

Supplementary Material for TridentAdapt: Learning Domain-invariance via Source-Target Confrontation and Self-induced Cross-domain Augmentation

Fengyi Shen^{1, 2}

fengyi.shen@tum.de

Akhil Gurram^{* 2, 3}

akhil.gurram@hauwei.com

Ahmet Faruk Tuna^{*† 2}

a.tuna@gmx.at

Onay Urfalioglu^{† 2}

onay.urfalioglu@gmail.com

Alois Knoll¹

knoll@in.tum.de

¹ Department of Informatics,

Technical University of Munich,
(TUM), 85748 Garching, Germany.

² Huawei Munich Research Center,
(MRC), 80992 Munich, Germany.

³ Department of Computer Science,
Universitat Autònoma de Barcelona,
(UAB), 08193 Bellaterra, Spain.

Abstract

In this Supplementary, we provide extensive experiments for evaluation as well as additional results obtained from our intermediate and final models of TridentAdapt or its variants.

1 Architecture Analysis

We compare TridentAdapt architecture with related previous methods which are based on image domain transfer and explain the major differences which lead our approach to outperform others. Fig. 1 illustrates how different methods behave when receiving an image from source domain. In Fig. 1 (a), CyCADA [2] employs a separate source domain segmentation network to compare prediction consistency before and after domain transfer. However, this semantic consistency loss could cause the source-to-target translation to be semantically biased towards source domain since E_{seg_s} is trained purely on source images. On the contrary, TridentAdapt computes semantic consistency loss by minimizing feature distances on the shared and adapted segmentation encoder, which avoids domain-specific bias during image translation. Moreover, instead of excluding source input image from the segmentation network in CyCADA [2] and BDL [3], TridentAdapt computes segmentation losses for both source input and source-to-target translated image, thus compensating the lost semantic information due to the adversarial disturbances during image translation.

Different from CyCADA [2] and BDL [3], for each source input, TridentAdapt introduces

* Equal contribution

† The majority of the work was done while working at Huawei MRC

© 2021. The copyright of this document resides with its authors.

It may be distributed unchanged freely in print or electronic forms.

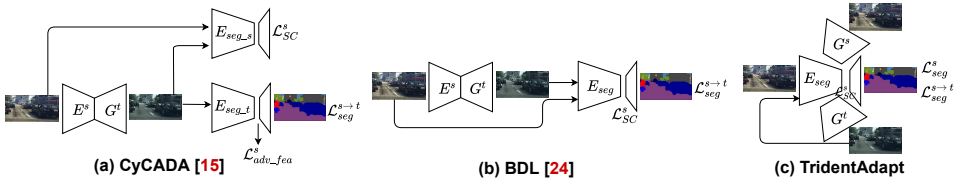


Figure 1: Architectural design comparison between related approaches and TridentAdapt.

Method	road	sdwk	blndg	wall	fence	pole	light	sign	veg	trm	sky	psn	rider	car	truck	bus	train	mcycl	beycl	mIoU
Source-only [15]	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
TridentAdapt (stage 1)	90.2	44.7	81.2	34.3	28.6	30.3	35.0	25.1	84.2	31.3	82.9	59.2	28.7	85.2	35.5	33.1	16.3	30.1	27.2	46.5
TridentAdapt (stage 2)	91.3	51.5	86.4	38.8	36.4	42.3	45.4	42.0	86.6	36.4	84.3	67.7	42.8	89.1	41.7	38.2	20.6	40.3	30.7	53.3

Table 1: GTA5-to-Cityscapes adaptation results. we provide class-wise performances of TridentAdapt in ‘stage 1’ and ‘stage 2’ which are trained with ResNet-101 [15] and Deeplab-V2 [24] based models.

source-target confrontation by allowing a simultaneous intro-domain reconstruction and cross-domain image transfer step, which forces domain-invariance on feature level. Most importantly, our architecture design brings two additional benefits in comparison with CyCADA [15] and BDL [24]. First, looping the cross-domain augmentation that is self-induced back to the encoder (see Fig. 1 (c)) ensures more coherent interaction between domain transfer module and segmentation encoder, resulting in better adapted segmentation model than CyCADA [15] and BDL [24] which build separate models for translation and segmentation; Second, learning image recovery together with segmentation from a shared encoder not only helps to obtain higher segmentation accuracy on class boundary pixels, but also yields better quality for image translation as rich semantic knowledge is incorporated. The points mentioned above are experimentally verified in the Sect. 5 & Sect. 6.

2 Pseudo-label quality

High quality pseudo-label generation triggers more efficient training on target domain. During ‘stage 1’ training of our proposed TridentAdapt framework, although adversarial learning takes the leading role, the model is still able to benefit a lot from source knowledge because of our proposed source-target confrontation concept and self-induced cross-domain data augmentation. In Fig. 2, we show through examples that our ‘stage 1’ training helps generate reliable pseudo-labels with few misclassifications, which provides strong support to ‘stage 2’ training.

3 Detailed information for two-stage training

Table 1 and Table 2 include the class-wise IoU and overall mIoU calculated for ‘stage 1’ and ‘stage 2’ training of TridentAdapt on GTA5-to-Cityscapes adaptation and SYNTHIA-to-Cityscapes adaptation, respectively. From the tables we can observe that our proposed source-target confrontation concept as well as self-induced cross-domain data augmentation brings strong constraints to enforce our shared encoder and the segmentation network to

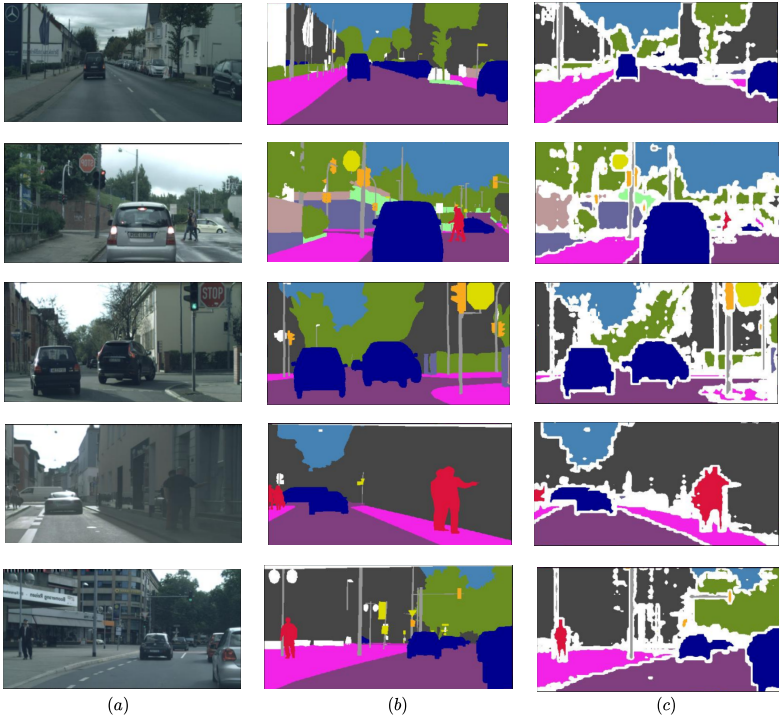


Figure 2: Qualitative results of pseudo-labels obtained from our ‘stage 1’ training on GTA5-to-Cityscapes adaptation. From (a) to (c) are: target domain inputs; ground-truth labels; pseudo-labels from ‘stage 1’.

produce domain-invariant outputs. Fig. 3 shows some output examples of our source and target modules during training. As a result of this proposal, ‘stage 1’, which serves as warm-up phase for training TridentAdapt, obtains 9.9 mIoU gain over source-only model on GTA5-to-Cityscapes adaptation and improves the source-only model by 10.4 mIoU for the 13-class setting on SYNTHIA-to-Cityscapes adaptation. ‘Stage 2’ training is thence able to benefit a lot from this warm-up phase, achieving state-of-art-results on both benchmark datasets for semantic segmentation. Fig. 4 and Fig. 5 show the mIoU curves of ‘stage 1’ and ‘stage 2’ on both benchmark datasets with the growth of training iterations. While ‘stage 1’ starts with 0, the mIoU of ‘stage 2’ training begins with a much higher number based on ‘stage 1’ result. The mIoU gets further improved with help of the self-training in ‘stage 2’. Interestingly,


Method	road	sdwk	blndg	wall*	fence*	pole*	light	sign	veg	sky	psn	rider	car	bus	mcycl	bcycl	mIoU	mIoU*
Source-only 	55.6	23.8	74.6	-	-	-	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	-	38.6
TridentAdapt (stage 1)	88.2	49.1	78.4	6.5	0.8	28.2	7.5	17.6	80.9	85.5	50.3	16.0	84.2	42.5	12.0	24.2	42.0	49.0
TridentAdapt (stage 2)	89.5	51.9	79.1	7.3	1.1	34.3	15.2	25.8	80.4	88.0	57.3	19.2	87.5	52.2	18.6	42.1	46.8	54.4

Table 2: Synthia-to-Cityscapes adaptation results. we provide class-wise performances of TridentAdapt in ‘stage 1’ and ‘stage 2’. mIoU, mIoU* refer to 16-class and 13-class experiment settings, respectively.



Figure 3: Outputs of our source and target modules during training with respect to source and target domain inputs (GTA5-to-Cityscapes adaptation). For any given input data (source or target domain), the simultaneous intra-domain reconstruction and cross-domain transfer impose confrontational constraints to the shared module, targeting at domain-invariance on feature level.

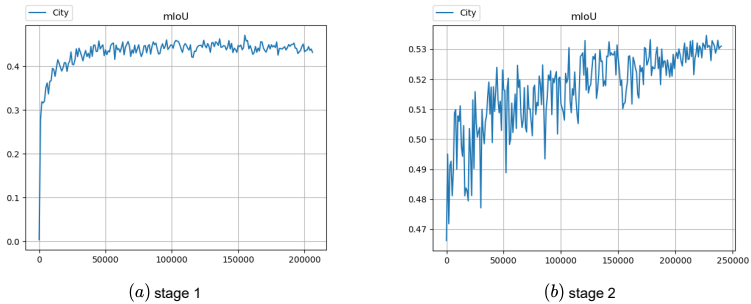


Figure 4: mIoU curve on the validation set of Cityscapes for GTA5-to-Cityscapes adaptation. The model performance is evaluated every 1000 iterations.

for bicycle class in GTA5-to-Cityscapes adaptation Table 1, source-only model performed better. The reason is that in this benchmark motorcycle and bicycle classes look visually similar, which makes the model classify many bicycles into motorcycles under the influence of cross-domain image transfer.

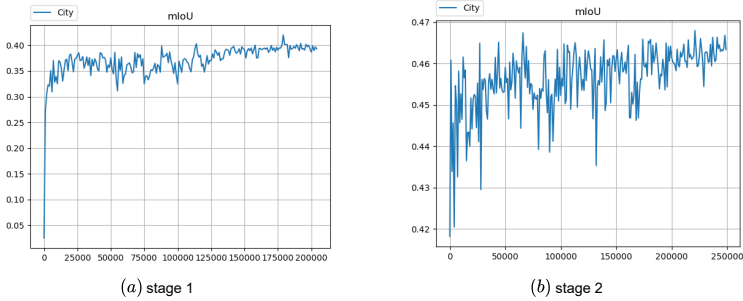


Figure 5: mIoU curve on the validation set of Cityscapes for Synthia-to-Cityscapes adaptation based on 16-class setting. The model performance is evaluated every 1000 iterations.

4 Combining with Mixup based data augmentation methods

We show that TridentAdapt, which is based on adversarial data augmentation, can be compatible with other popular Mixup based data augmentation methods [9, 10, 11]. When trained with those data augmentation techniques, TridentAdapt achieves new milestones in terms of segmentation performance. To illustrate this, We take the augmentation strategy of DACS [11], the domain adaptive version of Classmix [9] as example and train our ‘stage 2’ model. However, we do not adopt teacher-student training pipeline as presented in DACS [11]. In TridentAdapt, data mixture needs to be applied on target domain input x^t and its domain transferred version $G^s(E(x^t))$ and acquire images M and M^* , on which segmentation losses are computed in place of $\hat{\mathcal{L}}_{seg}^t$ and $\hat{\mathcal{L}}_{seg}^{t \rightarrow s}$. Since source domain input x^s and its target-like version $G^t(E(x^s))$ can be both involved in data mixture, as shown in Fig. 6, two out of four possible augmentations can be available for random pick at each iteration. Combining adversarial data augmentation and Mixup based augmentation, as presented in Table 3, our framework reaches 54.2 mIoU for GTA5-to-Cityscapes adaptation. Likewise, Synthia-to-Cityscapes adaptation reaches 48.5 mIoU for 13-class setting and 56.2 for 16-class setting, respectively. mIoU curves can be observed in Fig. 7.

Method	GTA5	Synthia
TridentAdapt	53.3	46.8 (54.4*)
TridentAdapt + CutMix	53.7	47.4 (55.3*)
TridentAdapt + ClassMix	54.2	48.5 (56.2*)

Table 3: Quantitative study for combining adversarial augmentation of TridentAdapt and Mixup based augmentation on domain adaptive semantic segmentation.



Figure 6: Enumeration of data augmentations combining adversarial augmentation of TridentAdapt and Mixup based augmentation used in DACS [11]

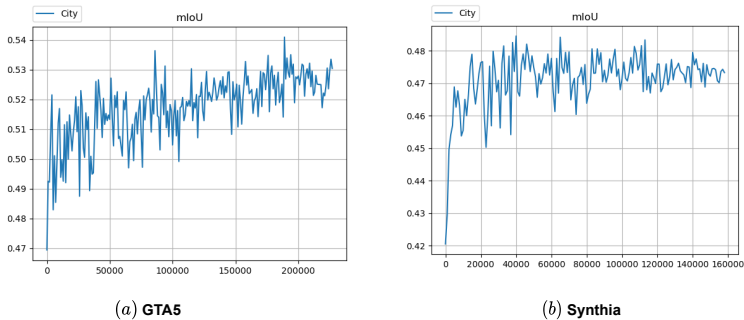


Figure 7: mIoU curves for ‘stage 2’ training after combining with Mixup based data augmentation technique.

5 Quality comparison of image domain transfer

In the main paper we claimed that, rather than building separate modules for segmentation and image translation, learning image translation by sharing the segmentation encoder and creating self-induced loop can ensure more sufficient mutual refinement between encoder and the generator, thus yielding higher image quality as rich semantic knowledge is incorporated. Here we evaluate images generated from CycleGAN [18] (the basic building block of CyCADA [4] and TIR [6]), BDL [4] and TridentAdapt. Since other approaches only involve source-to-target translation into training, for fairness we compare semantic segmentation models that are trained on target-like GTA5 images and validated on Cityscapes validation set. As shown in Table 4, training with TridentAdapt source-to-target translated images leads to a segmentation model of 44.5 mIoU, outperforming the corresponding results of CycleGAN [18] and BDL [4]. Therefore, we conclude that TridentAdapt’s image translation training pipeline incorporates more task-specific information into generator outputs, thus improving image quality over existing methods adopt separate networks for image translation and semantic segmentation. Qualitative visual comparison is presented in Fig. 8.

Translation Model	mIoU	Δ
Source-only (ours)	31.5	-
CycleGAN (S2T)	39.3	+7.8
BDL [4] (S2T)	42.7	+11.2
TridentAdapt (S2T)	44.5	+13.0

Table 4: Quantitative comparison of source-to-target image translation results for domain adaptive semantic segmentation.



Figure 8: comparison of image translation results. From top to bottom: input, CycleGAN, BDL [4] and TridentAdapt output.

Even though our source-to-target translated images show superiority over other methods on target domain when utilized solely to train a segmentation model, we still believe that involving original source domain data for training can help compensate the lost semantic

Method	TridentAdapt_S2T	TridentAdapt_S2T + S	TridentAdapt_S2T + Pseudo	TridentAdapt_S2T + S + Pseudo
mIoU	44.5	45.0	50.3	50.7

Table 5: Synthia-to-Cityscapes adaptation results. Ablation study of training a segmentation model on target-like source data with or without source data and pseudo-labelled target images.

information due to the adversarial disturbances during image translation. To this end, we conducted ablative experiment and present our findings in Table 5. It can be observed that involving original source domain image for training improves mIoU of the target-like-only model from 44.5 to 45.0. On the other hand, assisting the target-like-only model by training with its derived pseudo-labels on target domain dataset brings a larger improvement from 44.5 to 50.3. However, training with them all achieves obtains 50.7 mIoU. Therefore, we conclude that original source domain data should not be neglected for training even if we have an image translation model that produces high image domain transfer quality. And pseudo labels are always helpful for further improvement.

6 Comparison with SOTA methods based on image domain transfer

Now we compare domain adaptive semantic segmentation performance among TridentAdapt and relevant works, and analyse the superiority of TridentAdapt over others. In Fig. 10, we visually demonstrate the performance of TridentAdapt for domain adaptation by including additional segmentation results on Cityscapes validation set. Qualitatively, We compare TridentAdapt results with source-only model, BDL [2] and TIR [6]. Interestingly, we can observe in Table 6 that, TridentAdapt outperforms other methods in preserving class boundaries (metric presented in Fig. 9) for semantic segmentation, obtaining 37.59 mIoU for predicting correct semantic edges. We believe that this is owing to our image recovery based training strategy, such that the learned knowledge in our generators for reconstructing image edges will be reflected on the shared encoder and force the segmentation module to better preserve edge or class boundary pixels in segmentation maps.

Methods	overall	edges	non-edges
source-only	31.5	22.9	38.4
BDL [2]	48.5	31.3	56.7
TIR [6]	50.2	35.0	59.7
TridentAdapt	53.3	37.6	63.1

Table 6: Quantative results of GTA5-to-Cityscapes adaptation between source-only; BDL [2]; TIR [6]; TridentAdapt on overall segmentation mask, around semantic edges and non-edges regions.

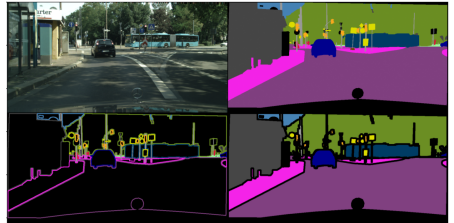


Figure 9: Visual representation of semantic-edges (bottom-left) and non-edges (bottom-right) region. The 1st row displays the RGB image (top-left) and respective semantic segmentation mask (top-right).

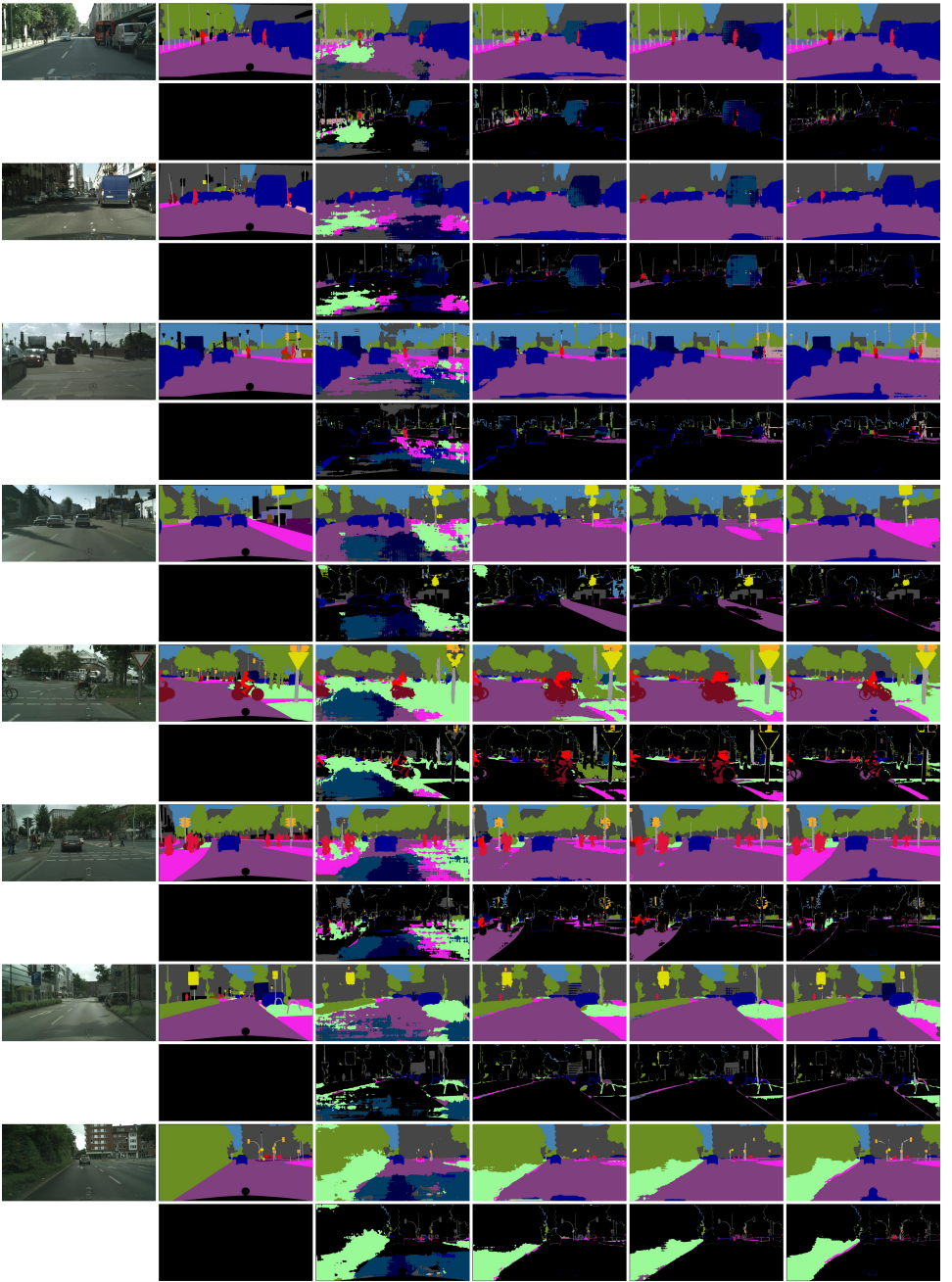


Figure 10: Additional qualitative results of GTA5-to-Cityscapes adaptation on Cityscapes validation set. For each group, the upper row; From left to right: target domain input image; their segmentation ground-truth labels; predictions of source-only model; BDL [0]; TIR [0]; TridentAdapt and the lower row; the error map of their respective approaches. The error map displays only the segmentation class for wrongly labeled pixels and black mask for correct predictions.

7 Analysis on edge-preserving in image reconstruction loss

In the paper, our motivation for up-weighting the sobel edges in reconstructed images is to compensate the pixel information that are lost due to maxpooling operators in the backbone encoder. To investigate how up-weighting the edges in the reconstruction loss could affect the performance of TridentAdapt, we plot an ablative study for ‘stage 1’ mIoU by changing the weight for edge pixels in our image reconstruction loss (see Fig. 11), and empirically we find that setting the weight to 0.5 produces the best result.

In addition, we also consider to preserve segmentation edges instead of sobel edges for image reconstruction. In this case, since no ground truth for target domain is available, meanwhile segmentation edges for adapted predictions on target domain are not accurate enough to be considered (see row 2 of Fig. 12), we only take segmentation edges from source domain for computing image reconstruction loss. However, training ‘stage 1’ with help of source segmentation edges gives 46.3 mIoU, which does not outperform our sobel edge setting. One reason is that lots of segmentation edge pixels are overlapping with sobel edges, thus having similar effect for edge preserving. Another reason is that the source domain segmentation edges are overly fine-grained, such that some unwanted semantic edge information are incorporated into image reconstruction phase, bringing some disturbances which counteract the benefits brought by using the segmentation edges. (See circles on road and sidewalk region of source edge map in Fig. 12)

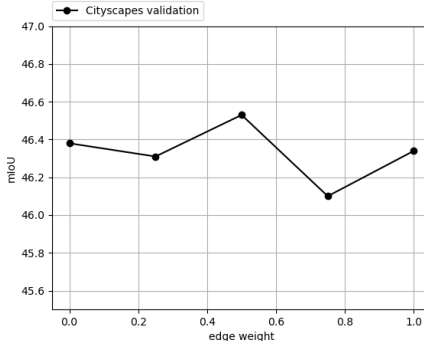


Figure 11: Ablation study for up-weighting the sobel edges in the reconstruction loss

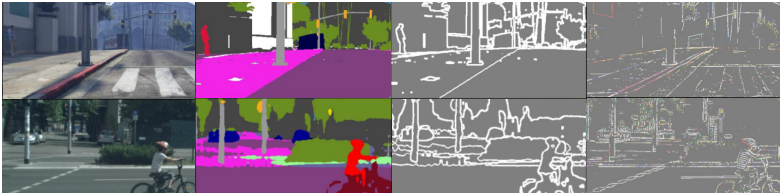


Figure 12: Visualization of semantic edges and sobel edges. First row from left to right: source domain input, segmentation ground truth, segmentation edge map, sobel edge map; Second row from left to right: target domain input, target segmentation prediction, segmentation edge map, sobel edge map.

8 Discussion

Why does TridentAdapt outperform other image translation based methods for domain adaptation? First, image-to-image translation models in existing approaches [2, 4, 5, 2, 3, 16] are trained based on CycleGAN [15] settings, which adopt different encoders and decoders for forward and backward translations. However, for the purpose of domain adaptation, image-to-image translation can be more properly explored with source-target confrontation concept. Therefore, in TridentAdapt all input images are passed to a shared encoder. In this way, for each input we are able to introduce an intro-domain image recovery step (i.e., self-reconstruction) and a cross-domain image transfer step, thus placing confrontational constraints on the encoder to force domain-invariance on feature level. However, this cannot be achieved following CycleGAN training settings.

Second, the above approaches seek to build separate networks for image translation and semantic segmentation purposes, where domain transfer modules and segmentation encoder do not act appropriate constraints on each other, thus lacking coherent sufficient interaction. Therefore, the potential of image translation is not fully explored to support domain adaptive semantic segmentation. TridentAdapt tackles this by sharing the encoder of semantic segmentation network with image translation networks, thus conveying the learned semantic knowledge into image translation step to acquire better generators G^s and G^t . The improved generators will, in return, produce stronger confrontational constraints to the shared encoder. Moreover, better quality outputs from generators can well support the subsequent steps.

Third, the trident-like design enables self-induced cross-domain data augmentation on-the-fly in a forward pass. This means, for each input there will be an extra domain transferred version that is utilized for segmentation sharing the same label (or pseudo-label). The cross-domain views augmented by both generators are proven to further contribute to domain gap reduction. In previous approaches, however, only source-to-target translated images produced by a separated generator network are considered into training. With TridentAdapt, we confirm experimentally that target-to-source transfers can also be beneficial to domain adaptation by widening input data coverage on domain level.

Does the trident-like design of TridentAdapt introduce large computational complexity for training? The computational complexity that is introduced in TridentAdapt is endurable in comparison with previous approaches which either involves depth modality for training [13], or constructing a teacher module which has the same amount of parameters as the student [10]. It can be observed in Table 7 that our generators and discriminators are relatively smaller networks compared to the backbone encoder and segmentation network. Therefore, the major computation costs are on E and Φ . Additionally, since we do not use separate encoders for image translation, if trained on the same backbone, our architecture actually has less total parameters than conventional image translation based approaches which have to build additional image encoders for both domains. In fact, the inference model after training is the same for all SOTA approaches if the same backbone is adopted. Hence, computational complexity has no impact on the final models.

What else can be investigated to improve TridentAdapt? First, to achieve our design purpose, we assume that our source and target modules should be effective enough for approximating domain data distributions. Therefore, seeking better discriminator solutions will enforce our generators to output higher fidelity cross-domain transfer results, yielding stronger constraints regarding source-target confrontation as well as inducing higher quality cross-domain augmented views. To this end, a recent work proposing UNet based discriminator [14] which gives pixel-wise judgement of image realism can be interesting to try. Other

Sub-modules	$E + \Phi$	G^s	G^t	D^s	D^t
params.	44.6 M	5.11 M	5.11 M	8.27 M	8.27 M

Table 7: Number of trainable parameters (millions) for each sub-module in TridentAdapt. We adopt ResNet-101 [9] for E and Deeplab-V2 [10] for Φ .

than that, relativistic discriminator loss [9] has shown promising outcomes when applied to image super-resolution tasks [15], and can also be a helpful building block for TridentAdapt. Second, the design of TridentAdapt seeks to learn an encoder which maps each input to a shared feature space to reduce the domain gap. Adversarial learning is taking the leading role in this process. However, TridentAdapt is not supposed to create a structured feature space whose class-wise spatial regions are discriminative among each other. This is the reason for the vagueness between motorcycle and bicycle classes. To resolve this, pixel contrast loss [14] is a good option for future experiments.

9 Limitation

Synthia dataset has been considered as a more challenging dataset for domain adaptive semantic segmentation. Although our TridentAdapt framework shows promising results on Synthia-to-Cityscapes adaptation in comparison with prior works, we would like to briefly analysis why there is a performance gap between adapted results using GTA5 and Synthia dataset. As shown in Fig. 13 (a), Synthia dataset contains training images with various weather, illumination and camera pose&distance changes. In image-to-image translation, datasets like this are often considered for multi-domain setting, however, for domain adaptive segmentation these images are considered as a single source domain, which makes it quite challenging for the network to learn what kind of domain specific information should be incorporated into the image translator. Therefore, in some cases, image-to-image translation modules fail to capture the correct cross-domain mapping. For example, in Fig. 13 (c), when an image of low visibility is translated into target domain using our G^t , the rider and bicycle are hidden into the road pixels, which does not bring benefit to semantic segmentation. Moreover, as Fig. 13 (b) shows, the Synthia class distribution in the label map is quite different from that of Cityscapes which are closer to GTA5 (e.g, less pedestrian pixels in a single image). Therefore, for TridentAdapt which is trained based on image translation, adapting Synthia as source domain to Cityscapes is more challenging than doing it for GTA5.

10 Outlook

We believe that the TridentAdapt design is highly flexible and can be incorporated on a broader range of tasks such as domain adaptation for virtual-to-real depth estimation and day-to-night perception.

References

- [1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous

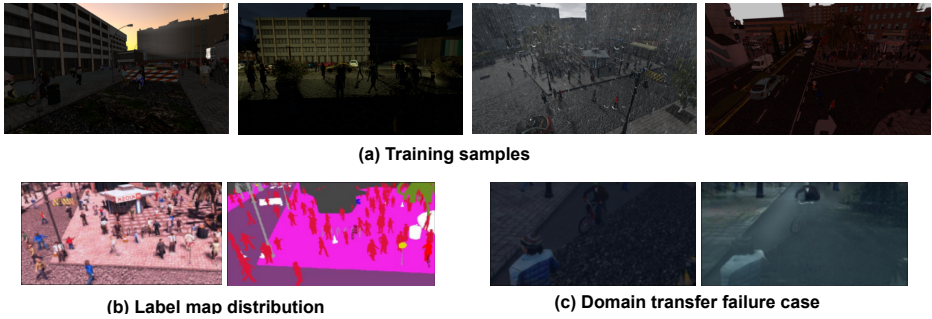


Figure 13: Examples of training images with various challenging conditions from Synthia dataset, which poses large difficulty for learning image domain transfer.

convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

- [2] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1791–1800, 2019.
- [3] 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.
- [4] 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. PMLR, 2018.
- [5] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard GAN. In *International Conference on Learning Representations*, 2019.
- [6] 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.
- [7] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019.
- [8] Luigi Musto and Andrea Zinelli. Semantically adaptive image-to-image translation for domain adaptation of semantic segmentation. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*, 2020.
- [9] Viktor Olsson, Wilhelm Tranheden, Julianio Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021.

- [10] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8207–8216, 2020.
- [11] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021.
- [12] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schuster, 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.
- [13] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7364–7373, 2019.
- [14] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *arXiv preprint arXiv:2101.11939*, 2021.
- [15] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.
- [16] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2100–2110, 2019.
- [17] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.
- [18] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.