020). A survey on semi-supervised learning. *Mach. Learn.*, 109(2):373–440. 2

Wang, H., Shen, T., Zhang, W., Duan, L., and Mei, T. (2020a). Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J., editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV*, volume 12359 of *Lecture Notes in Computer Science*, pages 642–659. Springer. 6

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., and Xiao, B. (2021). Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3349–3364. 2

Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J. (2019). Symmetric cross entropy for robust learning with noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 322–330. IEEE. 4

Wang, Z., Yu, M., Wei, Y., Feris, R., Xiong, J., Hwu, W., Huang, T. S., and Shi, H. (2020b). Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12632–12641. Computer Vision Foundation / IEEE. 6

Wu, Z., Han, X., Lin, Y., Uzunbas, M. G., Goldstein, T., Lim, S., and Davis, L. S. (2018). DCAN: dual channel-wise alignment networks for unsupervised scene adaptation. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, volume 11209 of *Lecture Notes in Computer Science*, pages 535–552. Springer. 6

Xie, Q., Luong, M., Hovy, E. H., and Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10684–10695. Computer Vision Foundation / IEEE. 2

Yang, Y., Lao, D., Sundaramoorthi, G., and Soatto, S. (2020). Phase consistent ecological domain adaptation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9008–9017. Computer Vision Foundation / IEEE. 6

Yu, F. and Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. 2

Zhang, Y., David, P., and Gong, B. (2017). Curriculum domain adaptation for semantic segmentation of urban scenes. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2039–2049. IEEE Computer Society. 7

Zhou, Q., Feng, Z., Cheng, G., Tan, X., Shi, J., and Ma, L. (2020). Uncertainty-aware consistency regularization for cross-domain semantic segmentation. *CoRR*, abs/2004.08878. 5, 6, 7, 8

Zhu, J., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251. IEEE Computer Society. 2

Zou, Y., Yu, Z., Kumar, B. V. K. V., and Wang, J. (2018). Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*, volume 11207 of *Lecture Notes in Computer Science*, pages 297–313. Springer. 6