Phase Consistent Ecological Domain Adaptation

Yanchao Yang* UCLA Vision Lab
Dong Lao* KAUST
Ganesh Sundaramoorthi KAUST & UTRC
Stefano Soatto UCLA Vision Lab

Abstract

We introduce two criteria to regularize the optimization involved in learning a classifier in a domain where no annotated data are available, leveraging annotated data in a different domain, a problem known as unsupervised domain adaptation. We focus on the task of semantic segmentation, where annotated synthetic data are aplenty, but annotating real data is laborious. The first criterion, inspired by visual psychophysics, is that the map between the two image domains be phase-preserving. This restricts the set of possible learned maps, while enabling enough flexibility to transfer semantic information. The second criterion aims to leverage ecological statistics, or regularities in the scene which are manifest in any image of it, regardless of the characteristics of the illuminant or the imaging sensor. It is implemented using a deep neural network that scores the likelihood of each possible segmentation given a single unannotated image. Incorporating these two priors in a standard domain adaptation framework improves performance across the board in the most common unsupervised domain adaptation benchmarks for semantic segmentation.[1]

1. Introduction

Unsupervised domain adaptation (UDA) aims to leverage an annotated "source" dataset in designing learning schemes for a "target" dataset for which no ground-truth is available. This problem arises when annotations are easy to obtain in one domain (*e.g.*, synthetic images) but expensive in another (*e.g.*, real images), and is exacerbated in tasks where the annotation is laborious, as in semantic segmentation where each pixel in an image is assigned one of K labels. If the two datasets are sampled from the same distribution, this is a standard semi-supervised learning problem. The twist in UDA is that the distributions from which source and target data are drawn differ enough that a model trained on the former performs poorly, out-of-the-box, on the latter.

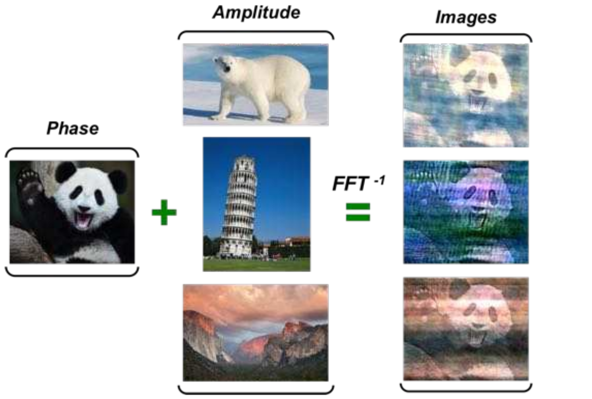

Figure 1. **Semantic content is mainly carried by the phase component of the Fourier Transform.** *Replacing the amplitude component of the image depicting a panda by the amplitude from other images within a wide range, will not prevent us from recognizing a panda in the images generated by the inverse Fourier Transform.*

Typical domain adaptation work employing deep neural networks (DNNs) proceeds by either learning a map that aligns the source and target (marginal) distributions, or by training a backbone to be insensitive to the domain change through an auxiliary discrimination loss for the domain variable. Either way, these approaches operate on the marginal distributions, since the labels are not available in the target domain. However, the marginals could be perfectly aligned, yet the labels could be scrambled: Trees in one domain could map to houses in another, and vice-versa. Since we want to transfer information about the classes, ideally we would want to align the class-conditional distributions, which we do not have. Recent improvements in UDA, for instance cycle-consistency, only enforce the invertibility of the map, but not preservation of semantic information such as the class identity, see Fig. 2. Since the problem is ill-posed, constraints or prior have to be enforced in UDA.

We introduce two priors or constraints, one on the map between the domains, the other on the classifier in the target domain, both unknown at the outset.

For the map between domains, we know from visual psychophysics that semantic information in images tends to be associated with the phase of its Fourier transform. Changes in the amplitude of the Fourier transform can significantly alter the appearance but not the interpretation of the image.

*These two authors contributed equally. Please send correspondence to yanchao.yang@cs.ucla.edu and dong.lao@kaust.edu.sa.

[1] Code available at: https://github.com/donglao/PCEDA

This suggests placing an incentive for the transformation between domains to be phase-preserving. Indeed, we show from ablation studies that imposing phase consistency improves the performance of current UDA models.

For the classifier in the target domain, even in the absence of annotations, a target image informs the set of possible hypotheses (segmentations), due to the statistical regularities of natural scenes (ecological statistics, [3, 11]). Semantic segments are unlikely to straddle many boundaries in the image plane, and their shape is unlikely to be highly irregular due to the regularity of the shape of objects in the scene. Such generic priors, informed by each single un-annotated images, could be learned from other (annotated) images and transfer across image domains, since they arise from properties of the scene they portray. We use a Conditional Prior Network [42] to learn a data-dependent prior on segmentations that can be imposed in an end-to-end framework when learning a classifier in the target domain in UDA.

These two priors yield improvement in UDA benchmarks. We conduct ablation studies to quantify the effect of each prior on the overall performance of learned classifiers (segmentation networks).

In the next section, we describe current approaches to UDA and then describe our method, which is summarized in Sect. 2.5, before testing it empirically in Sect. 3.

1.1. Related Work

Early works on UDA mainly focus on image classification [15, 12, 1], by minimizing a discrepancy measure between two domains [14]. Recent methods apply adversarial learning [13, 39] for classification, by instantiating a discriminator that encourages the alignment in feature space [33, 23, 35]. Unfortunately, none of these methods achieves the same success on semantic segmentation tasks.

Recent progress in image-to-image transformation techniques [46, 25] aligns domains in image space, with some benefit to semantic segmentation [20, 19]. [20] is the first UDA semantic segmentation method utilizing both global and categorical adaptation techniques. CyCADA [19] adapts representations in both image and feature space while enforcing cycle-consistency to regularize the image transformation network. [32] also applies image alignment by projecting the learned intermediate features into the image space. [45] proposes curriculum learning to gradually minimize the domain gap using anchor points. [41] reduces domain shift at both image and feature levels by aligning statistics in each channel of CNN feature maps in order to preserve spatial structures. [16] generates a sequence of intermediate shifted domains from source to target to further improve the transferability by providing multi-style translations. [26] introduces a category-level adversarial network to prevent the degeneration of well-aligned categories during global alignment. [21] conditions on both source images and random noise to produce samples that appear similar to the target. Despite the difficulty in training the domain discriminators, generally, the alignment criteria provided by the domain discriminators do not guarantee consistency of the semantic content between the original and transformed images. In addition to cycle-consistency, [24, 8] propose using the segmentation network on the target domain to encourage better semantic consistency. However, this will make the performance depend highly on the employed surrogate network.

In psychophysics, [28] demonstrates that certain phase modifications can hinder or prevent the recognition of visual scenes. [27] shows that many important features of a signal can be preserved by the phase component of the Fourier Transform, and under some conditions a signal can be completely reconstructed with only the phase. Moreover, [17] shows psychophysically that the Fourier phase spectrum plays a critical role in human vision. The concurrent work [43] shows that swapping the amplitude component of an image with one from the other domain preserves the semantic content while aligning the two domains. With all these observations, we propose to use phase to provide an effective semantic consistency constraint that does not depend on any surrogate networks.

Besides discriminators applied to the image or in feature space, [37, 38] find that adaptation on the structured output space is also beneficial for semantic segmentation. [7] proposes spatially-aware adaptation along with target guided distillation using activation supervision with a pretrained classification network. Further, [6] proposes a geometrically guided adaptation aided with depth in a multitask learning framework. [4] extracts the domain invariant structure from the image to disentangle images into domain invariant structure and domain-specific variations. [47] performs iterative class-balanced self-training as well as refinement of the generated pseudo-labels using a spatial prior. A similar strategy is also applied in [24, 38]. [40] approaches UDA for semantic segmentation by entropy minimization of the pixel-wise predictions. An adversarial loss on the entropy map is also used to introduce regularity in the output space. However, none of them explicitly models the scene compatibility that regularizes the training of the target domain segmentation network.

2. Method

We first describe general image translation for unsupervised domain adaptation (UDA) and how it is used in semantic segmentation. We point to some drawbacks as inspiration for the two complementary constraints, which we introduce in Sect. 2.3 and 2.4, and incorporate into a model of UDA for semantic segmentation in Sect. 2.5, which we validate empirically in Sect. 3.

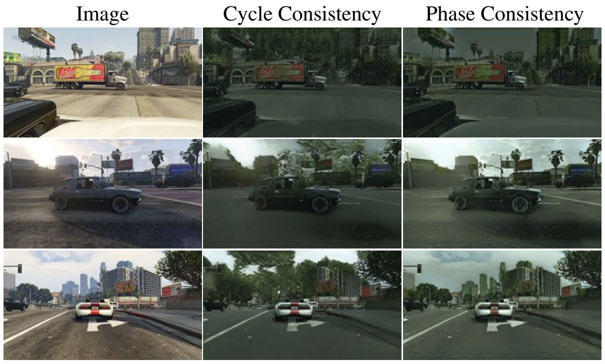

Figure 2. **Cycle consistency does not guarantee semantic consistency, but phase does.** *Note the sky is transformed to trees (1st row), the cloud is transformed to a mountain (2nd row), and the buildings are also transformed to trees (3rd row) even if cycle consistency is enforced. Phase consistency enforces the semantic information to be preserved and provides enough flexibility to align the two domains.*

2.1. Preliminaries: Image Translation for UDA

We consider two probabilities, a source P^s and a target P^τ, which are generally different (covariate shift), as measured by the Kullbach-Liebler divergence $KL(P^s||P^\tau)$. In UDA, we are given ground-truth annotation in the source domain only. So, if $x \in \mathbb{R}^{H\times W\times 3}$ are color images, and $y \in [1,\dots,K]^{H\times W}$ are segmentation masks where each pixel has an associated label, we have images and segmentations in the source domain, $D^s = \{(x_i^s, y_i^s) \sim P^s(x,y)\}_{i=1}^{N_s}$ but only images in the target domain, $\{x_i^\tau \sim P^\tau(x)\}_{i=1}^{N_\tau}$. The goal of UDA for semantic segmentation is to train a model ϕ^τ, for instance a deep neural network (DNN), that maps target images to estimated segmentations, $x^\tau \mapsto \hat{y}^\tau = \arg\max_y \phi^\tau(x^\tau)_y$, leveraging source domain annotations. Because of the covariate shift, simply applying to the target data a model trained on the source generally yields disappointing results. As observed in [2], the upper bound on the target domain risk can be minimized by reducing the gap between two distributions. Any invertible map T between samples in the source and target domains, for instance $x^s \mapsto T(x^s)$ induces a (pushforward) map between their distributions $P^s \mapsto T_*P^s$ where $T_*P^s(x^\tau) = P^s(T^{-1}(x^\tau))$. The map can be implemented by a "transformer" network, and the target domain risk is minimized by the cross-entropy loss, whose empirical approximation is:

$$\mathcal{L}_{ce}(\phi^\tau, T; D^s) = \sum_{(x_i,y_i)\in D^s} -\log[\phi^\tau(T(x_i))]_{y_i} \qquad (1)$$

where T maps data sampled from the source distribution to the target domain. The gap is measured by $KL(P^\tau||T_*P^s)$, and can be minimized by (adversarially) maximizing the domain confusion, as measured by a domain discriminator θ that maps each image into the probability of it coming from the source or target domains:

$$\mathcal{L}_D(\theta, T; x_i^s) = -\log[\theta(T(x_i^s))]. \qquad (2)$$

Ideally, θ returns 1 for images drawn from the target P^τ, and 0 otherwise.

2.2. Limitations and Challenges

Ideally, jointly minimizing the two previous equations would yield a segmentation model that operates in the target domain, producing estimated segmentations $y^\tau = \phi(x^\tau)$. Unfortunately, a transformation network T trained by minimizing Eq. (2) does not yield a good target domain classifier, as T is only asked by Eq. (2) to match the marginals, which it could do while scrambling all labels (images of class i in the source can be mapped to images of class j in the target). In other words, the transformation network can match the image statistics, but there is nothing that encourages it to match semantics. Cycle-consistency [46, 19] does not address this issue, as it only enforces the invertibility of T:

$$\mathcal{L}_{cyc}(T, T^{-1}; x_i^s) = \|x_i^s - T^{-1}\circ T(x_i^s)\|_1. \qquad (3)$$

Even after imposing this constraint, buildings in the source domain could be mapped to trees in the target domain, and vice-versa (Fig. 2). Ideally, if ϕ^s is a model trained on the source, and ϕ^τ the one operating on the target, we would like:

$$\phi^s(x_i^s) = \phi^\tau(T(x_i^s)), \forall i. \qquad (4)$$

Unfortunately, training ϕ^τ would require ground-truth in the target domain, which is unknown. We could use ϕ^s as a surrogate, apply ϕ^s on the target domain, and penalize the discrepancy between the two sides in Eq. (4) with respect to the unknowns. Absent any regularization, this yields the trivial result where $T(x) = x$ and $\phi^\tau = \phi^s$. While Eq. (4) is useless in providing information on T *and* ϕ^τ, it can be seen as a vehicle to transfer prior information *from one* (*e.g.*, T) *onto the other* (*e.g.*, ϕ^τ). In the next two sections we discuss additional constraints and priors that can be imposed on T (Sect. 2.3) and ϕ^τ (Sect. 2.4) that make the above constraint non-trivial, and usable in the context of UDA.

2.3. Phase Consistency

It is well known in perceptual psychology that manipulating the spectrum of an image can lead to different effects: Changes in the amplitude of the Fourier transform alters the image but does not affect its interpretation, whereas altering the phase produces uninterpretable images [22, 28, 27, 17]. This is illustrated in Fig. 1, where the amplitude of the Fourier transform of an image of a panda is replaced with the amplitude from an image of a bear, a tourist landmark and a landscape, yet the reconstructed images portray a panda. In other words, it appears that semantic information is included in the phase, not the amplitude, of the spectrum. This motivates us to hypothesize that the transformation T should be phase-preserving.

To this end, let $\mathcal{F} : \mathbb{R}^{H\times W} \rightarrow \mathbb{R}^{H\times W\times 2}$ be the Fourier Transform. Phase consistency, for a transformation T, for a single channel image x, is obtained by minimizing:

$$\mathcal{L}_{ph}(T; x) = -\sum_j \frac{\langle \mathcal{F}(x)_j, \mathcal{F}(T(x))_j \rangle}{\|\mathcal{F}(x)_j\|_2 \cdot \|\mathcal{F}(T(x))_j\|_2} \quad (5)$$

where $\langle , \rangle$ is the dot-product, and $\|\cdot\|_2$ is the L_2 norm. Note that Eq. (5) is the negative cosine of the difference between the original and the transformed phases, thus, by minimizing Eq. (5) we can directly minimize their difference and increase phase consistency. We demonstrate the effectiveness of phase consistency in the ablation studies in Sect. 3.

2.4. Prior on Scene Compatibility

While target images have no ground-truth labeling, not all semantic segmentations are equally likely at the outset. Given an unlabeled image, we may not know what classes $\{1, \ldots, K\}$ may interest a user, but we do know that objects in the scene have certain regularities, so it is unlikely that photometrically homogeneous regions are segmented into many pieces, or that a class segment straddles many image boundaries. It is also unlikely that the segmented map is highly irregular. These characteristics inform the probability of a segmentation given the image in the target domain, $Q(\phi(x)|x)$. Q can be thought of as a function that scores each hypothesis $\phi(x)$ based on the plausibility of the resulting segmentation given the input image x. The function can be learned using images for which the ground-truth segmentation is given, for instance the source dataset D^s, and then used at inference time as a scoring function. Such a scoring function can be implemented by a Conditional Prior Network (CPN) [42]. However, note that $D^s = \{(x_i^s, y_i^s)\}$ is sampled from $P^s(x, y)$. Simply training a CPN with D^s will make $Q(y|x)$ approximate $P^s(y|x)$[2], making the exercise moot. The CPN would capture both the domain-related unary prediction term and the domain irrelevant pairwise term that depends on the image structure. To make this point explicit, we can decompose $P^s(y|x)$ as follows:

$$P^s(y|x) \approx \prod_j P^s(y_j|x) \prod_{m<n} P(y_m = y_n|x) \quad (6)$$

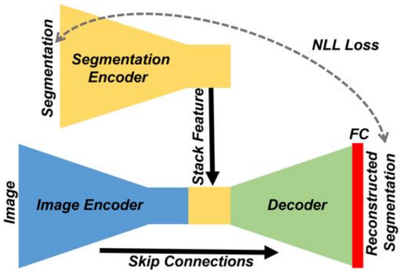

Figure 3. **CPN Architecture.** *To reconstruct the segmentation map that is encoded into a narrow bottleneck, the decoder needs to leverage structural information from the image. The CPN reconstructs better the prediction $\phi(x)$ with higher compatibility with x. Labels are randomly permuted during training to avoid overfitting to the domain dependent unary prediction terms in Eq. (6).*

where we omit higher-order terms for simplicity. The unary terms $P^s(y_j|x)$ measure the likelihood of the semantic label of a single pixel given the image; *e.g.*, pixels in a white region indicate sky in the source domain, which depends highly on the domains. The pairwise terms $P(y_m = y_n|x)$ measure the labeling compatibility between pixels, which would depend much less on the domain; *e.g.*, pixels in a white region may not be sky in the target domain, but they should be labeled the same. Absent at least binary terms, the unary terms would lead to overfitting the source domain. To prevent this, we randomly permute the labels in y^s according to a uniform distribution:

$$y^s \mid_{y^s=i} = \mathrm{PM}^K(i) \quad (7)$$

where PM^K is a random permutation of the class ID's for K classes, and we denote the permuted semantic segmentation masks as $\hat{y^s}$, which scales the original dataset up in size by a factor of $K!$. We denote the new source dataset with permuted ground-truth masks as $\hat{D^s}$, which will render the conditional distribution invariant to the domain-dependent unaries, *i.e.*:

$$\hat{P^s}(y|x) \approx \prod_{m<n} P(y_m = y_n|x) \quad (8)$$

Note, $\hat{P^s}(y|x)$ only evaluates the compatibility based on the segmentation layout but not the semantic meanings. Thus, we train a CPN Q using the following training loss [42] with an information capacity constraint:

$$\min_Q \mathbb{E}_x KL(\hat{P^s}(y|x), Q(y|x)) + \beta\mathbb{I}(y, Q^e(y)) \quad (9)$$

where $\mathbb{I}$ denotes the mutual information between y and its CPN encoding $Q^e(y)$. Then, we obtain a compatibility function

$$Q(y|x) \approx \prod_{m<n} P(y_m = y_n|x) \quad (10)$$

The proposed CPN architecture is illustrated in Fig. 3 and the training details, including the encoding metric, is described in Sect. 3. We now summarize the overall training loss, that exploits regularities implied by each constraint.

[2] We abuse the notation and use y to indicate both the class and the soft-max (log-likelihood) vector that approximates its indicator function.

Architecture	SSL	CPN	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	mIoU
A			88.2	41.3	**83.2**	28.8	21.9	**31.7**	**35.2**	28.2	**83.0**	26.2	**83.2**	57.6	27.0	77.1	27.5	34.6	2.5	28.3	**36.1**	44.3
A		✓	**91.4**	**47.2**	82.9	**29.2**	**22.9**	31.4	33.3	**30.2**	80.8	**27.8**	81.3	**59.1**	**27.7**	**84.4**	**31.5**	**40.9**	**3.2**	**30.2**	24.5	**45.3**
A	✓		91.2	46.1	83.9	31.6	20.6	29.9	36.4	31.9	**85.0**	39.7	**84.7**	57.5	29.6	83.1	38.8	**46.9**	**2.5**	27.5	38.2	47.6
A	✓	✓	**91.3**	**48.2**	**85.0**	**39.4**	**26.1**	**32.4**	**37.4**	**40.7**	84.9	**41.9**	83.0	**59.8**	**30.2**	**83.6**	**40.0**	46.1	0.1	**31.7**	**43.3**	**49.7**
B			**86.4**	39.5	79.2	27.4	**24.3**	**23.4**	29.0	18.0	80.5	33.2	70.1	47.2	18.1	75.4	20.6	23.3	0.0	**16.1**	5.4	37.7
B		✓	86.0	**39.9**	**80.6**	**32.3**	21.9	21.6	**29.5**	**23.9**	**83.1**	**37.5**	**75.9**	**53.2**	**24.4**	**79.3**	**22.8**	**32.4**	**0.9**	13.9	**18.9**	**40.9**
B	✓		89.2	40.9	**81.2**	29.1	19.2	14.2	29.0	19.6	**83.7**	35.9	**80.7**	**54.7**	23.3	**82.7**	**25.8**	28.0	**2.3**	**25.7**	19.9	41.3
B	✓	✓	**90.1**	**44.7**	81.0	**29.3**	**26.4**	**20.9**	**33.7**	**34.3**	83.4	**37.4**	71.2	54.0	**27.4**	79.9	23.7	**39.6**	1.1	18.5	**22.6**	**43.1**
C			79.1	33.1	77.9	**23.4**	**17.3**	32.1	33.3	31.8	81.5	26.7	69.0	**62.8**	**14.7**	74.5	**20.9**	**25.6**	6.9	**18.8**	**20.4**	39.5
C		✓	**89.1**	**41.4**	**81.2**	22.2	15.3	**34.0**	**35.0**	**37.1**	**84.8**	**32.1**	**76.2**	61.7	12.5	**82.1**	20.8	25.2	**7.3**	15.6	18.9	**41.7**

Table 1. **The learned Scene Compatibility improves segmentation accuracy.** *Training the segmentation model with the learned scene compatibility Q improves segmentation accuracy under all experimental settings, with different network backbones: A: ResNet-101, B: VGG-16, C: DRN-26. SSL: Self-supervised Learning. Note that whenever Q is added in the training loss while the other terms are fixed, the overall semantic segmentation performance gets improved.*

2.5. Overall Training Loss

Combining the adversarial losses and our novel constraints for both phase consistency and scene compatibility, we have the overall training loss for the proposed domain adaptation method for training the image transformation networks T, T^{-1} and the target domain segmentation network ϕ^τ:

$$\begin{aligned}
&\mathcal{L}(\phi^\tau, T, T^{-1}; \theta^s, \theta^\tau, x_i^s, y_i^s, x_i^\tau) = \\
&\qquad \lambda_D(\mathcal{L}_D(\theta^\tau, T; x_i^s) + \mathcal{L}_D(\theta^s, T^{-1}; x_i^\tau)) \\
&\qquad + \lambda_{cyc}(\mathcal{L}_{cyc}(T, T^{-1}; x_i^s) + \mathcal{L}_{cyc}(T^{-1}, T; x_i^\tau)) \\
&\qquad + \lambda_{ph}(\mathcal{L}_{ph}(T; x_i^s) + \mathcal{L}_{ph}(T^{-1}; x_i^\tau)) \\
&\qquad \mathcal{L}_{ce}(\phi^\tau, T; x_i^s, y_i^s) - \lambda_{cpn} \log[Q(\phi^\tau(x_i^\tau)|x_i^\tau)] \qquad (11)
\end{aligned}$$

with λ's the corresponding weights on each term (hyperparameters), whose values will be reported in Sect.3. Note when training ϕ^τ using Eq. (11), we do not permute its output to evaluate the scene compatibility term. And the scene compatibility Q is fixed after it is trained using Eq. (9). We follow the standard procedure in [19, 24] to train the domain discriminators.

3. Experiments

We evaluate the proposed UDA method on synthetic-to-real semantic segmentation tasks, where the source images (GTA5 [9] and Synthia [31]) and corresponding annotations are generated using graphics engines, and the adapted segmentation models are tested on real-world images. We use average intersection-over-union score (mIoU) across semantic classes as the evaluation metric in all experiments. Moreover, the frequency weighted IoU (fwIoU), which is the sum of the IoUs of different classes but weighted by how frequent a certain class appears in the dataset, is calculated and compared in the GTA5-to-Cityscapes experiments.

We first describe the data used for training and the implementation details, followed by a comprehensive ablation study demonstrating the effectiveness of each proposed component in our method. Then we show quantitative and qualitative comparisons against the state-of-the-art methods, using networks with different backbones, on the GTA5-to-Cityscapes and Synthia-to-Cityscapes benchmarks.

3.1. Datasets

Cityscapes [9] is a real-world semantic segmentation dataset containing 2975 street view training images and 500 validation images with original resolution 2048 × 1024, which is resized to 1024 × 512 for training. The images are collected during the day in multiple European cities and densely annotated. We train the image transformation network and the adapted segmentation network using the training set, and report the result on the validation set.

GTA5 [29] contains 24966 synthesized images from the Grand Theft Auto game with resolution 1914 × 1052. It exhibits a wide range of variations including weather and illumination. We resize the images to 1280 × 720 and use the 19 compatible classes for the training and evaluation.

Synthia [31] is a synthetic dataset focusing on driving scenarios rendered from a virtual city. We use the SYNTHIA-RAND-CITYSCAPES subset as source data, which contains 9400 images with the resolution of 1280 × 760 for training the 16 common classes with Cityscapes, and we evaluate the trained network using both the 16 classes or a subset of 13 classes following previous works [37, 24, 10].

3.2. Implementation Details

Image Transformation Network: We adapt the public CycleGAN [46] framework, and use the "cycle_gan" model therein. We set $\lambda_D = 1.0$, $\lambda_{cyc} = 10.0$ and $\lambda_{ph} = 5.0$ for training the image transformation networks T, T^{-1}. Images

Method	Surrogate	Output Space	mIoU
CyCADA [19]	✓		43.5
	✓	✓	43.1
AdaptSegNet [37]			36.6
	✓		39.3
	✓	✓	42.4
BDL [24]		✓	41.1
	✓*	✓	42.7
	✓†	✓	44.4
Ours		✓	44.8
			45.3

Table 2. **Phase consistency (ours) achieves better performance.** *Note our model trained only with the phase consistency outperforms other methods that utilize a surrogate network to impose semantic consistency (Surrogate), or employ output space regularization (Output Space). * and † : first and second round of improved image transformation using a self-trained surrogate network.*

from source and target domains are resized to 1024×512 and then cropped to 452×452 before feeding into the network. We set the batch-size to 1.0 and use "resnet_9blocks" as the backbone.

Conditional Prior Network: We adopt the standard UNet [30] architecture, and add the segmentation encoder branch. We instantiate 6 convolutional layers, whose channel numbers are $\{16, 32, 64, 128, 256, 256\}$, to encode the image. Each of the first five layers is followed by 2×2 max pooling, similarly, for semantic segmentation maps. Encoded image and segmentation are stacked at the bottleneck, then passed through a 6-layer decoder with channel numbers $\{512, 256, 128, 64, 32, 16\}$, followed by a fully connected layer for class prediction. Skip connections are enabled between the image encoder and the decoder. The network is trained with batch size four by ADAM optimizer with initial learning rate 1e-4. The learning rate is reduced by a factor of 10 after every 30000 iterations.

During training, the network aims at reconstructing the encoded $\hat{y}^s$, which is the randomly permutated ground-truth segmentation, by utilizing image information, leading to the training loss:

$$\mathcal{L}_{cpn}(Q; \hat{y}^s, x^s) = \text{NLL}(Q(\hat{y}^s|x^s), \hat{y}^s) \tag{12}$$

where NLL denotes the negative log likelihood loss derived from the KL-divergence term in the CPN training loss Eq. (9). Lower indicates better scene compatibility *i.e.* higher $Q(y|x)$. Note the information capacity constraint in Eq. (9) is implemented by a structural bottleneck as in [42].

Semantic Segmentation Network: We experiment with different segmentation network backbones. Due to memory constraint, we choose to train the segmentation network after the transformation networks are trained. We first train from scratch the segmentation network using transformed source images and the corresponding annotations using Eq. (11). We fix $\lambda_{cpn} = 0.5$ for all the experiments. Finally, we apply the self-supervised training technique as in [24, 38] to further improve the performance on the target domain. We accept the high confidence (> 0.9) predictions as the pseudo labels. All networks are trained using the ADAM optimizer, with learning rate 2.5e-4, 1e-5, and 1e-4 for ResNet-101, VGG-16, and DRN-26, respectively.

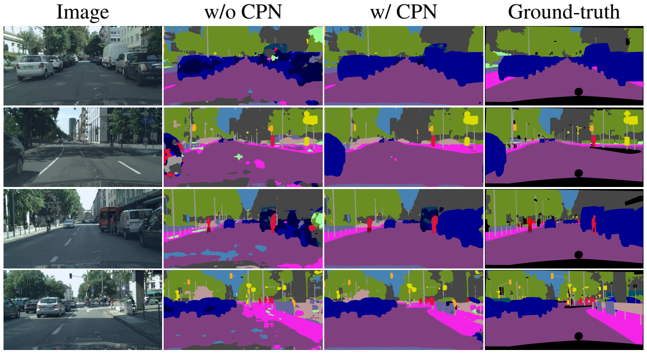

Figure 4. **The learned scene compatibility prior imposes regularity on the predictions.** *When the scene compatibility is added, the segmentation network yields predictions better aligned with object boundaries, and are more consistent within the objects.*

3.3. Ablation Study

Here we carry out an ablation study to investigate the effectiveness and robustness of the proposed priors.

Phase Consistency: Here we train the segmentation network Deeplab-V2 [5] on the transformed source dataset with phase consistency. To make the comparison fair, all competing methods also use the same segmentation network as ours. The results of [37] and [24] are reported by the original papers. We retrain [19] and report its best performance with hyperparameter tuning. The result is presented in Tab. 2. Without any surrogate semantic consistency provided by a surrogate semantic segmentation network, our segmentation model achieves higher accuracy. Note that introducing surrogate semantic consistency for regularizing the transformation networks will also incur more memory cost. Moreover, several rounds of training to improve the performance of the surrogate network can also be time-consuming. However, our phase consistency can be implemented at low computational overhead (see Sect. 3.5).

Interestingly, output space regularization, which aligns the marginal distributions of the segmentations, occasionally leads to worse performance in some settings, including [19] and ours. This is somewhat reasonable since aligning the marginal distributions does not guarantee the conditional alignment given the observations.

Scene Compatibility: To better understand the performance gain from the scene compatibility prior, we compare to competing methods on the same transformed source images. We collect the scores for all the other methods using the same setting as ours, if needed, we retrain their model.

Method	Architecture	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	mIoU	fwIoU
AdaptSegNet [37]	A	86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	**7.2**	29.5	32.5	41.4	75.5
DCAN [41]	A	85.0	30.8	81.3	25.8	21.2	22.2	25.4	26.6	83.4	36.7	76.2	58.9	24.9	80.7	29.5	42.9	2.5	26.9	11.6	41.7	76.2
CyCADA [19]	A	88.3	40.9	81.4	26.9	19.7	31.3	31.8	31.9	81.6	22.3	77.1	56.3	25.1	80.8	33.4	38.6	0.0	24.6	35.5	43.6	77.9
SSF-DAN [10]	A	90.3	38.9	81.7	24.8	22.9	30.5	37.0	21.2	84.8	38.8	76.9	58.8	30.7	**85.7**	30.6	38.1	5.9	28.3	36.9	45.4	79.6
BDL [24]	A	**91.0**	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	**43.6**	83.0	58.6	31.6	83.3	35.3	**49.7**	3.3	28.8	35.6	48.5	81.1
Ours	A	**91.0**	**49.2**	**85.6**	**37.2**	**29.7**	**33.7**	**38.1**	**39.2**	**85.4**	35.4	**85.1**	**61.1**	**32.8**	84.1	**45.6**	46.9	0.0	**34.2**	**44.5**	**50.5**	**82.0**
AdaptSegNet [37]	B	87.3	29.8	78.6	21.1	18.2	22.5	21.5	11.0	79.7	29.6	71.3	46.8	6.5	80.1	23.0	26.9	0.0	10.6	0.3	35.0	74.9
CyCADA [19]	B	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	73.8
DCAN [41]	B	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	72.9
SSF-DAN [10]	B	88.7	32.1	79.5	**29.9**	22.0	23.8	21.7	10.7	80.8	29.8	72.5	49.5	16.1	82.1	23.2	18.1	3.5	24.4	8.1	37.7	76.3
BDL [24]	B	89.2	40.9	81.2	29.1	19.2	14.2	29.0	19.6	**83.7**	35.9	**80.7**	**54.7**	23.3	**82.7**	**25.8**	28.0	2.3	**25.7**	19.9	41.3	78.4
Ours	B	**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**	**79.3**
CyCADA [19]	C	79.1	33.1	77.9	23.4	17.3	32.1	33.3	31.8	81.5	26.7	69.0	**62.8**	**14.7**	74.5	**20.9**	25.6	6.9	**18.8**	20.4	39.5	72.7
Ours	C	**90.7**	**49.8**	**81.9**	**23.4**	**18.5**	**37.3**	**35.5**	**34.3**	**82.9**	**36.5**	**75.8**	61.8	12.4	**83.2**	19.2	**26.1**	4.0	14.3	**21.8**	**42.6**	**79.7**

Table 3. **Quantitative Evaluation on the GTA5-to-Cityscapes benchmark.** *Our method achieves the best mIoU and fwIoU using different segmentation architectures: A (ResNet-101), B (VGG-16), C (DRN-26).*

Method	Architecture	road	sidewalk	building	wall*	fence*	pole*	light	sign	vegetation	sky	person	rider	car	bus	motorcycle	bicycle	mIoU	mIoU*
AdaptPatch [38]	A	82.4	38.0	78.6	8.7	0.6	26.0	3.9	11.1	75.5	84.6	53.5	21.6	71.4	32.6	19.3	31.7	40.0	46.5
AdaptSegNet [37]	A	84.3	42.7	77.5	-	-	-	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	-	46.7
SSF-DAN [10]	A	84.6	41.7	80.8	-	-	-	11.5	14.7	**80.8**	**85.3**	**57.5**	21.6	**82.0**	36.0	19.3	34.5	-	50.0
BDL [24]	A	**86.0**	**46.7**	80.3	-	-	-	14.1	11.6	79.2	81.3	54.1	27.9	73.7	**42.2**	25.7	45.3	-	51.4
Ours	A	85.9	44.6	**80.8**	**9.0**	**0.8**	**32.1**	**24.8**	**23.1**	79.5	83.1	57.2	**29.3**	73.5	34.8	**32.4**	**48.2**	**46.2**	**53.6**
AdaptSegNet [37]	B	78.9	29.2	75.5	-	-	-	0.1	4.8	72.6	76.7	43.4	8.8	71.1	16.0	3.6	8.4	-	37.6
AdaptPatch [38]	B	72.6	29.5	77.2	**3.5**	0.4	21.0	1.4	7.9	73.3	79.0	45.7	14.5	69.4	19.6	7.4	16.5	33.7	39.6
DCAN [41]	B	**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	41.7
BDL [24]	B	72.0	30.3	74.5	0.1	0.3	**24.6**	**10.2**	25.2	**80.5**	80.0	**54.7**	23.2	**72.7**	24.0	7.5	44.9	39.0	46.1
Ours	B	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**

Table 4. **Quantitative Evaluation on the Synthia-to-Cityscapes Benchmark.** *mIoU and mIoU* are the mean IoU computed on the 16 classes and the 13 subclasses respectively (* excluded). Our method achieves the best performance using different segmentation network backbones: A (ResNet-101), B(VGG-16).*

In Tab. 1, we show that under all experimental settings, scene compatibility prior improves accuracy for most of the semantic classes as well as the overall average. The performance gain is preserved during self-supervised learning. We present qualitative comparisons in Fig. 4, showing that the scene compatibility prior provides strong spatial regularity to align segmentation to the object boundaries. Incorporating the scene compatibility prior into the training process significantly improves the overall segmentation smoothness and integrity, resulting in more consistent label prediction within each object.

3.4. Benchmark Results

To recall, feature space alignment has been explored by **DCAN** [41] and **CyCADA** [19]. CyCADA also applies image level domain alignment by training cross-domain cycle consistent image transformation. Output space alignment methods include **AdaptSegNet** [37], **AdaptPatch** [38] and **SSF-DAN** [10], in which various ways of adversarial learning to the segmentation output are applied for better domain confusion. **BDL** [24] propagates information from semantic segmentation back to the image transformation network as semantic consistent regularization.

We apply ResNet-101 [18] based Deeplab-V2 [5] and VGG-16 [36] based FCN-8s [34] for the segmentation network to compare with [37, 41, 24, 40, 10] under the same experimental setting. To better understand the robustness to different neural network settings, We also apply our method to the retrain the DRN-26 [44] model from [19].

The result on the GTA5-to-Cityscapes benchmark is summarized in Tab. 3. Our method achieves state-of-the-art performance with all network backbones in terms of mIoU and fwIoU. Moreover, across different settings, our method achieves the best score for most of the classes, indicating that the proposed priors improve the segmentation accuracy consistently across different semantic categories. We also present a qualitative comparison in Fig. 5. Our proposed method outputs more spatially regularized predictions, which are also consistent with the scene structures. We relatively achieve 4.1% and 8.0% improvement over

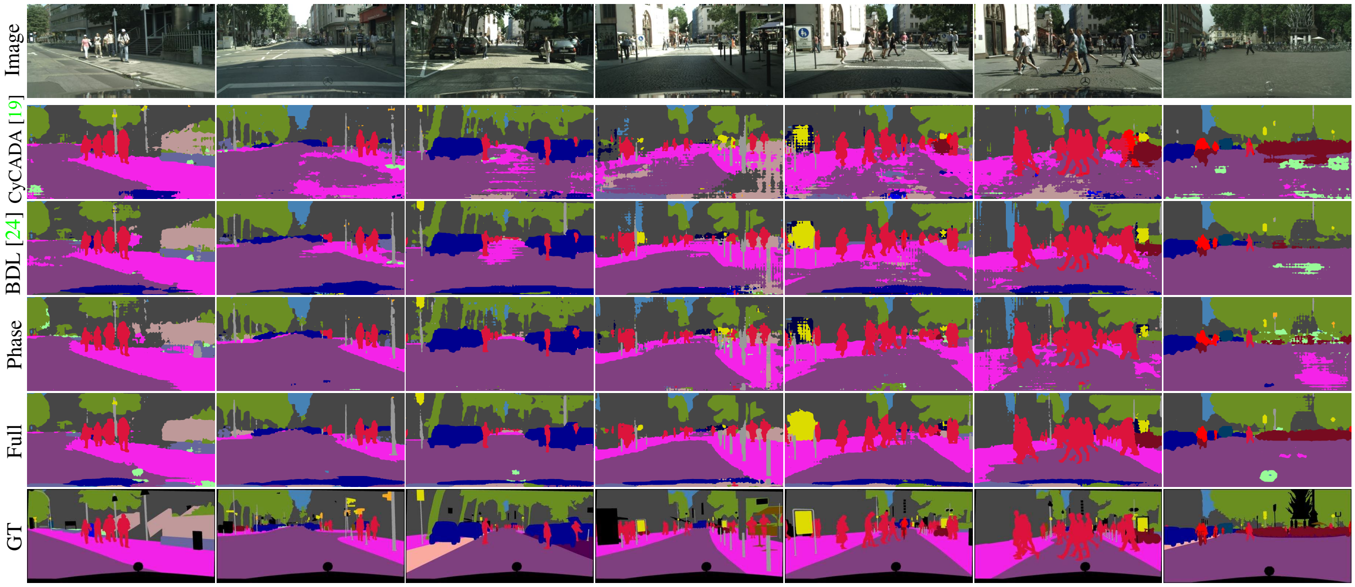

Figure 5. **Qualitative comparison with state-of-the-art methods.** *Our method outputs more spatially regularized segmentations align well with the underlying scene structure. All visualized models are based on DeepLab-V2 with ResNet-101 under the same setting. Phase: trained with phase consistency only; Full: our full model.*

the second-best method with the backbone ResNet-101 and VGG-16, respectively.

The result on the Synthia-to-Cityscapes benchmark can be found in Tab. 4. The mIoUs of either 13 or 16 classes are evaluated according to the evaluation protocol in the literature. Our method outperforms competing methods on both sets. It also achieves the best result on most of the semantic categories. Again, we relatively achieve 4.3% and 5.4% improvement over the second-best using different backbones.

3.5. Computational Cost

All networks are trained using a single Nvidia Titan Xp GPU. Enforcing the phase consistency will incur a <0.001s overhead for a 1024×512 image, which is negligible. Training the CPN for scene compatibility takes 2.5 seconds to process a batch of 4 images, given the images are cropped to 1280 × 768. Incorporating CPN into segmentation training adds 1.5 seconds overhead to each iteration. Note that CPN is not required at the time of inference to segment target images.

4. Discussion

It is empirically shown in Sect. 3 that the proposed priors improve UDA semantic segmentation accuracy under different settings, however, how to impose semantic consistency and ecological statistics priors to general UDA tasks besides semantic segmentation remains an open problem.

Analysis of the CPN is another unsolved task. Currently, the capacity of the CPN bottleneck is chosen empirically. In order to estimate the optimal bottleneck capacity for specific tasks, quantitative measurement of the information that CPN leverages from the image is necessary, which requires future exploration.

Unsupervised domain adaptation is key for semantic segmentation, where dense annotation in real images is costly and rare, but comes automatically in rendered images. UDA is a form of transfer learning that hinges on regularities and assumptions or priors on the relationship between the distributions from which the source and target data are sampled. We introduce two assumptions, and the corresponding priors and variational renditions that are integrated into end-to-end differential learning. One is that the transformations mapping one domain to another only affect the magnitude, but not the phase, of their spectrum. This is motivated by empirical evidence that image semantics, as perceived by the human visual system, go with the phase but not the magnitude of the spectrum. The other is a prior meant to capture the ecological statistics, that are characteristics of the images induced by regularities in the scene, and therefore shared across different imaging modalities and domains. We show that the resulting priors improve performance in UDA benchmarks, and quantify their impact through ablation studies.

Acknowledgements

Research supported by ARO W911NF-17-1-0304 and ONR N00014-19-1-2066. Dong Lao is supported by KAUST through the VCC Center Competitive Funding.

References

[1] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 769–776, 2013.

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

[3] Egon Brunswik and Joe Kamiya. Ecological cue-validity of'proximity'and of other gestalt factors. *The American journal of psychology*, 66(1):20–32, 1953.

[4] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1900–1909, 2019.

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

[6] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2019.

[7] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2018.

[8] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[11] James H Elder and Richard M Goldberg. Ecological statistics of gestalt laws for the perceptual organization of contours. *Journal of Vision*, 2(4):5–5, 2002.

[12] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967, 2013.

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.

[14] Bo Geng, Dacheng Tao, and Chao Xu. Daml: Domain adaptation metric learning. *IEEE Transactions on Image Processing*, 20(10):2980–2989, 2011.

[15] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520, 2011.

[16] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2477–2486, 2019.

[17] Bruce C Hansen and Robert F Hess. Structural sparseness and spatial phase alignment in natural scenes. *JOSA A*, 24(7):1873–1885, 2007.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1989–1998, 2018.

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

[21] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2018.

[22] Dorian Kermisch. Image reconstruction from phase information only. *JOSA*, 60(1):15–17, 1970.

[23] Abhishek Kumar, Prasanna Sattigeri, Kahini Wadhawan, Leonid Karlinsky, Rogerio Feris, Bill Freeman, and Gregory Wornell. Co-regularized alignment for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 9345–9356, 2018.

[24] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019.

[25] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[26] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.

[27] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.

[28] Leon N Piotrowski and Fergus W Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3):337–346, 1982.

[29] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016.

[30] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[31] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016.

[32] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018.

[33] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 2110–2118, 2016.

[34] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:640–651, 2014.

[35] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. In *Proc. 6th International Conference on Learning Representations*, 2018.

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[37] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018.

[38] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1456–1465, 2019.

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

[40] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[41] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–534, 2018.

[42] Yanchao Yang and Stefano Soatto. Conditional prior networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 271–287, 2018.

[43] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[44] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[45] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2030, 2017.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[47] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018.