Cross-View Regularization for Domain Adaptive Panoptic Segmentation

Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu*

School of Computer Science Engineering, Nanyang Technological University

{Jiaxing.Huang, Dayan.Guan, Aoran.Xiao, Shijian.Lu}@ntu.edu.sg

Abstract

Panoptic segmentation unifies semantic segmentation and instance segmentation which has been attracting increasing attention in recent years. However, most existing research was conducted under a supervised learning setup whereas unsupervised domain adaptive panoptic segmentation which is critical in different tasks and applications is largely neglected. We design a domain adaptive panoptic segmentation network that exploits inter-style consistency and inter-task regularization for optimal domain adaptive panoptic segmentation. The inter-style consistency leverages semantic invariance across the same image of the different styles which 'fabricates' certain self-supervisions to guide the network to learn domain-invariant features. The inter-task regularization exploits the complementary nature of instance segmentation and semantic segmentation and uses it as a constraint for better feature alignment across domains. Extensive experiments over multiple domain adaptive panoptic segmentation tasks (e.g. synthetic-to-real and real-to-real) show that our proposed network achieves superior segmentation performance as compared with the state-of-the-art.

1. Introduction

Panoptic segmentation unifies semantic segmentation and instance segmentation which aims to assign a semantic class and an instance ID to each image pixel concurrently. With a large amount of annotated training images, panoptic segmentation has recently made rapid progress under a supervised setup [32, 31, 39, 52, 41, 68, 37, 25, 38]. Unfortunately, collecting large-scale training images with pixel-level annotations is prohibitively expensive and time-consuming [10, 46]. One way to mitigate this constraint is to leverage synthetic images [53] that can be automatically annotated by graphic software. However, synthetic and natural images have clear domain gaps and panoptic segmentation models trained using synthetic images usually
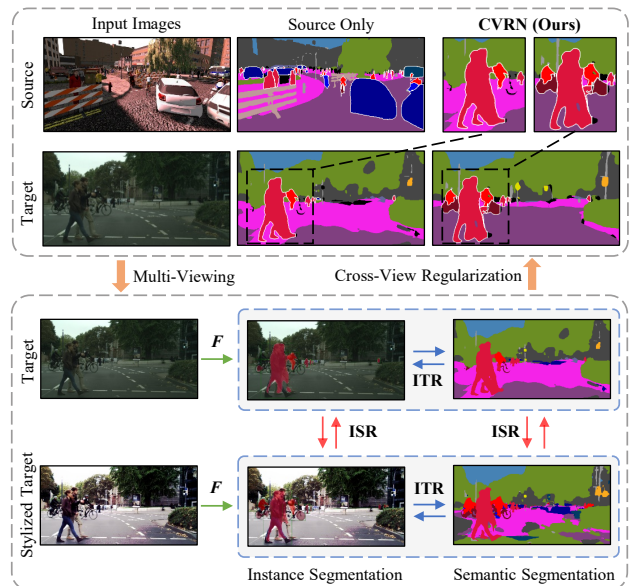

Figure 1. Our proposed cross-view regularization network (CVRN) tackles domain adaptive panoptic segmentation by exploring an inter-task regularization (ITR) and an inter-style regularization (ISR). As shown in the bottom part, ITR exploits the complementary nature of instance segmentation and semantic segmentation to regularize each other. ISR employs online image stylization to augments multiple views of the same image for regularization. Panoptic segmentations with and without our cross-view regularization are shown in the top part (the black dash lines highlight two close-up views). Best viewed in color.

experience sharp performance drop while applied to natural images, as shown in Figure 1.

One strategy that could better leverage synthetic images is domain adaptive panoptic segmentation that adopts certain unsupervised domain adaptation (UDA) techniques for adaptation from synthetic images to natural images. However, domain adaptive panoptic segmentation is largely neglected though domain adaptive semantic segmentation [76, 22, 59, 34, 24, 77, 67, 26, 70, 33] and domain adaptive instance detection/segmentation [3, 65, 72, 21, 79, 69, 17] have been investigated extensively. This could be due to a misconception that domain adaptive panoptic segmentation

---

*Corresponding author.

can be simply achieved by integrating domain adaptive semantic segmentation and domain adaptive instance segmentation. Nevertheless, semantic segmentation and instance segmentation are guided by different objectives which usually learn different feature representations from different perspectives. Learning the two tasks separately and then integrating the learned models is thus sub-optimal as it simply ignores the complementary nature of the two tasks.

We design CVRN, an innovative cross-view regularization network that addresses the challenge of domain adaptive panoptic segmentation through the regularization from different perspectives, as illustrated in Figure 1. Instead of treating semantic segmentation and instance segmentation as two independent tasks in training, we designed an inter-task regularizer that guides the two tasks to complement and regularize each other to compensate the lack of annotations (for target-domain data) in domain adaptive panoptic segmentation. This design is inspired by our observations that semantic segmentation usually performs clearly better for amorphous regions called "stuff" as compared with countable objects called "things" whereas instance segmentation usually performs in an opposite manner. In addition, we designed an inter-style regularizer that formulates the semantic consistency of the same image across different styles (in illumination, weather conditions, contrast, etc.) as supervision to regularize domain adaptation and mitigate missing annotations in target domain. The inter-style regularizer treats different styles as different views of the same image which regularizes the domain adaptation within each single image. For both inter-task and inter-style regularization, we predict pseudo labels for target-domain samples by adapting self-training ideas that have been widely adopted in many other domain adaptive computer vision tasks [81, 40, 80].

The contributions of this work can be summarized in three aspects. *First*, we designed a cross-view regularization network that addresses the challenge of domain adaptive panoptic segmentation effectively. To the best of our knowledge, this is the first work that tackles the challenging domain adaptive panoptic segmentation task. *Second*, we designed a novel inter-task regularizer that exploits the complementary nature of semantic segmentation and instance segmentation for optimal domain adaptive panoptic segmentation. In addition, we designed an inter-style regularizer that formulates consistency of the same image of different styles as supervision for better feature alignment across domains. *Third*, extensive experiments over multiple domain adaptive panoptic segmentation tasks show that our network achieves superior segmentation performance as compared with the state-of-the-art.

2. Related Works

The concept of **Panoptic segmentation** was introduced in [32] which handles the problem by fusing the predic-

tions of instance segmentation and semantic segmentation heuristically. Quite a number of relevant works have been reported since then. For example, [31] extends an instance segmentation model with a semantic segmentation branch and takes a shared feature pyramid network as backbone. [39] employs instance-level attention to transfer knowledge from an instance segmentation branch to a semantic segmentation branch. [41] presents a spatial ranking module to address occlusions between the predicted instances. [68] introduces a non-parametric panoptic head for resolving the conflicts between instance and semantic segmentation. [9] presents a bottom-up approach that employs a class-agnostic instance segmentation branch with center regression. Though panoptic segmentation has been studied extensively recently, most existing research was conducted under a supervised setup where all training data are fully annotated. We instead focus on a more challenging domain adaptive panoptic segmentation task that aims to adapt from an annotated source domain to an unlabelled target domain.

Multi-view learning trains a learner over two or more different views by incorporating confident predictions of target data iteratively [1, 11, 75, 15, 49]. In unsupervised domain adaptation, it generates pseudo labels for unlabelled target data for measuring and minimizing various task loss [44, 6, 4, 18]. In recent years, multi-view learning diversifies the learned parameters (*e.g.*, kernel weights) to enforce multiple classifiers via adversarial dropout [55, 36], classifier discrepancy maximization [57, 35], parameter diversification [73, 45], asymmetric classifier tri-training [54, 21], etc. Different from existed multi-view learning that employs multiple classifiers to create multiple views in the feature space, we exploit the complementary nature of semantic segmentation and instance segmentation and use their predictions as two views in panoptic segmentation. In addition, we adopt online image stylization to construct multiple views in the input space which enhances domain adaptation by enforcing semantic consistency across image styles.

Unsupervised domain adaptation aims to adapt a model from a labelled source domain to an unlabelled target domain. One typical domain adaptation approach leverages adversarial learning that employs a domain classifier to learn domain invariant features [13, 23, 42, 43, 64, 8, 62, 65, 16, 71, 27, 30]. Another typical approach exploits self-training that predicts pseudo labels for target-domain data and includes confident predictions in network training [81, 58, 78, 14, 18, 74]. The self-training approach has attracted increasing attention in recent years, and different strategies have been designed for predicting high-quality pseudo labels by incorporating class-balance thresholding [81], confidence-regularization [80], voting-based densification [48] or scale-invariance example exploration [61]. We adopt the self-training idea in the implementation of our proposed cross-view regularization over
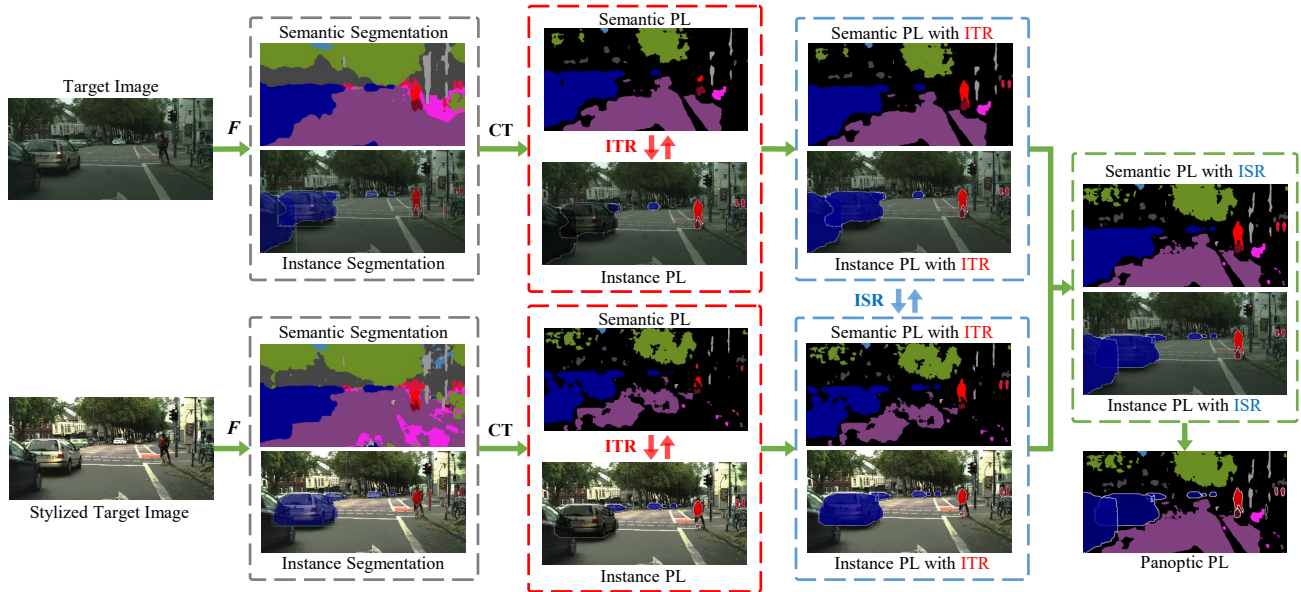
Figure 2. Overview of our proposed cross-view regularization network (CVRN): CVRN predicts multi-view co-regularized panoptic pseudo labels (PL) for learning from unlabelled target data. A target image and its stylized transformation are first fed to a panoptic segmentation model F to generate predictions (in gray boxes) and four sets of primary pseudo labels (in red boxes). With the four sets of primary pseudo labels, inter-task regularization (ITR) exploits the complementary nature of instance segmentation and semantic segmentation to co-regularize their pseudo labels, e.g. the confident predictions in instance segmentation can guide the unconfident predictions in semantic segmentation (e.g., the rider), and vice versa. The inter-task regularized pseudo labels are shown in the blue boxes. Similarly, inter-style regularization (ISR) exploits the complementary property of images of the same scene but different styles for regularization, where green box shows ISR regularized pseudo labels. Finally, the cross-view regularized instance and semantic segmentation pseudo labels are fused into panoptic segmentation pseudo labels to train unsupervised domain adaptive panoptic segmentation model with unlabeled target data.

different tasks and image styles. More specifically, we introduce multi-view learning (across tasks and image styles) into the self-training framework for unifying instance segmentation and semantic segmentation under the context of domain adaptive panoptic segmentation.

3. Method

This section presents the proposed cross-view regularization network (CVRN) as illustrated in Figure 2. Leveraging multi-task self-training (MTST), CVRN aims to adapt towards a domain-specific panoptic segmentation model by using unlabelled target images. It tackles cross-view regularization from two perspectives including inter-task regularization (ITR) and inter-style regularization (ISR), more details to be described in the ensuing subsections.

3.1. Multi-Task Self-Training

Our proposed cross-view regularization is implemented over a multi-task self-training (MTST) network. Self-training has been widely investigated to exploit unlabeled target-domain data in semi-supervised and unsupervised learning. It predicts pseudo labels for target-domain data and incorporates confident prediction for training stronger models iteratively. For the task of domain adaptive panoptic segmentation under study, we extend self-training to multi-task self-training and use it as a base to implement our proposed inter-task regularization between semantic segmentation and instance segmentation.

The problem setting is as follows. We have source-domain images X_s and the corresponding pixel-level semantic labels Y_s where C denotes the class number and N denotes the ID index of things i.e., countable objects. We also have unlabelled target-domain images X_t. The target is to learn a panoptic segmentation model F that performs well in the unlabeled target domain. With X_s and Y_s in the source domain, F can be optimized by a supervised panoptic segmentation loss $\mathcal{L}_{pan}$ that consists of a semantic segmentation loss $\mathcal{L}_{seg}$ and an instance segmentation loss $\mathcal{L}_{ins}$. During training, Y_s are converted to semantic segmentation labels Y_s^e and instance segmentation labels Y_s^d for computing $\mathcal{L}_{seg}$ and $\mathcal{L}_{ins}$ concurrently.

For x_t in the target domain, F will generate panoptic segmentation predictions which are a heuristic combination of semantic segmentation predictions p_t^e and instance segmentation predictions p_t^d. Based on generated predictions p_t^e and p_t^d, pseudo labels can be determined by a selection function $\mathcal{S}$ (confidence thresholding) that is defined as fol-

lows:

$$\mathcal{S}(p_t) = \underset{c \in C}{\arg\max}\, \mathbb{1}_{[p_t^{(c)} > \exp(-k_c)]}(p_t^{(c)}) \qquad (1)$$

where p_t refers to either p_t^e or p_t^d, $\mathbb{1}$ is a function that returns the input if the condition is true or an empty output otherwise, and k_c is the class-balanced weights [81]. The panoptic segmentation model F can be retained with target-domain images X_t and the predicted pseudo labels $\mathcal{S}(P_t^e)$ and $\mathcal{S}(P_t^d)$ via self-training.

Supervised loss: Given the panoptic segmentation model F, a source-domain image $x_s \subset X_s$ and its corresponding semantic segmentation and instance segmentation labels $\{y_s^e, y_s^d\} \subset \{Y_s^e, Y_s^d\}$, the supervised panoptic segmentation loss $\mathcal{L}_{pan}$ can be defined as follows:

$$\mathcal{L}_{pan}(x_s, y_s; F) = \mathcal{L}_{seg}(p_s^e, y_s^e) + \mathcal{L}_{ins}(p_s^d, y_s^d), \qquad (2)$$

where $\mathcal{L}_{seg}$ is cross-entropy loss, $\mathcal{L}_{ins}$ is instance segmentation loss as defined in [19], p_s^e and p_s^d represent source-domain semantic segmentation and instance segmentation predictions, respectively.

Multi-task self-training loss: Given the panoptic segmentation model F and a target image $x_t \subset X_t$, the multi-task self-training loss $\mathcal{L}_{mtst}$ can be defined by:

$$\mathcal{L}_{mtst}(x_t; F) = \mathcal{L}_{seg}(p_t^e, \mathcal{S}(p_t^e)) + \mathcal{L}_{ins}(p_t^d, \mathcal{S}(p_t^d)), \quad (3)$$

where p_t^e and p_t^d represent target-domain semantic segmentation and instance segmentation predictions, respectively.

3.2. Inter-Task Regularization

Semantic segmentation and instance segmentation learn from different perspectives which often produce complementary predictions under the context of panoptic segmentation. Under the framework of multi-task self-training, we observe that pseudo labels predicted by the two tasks often complement each other. As illustrated in Figure 2, semantic segmentation tends to predict high-quality pseudo labels (*i.e.*, diverse and accurate) for amorphous regions called "stuff" but low-quality pseudo labels (*i.e.*, sparse and inaccurate) for countable objects called "things". On the contrary, instance segmentation tends to generate high-/low-quality pseudo labels for things/stuff.

Based on this observation, we design an inter-task regularization (ITR) method that mutually regularize self-training in between the semantic segmentation and instance segmentation tasks in the target domain. Specifically, ITR employs high-certainty (*i.e.*, low entropy) pseudo label predictions of one task to regularize pseudo label predictions of the other task and vice versa. In this way, ITR is capable of predicting higher quality pseudo labels as compared with those predicted by each single task alone.

In the unannotated target domain, pseudo labels of semantic segmentation and instance segmentation (*i.e.*, $\mathcal{S}(p_t^e)$ and $\mathcal{S}(p_t^d)$) can be determined based on the predictions (*i.e.*, p_t^e and p_t^d as described in the last subsection. $\mathcal{S}(p_t^d)$ can then be regularized by $\mathcal{S}(p_t^e)$. This instance segmentation pseudo-label regularization function $\mathcal{R}_d$ that is defined as follows:

$$\mathcal{R}_d(p_t^d, p_t^e) = \mathbb{1}_{[\mathcal{E}(p_t^d) < \mathcal{E}(p_t^e)]}(\mathcal{S}(p_t^d)) +$$
$$\underset{c \in C}{\arg\max}\, \mathbb{1}_{[\mathcal{J}(p_t^d, \mathcal{S}(p_t^e))]}(p_t^{d(c)}), \qquad (4)$$

where $\mathcal{J}$ is a function to judge if the instance segmentation prediction p_t^d is highly consistent with the semantic segmentation pseudo label $\mathcal{S}(p_t^e)$ in the same image location. Semantic segmentation refines instance segmentation by the instance segmentation pseudo-label regularization function defined in Eq. (4). Semantic pseudo labels (PLs) can remove false positive from instance PLs via the first term in Eq. (4), and add false negative into instance PLs via the second term in Eq. (4). Thus, semantic PLs can refine instance PLs. Although instance PLs cover 'things' labels only, the class-agnostic 'stuff' mask can be easily got as it is a binary issue, *i.e.*, 'non-things' regions must be class-agnostic 'stuff'. Thus, more accurate Instance PL can improve learning both 'things' and class-agnostic 'stuff' by training on both positive/negative region-of-interests (ROIs) [19] selected from instance PLs. For the function $\mathcal{J}$ in Eq. (4), it re-identifies one prediction as an instance PLs, if its intersection with semantic PLs over itself is larger than the threshold (*i.e.*, 0.5).

Similarly, $\mathcal{S}(p_t^e)$ can be regularized by $\mathcal{S}(p_t^d)$. This semantic segmentation pseudo-label regularization function $\mathcal{R}_e$ is defined by:

$$\mathcal{R}_e(p_t^e, p_t^d) = \mathbb{1}_{[\mathcal{E}(p_t^e) < \mathcal{E}(p_t^d)]}(\mathcal{S}(p_t^e)) +$$
$$\mathcal{T}(\mathbb{1}_{[\mathcal{E}(p_t^d) < \mathcal{E}(p_t^e)]}(\mathcal{S}(p_t^d))), \qquad (5)$$

where $\mathcal{E}$ is the entropy function as defined in [60]. $\mathcal{T}$ is a label transformation function from instance segmentation to semantic segmentation, *i.e.*, ignoring ID indexes of instances of the same category.

Inter-task regularization loss: Given the panoptic segmentation model F and a target-domain image $x_t \subset X_t$, the inter-task regularization loss $\mathcal{L}_{itr}$ can be defined by:

$$\mathcal{L}_{itr}(x_t; F) = \mathcal{L}_{seg}(p_t^e, \mathcal{R}_e(p_t^e, p_t^d)) +$$
$$\mathcal{L}_{ins}(p_t^d, \mathcal{R}_d((p_t^d, p_t^e)), \qquad (6)$$

where p_t^e and p_t^d represent target-domain semantic segmentation and instance segmentation predictions, respectively.

3.3. Inter-Style Regularization

Besides inter-task regularization, we also design an inter-style regularization (ISR) method that further improves the

quality of pseudo labels by fusing confident (pseudo label) predictions of images of the same scene but different styles. The idea is that images of the same scene should share perfectly the same pixel-level semantics when they are captured under different conditions with different image styles (e.g. in different illumination, weather, etc.). A pixel with more confident pseudo-level prediction in one image view with one specific style can therefore be exploited to regularize the less confident prediction of the corresponding pixel in another image view with a different style.

We implement ISR based on online image stylization that generates images with new styles/views with histogram matching [51]. Specifically, the online image stylization first transfers a training image x_t to $\widetilde{x}_t$ based on another randomly selected target-domain image. It then forwards x_t and its style transformation $\widetilde{x}_t$ to the panoptic segmentation model F to predict semantic segmentation (i.e., p_t^e and $\widetilde{p}_t^e$) and instance segmentation (i.e., p_t^d and $\widetilde{p}_t^d$). Finally, we generate the pseudo labels by an inter-style pseudo-label unification function that is defined as follows:

$$\mathcal{U}(p_t, \widetilde{p}_t) = \mathbb{1}_{[\mathcal{E}(p_t) < \mathcal{E}(\widetilde{p}_t)]}(\mathcal{S}(p_t)) + \\ \mathbb{1}_{[\mathcal{E}(\widetilde{p}_t) < \mathcal{E}(p_t))]}(\mathcal{S}(\widetilde{p}_t)). \quad (7)$$

ISR thus enforces semantic consistency across images of the same contents but different styles by retraining the model F with unified pseudo labels in different style views.

Inter-style regularization loss: Given the panoptic segmentation model F and a target-domain image $x_t \subset X_t$ and its style transformation $\widetilde{x}_t$, the inter-style regularization loss $\mathcal{L}_{isr}$ can be formulated as follows:

$$\mathcal{L}_{isr}(x_t, \widetilde{x}_t; F) = \mathcal{L}_{seg}(\widetilde{p}_t^e, \mathcal{U}_e(p_t, \widetilde{p}_t)) + \\ \mathcal{L}_{ins}(\widetilde{p}_t^d, \mathcal{U}_d(p_t, \widetilde{p}_t)) + \\ \mathcal{L}_{seg}(p_t^e, \mathcal{U}_e(p_t, \widetilde{p}_t)) + \\ \mathcal{L}_{ins}(p_t^d, \mathcal{U}_d(p_t, \widetilde{p}_t)), \quad (8)$$

where $\widetilde{p}_t^e$ and $\widetilde{p}_t^d$ represent semantic segmentation and instance segmentation predictions generated from style-transformed images, respectively.

Training objective: The overall objective function of the proposed cross-view regularization network (CVRN) can thus be formulated by summing up the four training losses as follows:

$$\mathcal{L}_{cvrn} = \mathcal{L}_{pan} + \lambda_{mt}\mathcal{L}_{mtst} + \lambda_{it}\mathcal{L}_{itr} + \lambda_{is}\mathcal{L}_{isr}, \quad (9)$$

where λ_{mt}, λ_{it} and λ_{is} are the balancing weights.

4. Experiment

4.1. Datasets and Evaluation Metrics

We evaluated our proposed cross-view regularization technique over three widely used datasets, i.e., **SYNTHIA**

SYNTHIA → Cityscapes							
Method	L_{sup}	L_{mtst}	L_{itr}	L_{isr}	mSQ	mRQ	PQ
Source only	✓				59.0	27.8	20.1
MTST	✓	✓			59.5	32.9	24.8
MTST + ITR	✓	✓	✓		61.1	36.4	28.0
MTST + ISR	✓	✓		✓	62.3	36.9	28.9
CVRN	✓	✓	✓	✓	**66.6**	**40.9**	**32.1**

Table 1. Ablation study of CVRN over domain adaptive panoptic segmentation task SYNTHIA → Cityscapes: The proposed inter-task regularization (ITR) and inter-style regularization (ISR) both outperforms the base network MTST (multi-task self-training) greatly. ITR and ISR complement with each other clearly.

[53], **Cityscapes** [10] and **Mapillary Vistas** [46]. Please refer to the appendix for dataset details. We evaluate panoptic segmentation by three widely used metrics including semantic quality (SQ), recognition quality (RQ), and panoptic quality (PQ) [32].

4.2. Implementation Details

We adopted PSN [32] as the panoptic segmentation architecture that consists of a semantic segmentation branch (i.e., *Deeplab-V2* [5]) and an instance segmentation branch (i.e., *Mask R-CNN* [19]). All the networks in the experiments use ResNet-101 [20] pre-trained on ImageNet [12] as backbone. We implemented CVRN by using PyTorch [50] and trained it with a single NVIDIA 2080TI GPU with 11GB memory. The networks are trained with a standard Stochastic Gradient Descent optimizer [2] with learning rate 2.5×10^{-4}, momentum 0.9, and weight decay 10^{-4}.

4.3. Ablation studies

We performed extensive ablation studies to investigate how our designs contribute to domain adaptive panoptic segmentation. The ablation studies were conducted over the task SYNTHIA → Cityscapes as shown in Table 1. It can be seen that *MTST* outperforms *Source only* clearly as it exploits target-domain knowledge via self-training. In addition, *MTST+ITR* outperforms *MTST* by a clear margin, demonstrating the effectiveness of the proposed inter-task regularization in between instance segmentation and semantic segmentation. At the same time, *MTST+ISR* outperforms *MTST* significantly as well which is largely attributed to the proposed inter-style regularization design that aggregates rich and complementary knowledge from different image views. Further, *CVRN* produces the best panoptic segmentation, which shows that the proposed inter-task regularization and inter-style regularization are orthogonal and complementary to each other.

SYNTHIA → Cityscapes Panoptic Segmentation																		
Methods	road	side.	buil.	wall	fence	pole	light	sign	vege.	sky	pers.	rider	car	bus	mot.	bike	mSQ	mRQ	mPQ
Source only	32.3	5.1	58.5	0.9	0.0	0.9	0.0	4.6	61.7	61.3	27.6	9.5	32.8	22.6	1.0	2.7	59.0	27.8	20.1
FDA [71]	79.0	22.0	61.8	1.1	0.0	5.6	5.5	9.5	51.6	70.7	23.4	16.3	**34.1**	31.0	5.2	8.8	65.0	35.5	26.6
CRST [80]	75.4	19.0	70.8	1.4	0.0	7.3	0.0	5.2	74.1	69.2	23.7	**19.9**	33.4	26.6	2.4	4.8	60.3	35.6	27.1
SVMin [16]	82.6	39.5	76.5	5.9	0.0	34.0	7.5	11.5	81.1	82.8	56.5	12.2	80.3	35.7	18.9	20.3	64.8	35.9	27.8
AdvEnt [65]	**87.1**	32.4	69.7	1.1	0.0	3.8	0.7	2.3	71.7	72.0	**28.2**	17.7	31.0	21.1	6.3	4.9	65.6	36.3	28.1
CrCDA [28]	84.4	21.9	69.6	3.1	0.0	3.5	5.3	6.0	75.3	68.7	34.4	17.8	32.8	27.6	2.3	4.9	64.8	37.1	28.6
CVRN (Ours)	86.6	**33.8**	**74.6**	**3.4**	0.0	**10.0**	5.7	**13.5**	80.3	76.3	26.0	18.0	**34.1**	37.4	7.3	6.2	**66.6**	**40.9**	**32.1**
Oracle	96.0	60.7	83.0	17.9	19.0	16.9	15.6	38.1	83.4	83.8	44.0	41.3	57.9	39.0	33.3	32.8	75.5	60.2	47.7

SYNTHIA → Mapillary Panoptic Segmentation																		
Methods	road	side.	buil.	wall	fence	pole	light	sign	vege.	sky	pers.	rider	car	bus	mot.	bike	mSQ	mRQ	mPQ
Source only	22.1	5.5	34.5	0.2	0.0	2.8	0.0	1.5	40.3	**79.5**	18.9	9.2	35.6	3.9	0.9	0.5	59.4	21.3	16.0
AdvEnt [65]	27.7	6.1	28.1	0.3	0.0	3.4	1.6	5.2	48.1	66.5	28.4	13.4	40.5	14.6	5.2	3.3	63.6	24.7	18.3
CRST [80]	36.0	6.4	29.1	0.2	0.0	2.8	0.5	4.6	47.7	68.9	28.3	13.0	42.4	13.6	5.1	2.0	63.9	25.2	18.8
FDA [71]	**44.1**	7.1	26.6	1.3	0.0	3.2	0.2	5.5	45.2	61.3	30.1	13.9	39.4	12.1	**8.5**	7.0	63.8	26.1	19.1
CrCDA [28]	50.5	7.6	32.5	0.7	0.0	2.9	1.8	1.9	44.1	66.8	28.3	13.0	42.4	13.6	5.1	2.0	64.4	26.1	19.6
SVMin [17]	21.9	5.6	35.0	1.0	0.0	3.5	0.3	6.1	47.9	79.9	29.2	15.5	41.8	19.8	5.5	5.7	64.5	26.4	19.9
CVRN (Ours)	33.4	**7.4**	**32.9**	**1.6**	0.0	**4.3**	0.4	**6.5**	50.8	76.8	30.6	15.2	44.8	18.8	7.9	**9.5**	**65.3**	**28.1**	**21.3**
Oracle	80.4	37.6	57.0	13.4	20.3	15.7	24.1	33.5	66.2	90.5	30.1	15.7	54.9	28.7	10.8	9.3	73.5	46.9	36.8

Cityscapes → Mapillary Panoptic Segmentation																		
Methods	road	side.	buil.	wall	fence	pole	light	sign	vege.	sky	pers.	rider	car	bus	mot.	bike	mSQ	mRQ	mPQ
Source only	75.6	15.8	40.4	5.1	8.9	3.2	2.1	13.4	55.2	81.8	24.4	15.1	51.4	4.4	12.8	13.6	71.2	34.3	26.4
CRST [80]	77.0	22.6	40.2	7.8	10.5	5.5	11.3	21.8	56.5	77.6	29.4	18.4	56.0	27.7	11.9	18.4	72.4	39.9	30.8
FDA [71]	74.3	**23.4**	42.3	9.6	11.2	6.4	**15.4**	23.5	60.4	78.5	33.9	19.9	52.9	8.4	**17.5**	16.0	72.3	40.3	30.9
SVMin [17]	70.5	18.9	41.4	5.5	9.6	4.3	2.9	15.7	56.4	79.3	38.4	25.5	57.1	29.2	19.3	19.6	72.6	39.8	30.9
CrCDA [28]	76.8	22.2	43.3	6.7	11.4	6.9	7.7	19.8	60.1	79.2	29.4	18.4	56.0	27.7	11.9	18.4	72.9	39.8	31.0
AdvEnt [65]	76.2	20.5	42.6	6.8	9.4	4.6	12.7	24.1	59.9	83.1	34.1	**22.9**	54.1	16.0	13.5	18.6	72.7	40.3	31.2
CVRN (Ours)	**77.3**	21.0	**47.8**	**10.5**	**13.4**	7.5	14.1	**25.1**	62.1	**86.4**	37.7	20.4	**55.0**	21.7	14.3	**21.4**	**73.8**	**42.8**	**33.5**
Oracle	80.4	37.6	57.0	13.4	20.3	15.7	24.1	33.5	66.2	90.5	30.1	15.7	54.9	28.7	10.8	9.3	73.5	46.9	36.8

Table 2. Comparing CVRN with state-of-the-art domain adaptive panoptic segmentation: CVRN outperforms the state-of-the-art across three tasks. PQ is computed for each category. Mean SQ (mSQ), mean RQ (mSQ), mean PQ (mPQ) are computed over all categories.

4.4. Comparison with state-of-art

Since there is little domain adaptive panoptic segmentation work, we compare our method with a number of domain adaptation methods [65, 80, 28, 16, 71] that achieved state-of-the-art performance in both semantic segmentation and instance detection/segmentation. These methods can be easily applied to the panoptic segmentation task by heuristically combining predictions from semantic and instance segmentation [32] The comparisons were conducted over three domain adaptive panoptic segmentation tasks as shown in Table 2. We can observe that CVRN achieves the best panoptic segmentation across all three evaluation metrics for the task "SYNTHIA → Cityscapes" (synthetic-to-real). The superior panoptic segmentation is largely attributed to the proposed cross-view regularization that

guides to exploit more confident samples which leads to more true positives. For another two tasks "SYNTHIA → Mapillary Vistas" (synthetic-to-real) and "SYNTHIA → Mapillary Vistas" (real-to-real), CVRN outperforms the state-of-the-art consistently due to similar reasons.

We also compare CVRN with state-of-the-art domain adaptive semantic segmentation methods [62, 63, 45, 65, 47, 29, 28, 80, 40, 66, 71] and domain adaptive instance detection/segmentation methods [7, 65, 56, 80, 16, 71] that are dedicated to the two specific tasks, respectively. We compare with the two categories of methods because there is little domain adaptive panoptic segmentation work but panoptic segmentation actually consists of semantic segmentation and instance detection/segmentation. Tables 3 and 4 show experimetal results over the task "SYNTHIA → Cityscapes". As Table 3 shows, CVRN outperforms state-
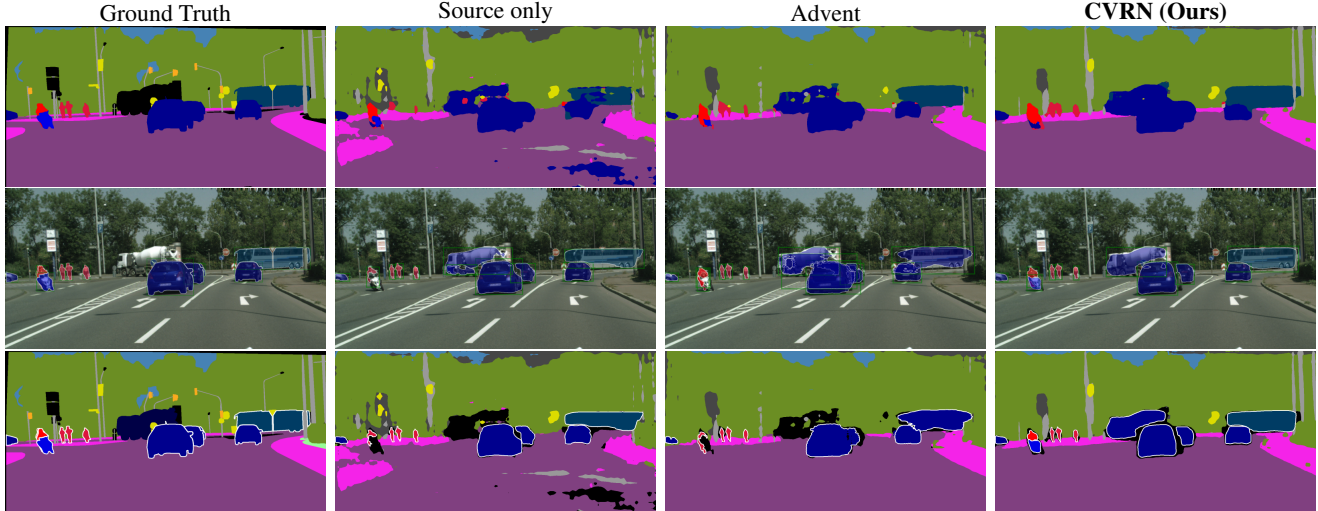
Figure 3. Qualitative comparison of CVRN with *Source only* (with no adaptation) and state-of-the-art method *AdvEnt* [65] for domain adaptive semantic segmentation, instance segmentation and panoptic segmentation as shown in Rows 1-3, respectively. The comparison was conducted over the task "SYNTHIA → Cityscapes". Best viewed in color.

SYNTHIA → Cityscapes Semantic Segmentation																	
Methods	road	side.	buil.	wall	fence	pole	light	sign	vege.	sky	pers.	rider	car	bus	mot.	bike	mIoU	mIoU*
Source only	56.8	26.2	68.6	6.3	0.4	23.9	3.6	11.9	76.5	75.7	37.4	14.4	50.5	12.7	8.2	33.0	31.6	36.6
PatAlign [63]	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
AdaptSeg [62]	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
CLAN [45]	81.3	37.0	80.1	-	-	-	16.1	13.7	78.2	81.5	53.4	21.2	73.0	32.9	22.6	30.7	-	47.8
AdvEnt [65]	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2	48.0
IDA [47]	84.3	37.7	79.5	5.3	0.4	24.9	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	41.7	48.9
TIR [29]	**92.6**	**53.2**	79.2	-	-	-	1.6	7.5	78.6	84.4	52.6	20.0	82.1	34.8	14.6	39.4	-	49.3
CrCDA [28]	86.2	44.9	79.5	8.3	0.7	27.8	9.4	11.8	78.6	**86.5**	57.2	26.1	76.8	39.9	21.5	32.1	42.9	50.0
CRST [80]	67.7	32.2	73.9	10.7	**1.6**	**37.4**	22.2	**31.2**	80.8	80.5	60.8	29.1	**82.8**	25.0	19.4	45.3	43.8	50.1
BDL [40]	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
SIM [66]	83.0	44.0	80.3	-	-	-	17.1	15.8	80.5	81.8	59.9	**33.1**	70.2	37.3	28.5	45.8	-	52.1
FDA [71]	79.3	35.0	73.2	9.1	0.3	33.5	19.9	24.0	61.7	82.6	**61.4**	31.1	83.9	40.8	**38.4**	51.1	45.3	52.5
CVRN (Ours)	87.5	45.5	**83.5**	**12.2**	0.5	**37.4**	**25.1**	29.6	**85.9**	86.0	61.1	25.9	80.9	34.7	33.8	**53.5**	**49.0**	**56.6**
Oracle	96.5	76.7	88.7	47.0	46.4	44.6	43.9	58.9	89.8	92.9	65.8	49.8	89.8	77.6	48.1	60.6	67.3	72.2

Table 3. Comparisons of CVRN with state-of-the-art domain adaptive semantic segmentation: CVRN outperforms the state-of-the-art by large margins. IoU is evaluated for each of 16 pixel categories, *mIoU** is evaluated for 13 pixel classes following [62, 45, 40, 29, 66].

of-the-art domain adaptive semantic segmentation methods by a large margin (over 3.7 in mIoU). For domain adaptive instance detection and segmentation task, CVRN outperforms the state-of-the-art with a mAP of 2.7 for instance detection and a mAP of 5.4 for instance segmentation, respectively, as shown in Table 4.

The qualitative segmentation is well aligned with the quantitative experimental results as illustrated in Figure 3. We compared CVRN with the baseline *Source only* and state-of-the-art method *Advent* [65] qualitatively over the domain adaptation task "SYNTHIA → Cityscapes". As Figure 3 shows, CVRN identifies and detects more correct "thing" instances and segments more accurate "stuff"

regions across semantic segmentation, instance detection/segmentation and panoptic segmentation consistently. The superior segmentation performance is largely attributed to the proposed cross-view regularization that guides to learn rich and complementary semantic information from multiple views.

4.5. Discussion

We studied whether the proposed CVRN is complementary with state-of-the-art UDA methods for the task of domain adaptive panoptic segmentation. In the experiments, we incorporated our designed cross-view regularizers into each studied domain adaptation method and

SYNTHIA → Cityscapes Instance Detection

Methods	pers.	rider	car	bus	mot.	bike	mAP
Source only	42.9	23.4	38.4	20.4	3.5	17.1	24.3
DA [7]	43.2	44.0	39.9	22.8	7.6	23.6	30.2
SWDA [56]	43.9	35.4	42.1	28.5	11.3	26.4	31.3
AdvEnt [65]	43.9	39.6	44.0	22.2	11.2	26.8	31.3
SVMin [16]	42.7	37.9	44.7	28.1	7.3	27.2	31.3
UaDAN [17]	45.6	34.6	44.9	27.5	12.3	28.0	32.1
CRST [80]	40.3	**48.6**	34.8	29.7	10.7	28.6	32.1
FDA [71]	42.6	43.9	42.5	24.9	10.6	**30.1**	32.4
CRDA [69]	**45.5**	37.9	**45.6**	28.2	9.1	29.1	32.6
CVRN (Ours)	43.5	41.9	45.2	**39.3**	**14.8**	27.1	**35.3**
Oracle	60.1	66.2	76.4	52.4	49.8	57.4	60.4

SYNTHIA → Cityscapes Instance Segmentation

Methods	pers.	rider	car	bus	mot.	bike	mAP
Source only	30.4	9.2	32.5	20.4	3.1	1.3	16.2
DA [7]	30.4	**25.9**	28.4	13.6	4.3	1.0	17.3
SWDA [56]	30.3	16.8	31.8	25.6	4.1	0.7	18.2
AdvEnt [65]	30.8	19.9	32.0	19.6	7.0	4.2	18.9
SVMin [16]	32.6	15.1	36.6	27.3	4.1	2.2	19.6
FDA [71]	27.7	24.6	33.9	22.5	5.5	5.4	19.9
CRDA [69]	33.9	16.7	34.3	27.7	3.8	3.0	19.9
UaDAN [17]	34.0	19.4	35.9	26.9	1.5	3.3	20.2
CRST [80]	26.4	20.3	31.5	27.9	8.4	8.6	20.5
CVRN (Ours)	**34.4**	25.3	**38.7**	**38.1**	**10.1**	**8.7**	**25.9**
Oracle	51.2	54.9	65.8	50.9	33.4	38.0	49.0

Table 4. Quantitative comparison of CVRN with state-of-the-art domain adaptive instance detection and instance segmentation methods: Class-wise AP (%) and mAP (%) over all classes are computed with an IoU threshold of 0.5 following [7, 56, 69].

train panoptic segmentation models over the task SYN-THIA → Cityscapes. Table 5 shows experimental results. We can observe that the incorporation of CVRN improves segmentation consistently across all studied UDA methods [65, 80, 71] and evaluation metrics. For metric mPQ that is more relevant to panoptic segmentation, the performance gains are above 3.3 for all three UDA methods. Such experimental results show that CVRN is complementary to existing UDA methods that work by image translation (*e.g.*, FDA [71]), adversarial learning (*e.g.*, AdvEnt [65]) and single-view self-training (*e.g.*, CRST [80]).

We also investigate whether CVRN performs stably with different panoptic segmentation architectures. We evaluated over two network architectures PFPN [31] and UPSNet [68] that have been widely adopted in supervised panoptic segmentation. For each architecture, we trained 3 models including *Source only*, *MTST* and *CVRN*, where *MTST* is the base network that our proposed cross-view regularization is built upon. Table 6 shown experimental results. It can be

SYNTHIA → Cityscapes Panoptic Segmentation

Method	Base mSQ mRQ mPQ	+ CVRN mSQ mRQ mPQ	Gain mSQ mRQ mPQ
FDA [71]	65.0 35.5 26.6	66.0 39.5 30.9	+1.0 +4.0 +3.3
AdvEnt [65]	65.6 36.3 28.1	66.6 40.5 32.0	+1.0 +4.2 +3.9
CRST [80]	60.3 35.6 27.1	66.6 41.5 32.4	+6.3 +5.9 +5.3

Table 5. CVRN complements with existing UDA methods: *CVRN* can be easily incorporated into state-of-the-art representative UDA methods [65, 80, 71] with consistent performance improvement (evaluated over domain adaptive panoptic segmentation task SYN-THIA → Cityscapes).

PFPN		UPSNet	
Method	mSQ mRQ mPQ	Method	mSQ mRQ mPQ
Source only	54.3 29.0 21.5	Source only	58.7 31.5 23.3
MTST	64.8 36.1 26.8	MTST	66.2 36.1 27.5
CVRN	**66.4 40.0 31.4**	**CVRN**	**68.2 43.4 34.0**

Table 6. CVRN works with different panoptic segmentation architectures well: *CVRN* can work with different panoptic segmentation architectures (e.g. PFPN [31] and UPSNet [68]) with consistent performance improvement as compared with *Source only* and the baseline network *MTST* (evaluated on domain adaptive panoptic segmentation task SYNTHIA → Cityscapes).

seen that *CVRN* outperforms *Source only* and *MTST* with a large margin (over 4.6 in mPQ) in both panoptic segmentation architectures. Such experimental results show that our proposed cross-view regularization can work well with different panoptic segmentation architectures.

5. Conclusion

This paper presents a cross-view regularization network that tackles domain adaptive panoptic segmentation by exploring an inter-task regularization (ITR) and an inter-style regularization (ISR). Specifically, ITR exploits the complementary nature of instance segmentation and semantic segmentation to regularize their self-training. ISR employs online image stylization to augments multiple views of the same image for self-training regularization. The proposed method have been evaluated extensively over multiple domain adaptive panoptic segmentation settings and experiments demonstrate its superior performance as compared with state-of-the-art.

References

[1] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998. 2

[2] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 5

[3] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019. 1

[4] Yanpeng Cao, Dayan Guan, Weilin Huang, Jiangxin Yang, Yanlong Cao, and Yu Qiao. Pedestrian detection with unsupervised multispectral feature learning using deep neural networks. *information fusion*, 46:206–217, 2019. 2

[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. 5

[6] Minmin Chen, Kilian Q Weinberger, and John Blitzer. Co-training for domain adaptation. In *Advances in neural information processing systems*, pages 2456–2464, 2011. 2

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 6, 8

[8] 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. 2

[9] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12485, 2020. 2

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 5

[11] Sanjoy Dasgupta, Michael L Littman, and David A McAllester. Pac generalization bounds for co-training. In *Advances in neural information processing systems*, pages 375–382, 2002. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014. 2

[14] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073. IEEE, 2012. 2

[15] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019. 2

[16] Dayan Guan, Jiaxing Huang, Shijian Lu, and Aoran Xiao. Scale variance minimization for unsupervised domain adaptation in image segmentation. *Pattern Recognition*, 112:107764, 2021. 2, 6, 8

[17] Dayan Guan, Jiaxing Huang, Shijian Lu, Aoran Xiao, and Yanpeng Cao. Uncertainty-aware unsupervised domain adaptation in object detection. *arXiv preprint arXiv:2103.00236*, 2021. 1, 6, 8

[18] Dayan Guan, Xing Luo, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, George Vosselman, and Michael Ying Yang. Unsupervised domain adaptation for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 2

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4, 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[21] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In *European conference on computer vision*, 2020. 1, 2

[22] 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. 1

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

[24] 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. 1

[25] Rui Hou, Jie Li, Arjun Bhargava, Allan Raventos, Vitor Guizilini, Chao Fang, Jerome Lynch, and Adrien Gaidon. Real-time panoptic segmentation from dense detections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[26] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. *arXiv preprint arXiv:2103.02370*, 2021. 1

[27] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Mlan: Multi-level adversarial network for domain adaptive semantic segmentation. *arXiv preprint arXiv:2103.12991*, 2021. 2

[28] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In *European Conference on Computer Vision*, pages 705–722. Springer, 2020. 6, 7

[29] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12975–12984, 2020. 6, 7

[30] Minsu Kim, Sunghun Joung, Seungryong Kim, JungIn Park, Ig-Jae Kim, and Kwanghoon Sohn. Cross-domain grouping and alignment for domain adaptive semantic segmentation. *arXiv preprint arXiv:2012.08226*, 2020. 2

[31] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019. 1, 2, 8

[32] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 1, 2, 5, 6

[33] Marvin Klingner, Jan-Aike Termöhlen, Jacob Ritterbach, and Tim Fingscheidt. Unsupervised batchnorm adaptation (ubna): A domain adaptation method for semantic segmentation without using source domain representations. *arXiv preprint arXiv:2011.08502*, 2020. 1

[34] Jogendra Nath Kundu, Nishank Lakkakula, and R. Venkatesh Babu. Um-adapt: Unsupervised multitask adaptation using adversarial cross-task distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[35] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019. 2

[36] Seungmin Lee, Dongwan Kim, Namil Kim, and Seong-Gyun Jeong. Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[37] Qizhu Li, Xiaojuan Qi, and Philip H.S. Torr. Unifying training and inference for panoptic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[38] Qizhu Li, Xiaojuan Qi, and Philip HS Torr. Unifying training and inference for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13328, 2020. 1

[39] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7026–7035, 2019. 1, 2

[40] 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. 2, 6, 7

[41] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6181, 2019. 1, 2

[42] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015. 2

[43] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016. 2

[44] Ping Luo, Fuzhen Zhuang, Hui Xiong, Yuhong Xiong, and Qing He. Transfer learning from multiple source domains via consensus regularization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 103–112, 2008. 2

[45] 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. 2, 6, 7

[46] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999, 2017. 1, 5

[47] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020. 6, 7

[48] Inkyu Shin Sanghyun Woo Fei Pan and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. *European conference on computer vision*, 2020. 2

[49] Sungrae Park, JunKeon Park, Su-Jin Shin, and Il-Chul Moon. Adversarial dropout for supervised and semi-supervised learning. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 3917–3924. AAAI press, 2018. 2

[50] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[51] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987. 5

[52] Lorenzo Porzi, Samuel Rota Bulo, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8277–8286, 2019. 1

[53] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of

urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 1, 5

[54] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 2988–2997, 2017. 2

[55] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. In *International Conference on Learning Representations*, 2018. 2

[56] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019. 6, 8

[57] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 2

[58] Fatemeh Sadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, and Jose M Alvarez. Effective use of synthetic data for urban scene semantic segmentation. In *European Conference on Computer Vision*, pages 86–103. Springer, 2018. 2

[59] 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. 1

[60] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948. 4

[61] M Naseer Subhani and Mohsen Ali. Learning from scale-invariant examples for domain adaptation in semantic segmentation. *arXiv preprint arXiv:2007.14449*, 2020. 2

[62] 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. 2, 6, 7

[63] 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. 6, 7

[64] 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. 2

[65] 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. 1, 2, 6, 7, 8

[66] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12635–12644, 2020. 6, 7

[67] 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. 1

[68] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8818–8826, 2019. 1, 2, 8

[69] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020. 1, 8

[70] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. In *European Conference on Computer Vision*, pages 480–498. Springer, 2020. 1

[71] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2, 6, 7, 8

[72] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, and Shijian Lu. Meta-detr: Few-shot object detection via unified image-level meta-learning, 2021. 1

[73] Junting Zhang, Chen Liang, and C-C Jay Kuo. A fully convolutional tri-branch network (fctn) for domain adaptation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3001–3005. IEEE, 2018. 2

[74] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–607, 2019. 2

[75] Zhi-Hua Zhou and Ming Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541, 2005. 2

[76] Hongyuan Zhu, Fanman Meng, Jianfei Cai, and Shijian Lu. Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. *Journal of Visual Communication and Image Representation*, 34:12–27, 2016. 1

[77] Xinge Zhu, Hui Zhou, Ceyuan Yang, Jianping Shi, and Dahua Lin. Penalizing top performers: Conservative loss for semantic segmentation adaptation. In *Proceedings of the Eu-*

ropean Conference on Computer Vision (ECCV), pages 568–583, 2018. 1

[78] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005. 2

[79] Chenfan Zhuang, Xintong Han, Weilin Huang, and Matthew R Scott. ifan: Image-instance full alignment networks for adaptive object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 1

[80] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5982–5991, 2019. 2, 6, 7, 8

[81] 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. 2, 4