Supplementary Material: Self-Supervised Learning of Domain Invariant Features for Depth Estimation

Hiroyasu Akada[1,2] Shariq Farooq Bhat[1] Ibraheem Alhashim[3] Peter Wonka[1]

[1]KAUST, [2]Keio University [3]National Center for Artificial Intelligence (NCAI), Saudi Data and Artificial Intelligence Authority (SDAIA)

hiroyasu5959@keio.jp shariq.bhat@kaust.edu.sa {ibraheem.alhashim, pwonka}@gmail.com

A. Training stages and loss functions

In this section, we describe the style transfer stage, the depth estimation stage, and relevant loss functions in more details. Please also see Section 3 in our main paper.

A.1. Style transfer stage

In this stage, we aim to train the image-to-image translation networks $G_{S\to T}$ and $G_{T\to S}$. $G_{S\to T}$ learns a mapping between the synthetic (*i.e*. source) and realistic (*i.e*. target) domain. $G_{T\to S}$ learns a mapping in the opposite direction. Previous work [17, 16] has shown that the depth supervision can help to improve the quality of style transfer compared to the standalone training of the translation networks. Although the effect of the depth supervision seems minor based on our experiment as shown in Table 7, we adopt it for all experiments and jointly train the image-to-image translation networks $G_{S\to T}$ and $G_{T\to S}$, and the image-to-depth task network F with synthetic depth labels. Here, based on previous work [17, 2], we adopt six loss functions: adversarial loss L_{adv}, cycle consistency loss L_{cycle}, identity mapping loss $L_{identity}$ (or reconstruction loss in [17]), task loss L_{task}, smooth loss L_{smooth}, and cross domain consistency loss L_{crdoco}. These losses are described in the following.

Adversarial loss We utilize adversarial training for the image translation. Following CycleGAN [18], we employ two generators $G_{S\to T}$ and $G_{T\to S}$, and two discriminators D_s and D_t for the source and target domains, respectively. $G_{S\to T}$ tries to learn the mapping from the source to the target domain, *i.e*. $G_{S\to T} : I_S \to I_{S\to T}$, such that the data distribution of the translated images from the source domain $I_{S\to T}$ is indistinguishable from that of the target domain I_T. Then, D_T aims to distinguish between the images in the target domain I_T and the translated images from the source domain $I_{S\to T}$. Thus, using the technique of a least-square loss [11] for stable training, we define adversarial loss [7] in the target domain as

$$\begin{aligned} &L_{adv}(G_{S\to T}, D_T, X_S, X_T) \\ &\quad = \mathbb{E}_{I_T\sim X_T}[(D_T(I_T)-1)^2] \\ &\quad + \mathbb{E}_{I_S\sim X_S}[(D_T(G_{S\to T}(I_S)))^2]. \end{aligned} \tag{3}$$

Similarly, the adversarial loss for the mapping function $G_{T\to S} : I_T \to I_{T\to S}$ is introduced as $L_{adv}(G_{T\to S}, D_S, X_T, X_S)$. Please note that as a result of the bidirectional adversarial training, we obtain two labeled types of images I_S and $I_{S\to T}$, and two unlabeled types of images I_T and $I_{T\to S}$ as shown in Fig 2 in our main paper.

Cycle consistency loss L_{cycle}. Generally, training $G_{S\to T}$ and $G_{T\to S}$ with only the adversarial loss are highly underconstrained. To further regularize the translation network, we use a cycle consistency loss [18, 15]. This loss function is based on the idea that when images are translated from one domain to another, followed by an inverse translation, the reconstructed images should be the same as the original, *i.e*. $G_{T\to S}(G_{S\to T}(I_S)) \approx I_S$ and $G_{S\to T}(G_{T\to S}(I_T)) \approx I_T$. Therefore, we define the cycle consistency loss as

$$\begin{aligned} &L_{cycle}(G_{S\to T}, G_{T\to S}, X_S, X_T) \\ &\quad = \mathbb{E}_{I_S\sim X_S}[\|G_{T\to S}(G_{S\to T}(I_S))) - I_S\|_1] \\ &\quad + \mathbb{E}_{I_T\sim X_T}[\|G_{S\to T}(G_{T\to S}(I_T)) - I_T\|_1]. \end{aligned} \tag{4}$$

Identity mapping loss $L_{identity}$. In addition to L_{cycle}, we also use an identity mapping loss [18, 13] to regularize the training of $G_{S\to T}$ and $G_{T\to S}$. Note that in [17], this loss function is named as 'reconstruction loss'. The identity mapping loss encourages $G_{S\to T}$ and $G_{T\to S}$ to preserve image styles when the input images already belong to the translation-target domain, *i.e*. $G_{S\to T}(I_T) \approx I_T$ and

Method	Lower is better				Higher is better		
	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
No depth supervision	**0.174**	1.490	5.751	0.244	0.766	0.910	0.964
depth supervision	**0.174**	**1.439**	**5.701**	**0.241**	**0.770**	**0.914**	**0.967**

Table 7: Ablation studies on the effect of the depth supervision in the style transfer stage using CrDoCo* [2] on KITTI [5].

Method	Lower is better				Higher is better		
	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
$\lambda_{crdoco} = 10.0$	0.178	1.269	5.845	0.245	0.742	0.910	0.970
$\lambda_{crdoco} = 1.0$	**0.168**	**1.228**	**5.498**	**0.235**	**0.771**	**0.921**	**0.973**
w/o λ_{crdoco}	0.176	1.443	5.676	0.240	0.768	0.917	0.971

Table 8: Ablation studies with the different number of λ_{crdoco} on KITTI [5].

$G_{T\to S}(I_S) \approx I_S$. Specifically, we define the identity mapping loss as

$$\begin{aligned} &L_{identity}(G_{S\to T}, G_{T\to S}, X_S, X_T) \\ &\quad = \mathbb{E}_{I_S\sim X_S}[\|G_{T\to S}(I_S)) - I_S\|_1] \\ &\quad + \mathbb{E}_{I_T\sim X_T}[\|G_{S\to T}(I_T) - I_T\|_1]. \end{aligned} \tag{5}$$

Task loss L_{task}. To train the image-to-depth task network F, we provide supervision to F using synthetic ground truth depth maps $I_{S,lab}$. Specifically, we pass the two labeled types of images I_S and $I_{S\to T}$ to F to obtain corresponding depth maps $P_S = F(I_S)$ and $P_{S\to T} = F(I_{S\to T})$ as shown in Fig 1 in our main paper. Since these depth maps should have the same label $I_{S,lab}$, we define the task loss as

$$\begin{aligned} &L_{task}(G_{S\to T}, F, X_S) \\ &\quad = \mathbb{E}_{I_S\sim X_S}[\|F(I_S) - I_{S,lab}\|_1] \\ &\quad + \mathbb{E}_{I_S\sim X_S}[\|F(G_{S\to T}(I_S)) - I_{S,lab}\|_1]. \end{aligned} \tag{6}$$

Smooth loss L_{smooth}. Following previous works [17, 4, 6, 8, 9, 16, 2] we utilize a a smooth loss to guide a more reasonable depth estimation using the unlabeled images I_T and $I_{S\to T}$. Specifically, we use a robust penalty with an edge-aware term for I_T as

$$\begin{aligned} L_{smooth}(F, X_T) = &\mathbb{E}_{I_T\sim X_T}[|\partial_x F(I_T)|e^{-|\partial_x I_T|}] \\ &+ \mathbb{E}_{I_T\sim X_T}[|\partial_y F(I_T)|e^{-|\partial_y I_T|}]. \end{aligned} \tag{7}$$

Similarly, the smooth loss for $I_{S\to T}$ is introduced, *i.e.* $L_{smooth}(F, X_{S\to T})$.

Cross domain consistency loss L_{crdoco}. Following the previous work [2], we also introduce a cross domain consistency loss to enforce the consistency between the depth predictions of the two unlabeled types of images $P_T = F(I_T)$ and $P_{T\to S} = F(I_{T\to S})$. More specifically, we define the cross domain consistency loss as

$$\begin{aligned} &L_{crdoco}(G_{T\to S}, F, X_T) \\ &\quad = \mathbb{E}_{I_T\sim X_T}[\|F(I_T) - F(G_{T\to S}(I_T))\|_1]. \end{aligned} \tag{8}$$

Full objective. The overall objective function in the style transfer stage is defined as

$$\begin{aligned} L_{style\_transfer} = &L_{adv} + \lambda_{cycle}\cdot L_{cycle} \\ &+ \lambda_{identity}\cdot L_{identity} \\ &+ \lambda_{task}\cdot L_{task} + \lambda_{smooth}\cdot L_{smooth} \\ &+ \lambda_{crdoco}\cdot L_{crdoco}, \end{aligned} \tag{9}$$

where each λ controls the relative importance of each objective. Then, we optimize the following min-max problem in the style transfer stage:

$$F^* = \arg\min_{F} \min_{\substack{G_{S\to T}\\ G_{T\to S}}} \max_{\substack{D_S,\\ D_T}} L_{style\_transfer}. \tag{10}$$

In later stages, we leverage $G_{S\to T}$ and $G_{T\to S}$ pretrained in this style transfer stage.

A.2. Depth estimation stage

As the last stage in our UDA framework, we fine-tune the image-to-depth task network F trained in the SSRL stage, by using both synthetic and real-world datasets as shown in Fig 2 in our main paper. The network architecture in this stage is similar to that in the style transfer stage but the weights of the image-to-image translation networks $G_{S\to T}$ and $G_{T\to S}$ are fixed for faster training. Specifically, We train F, *i.e.* both F_{enc} and F_{dec}, by minimizing the follow-

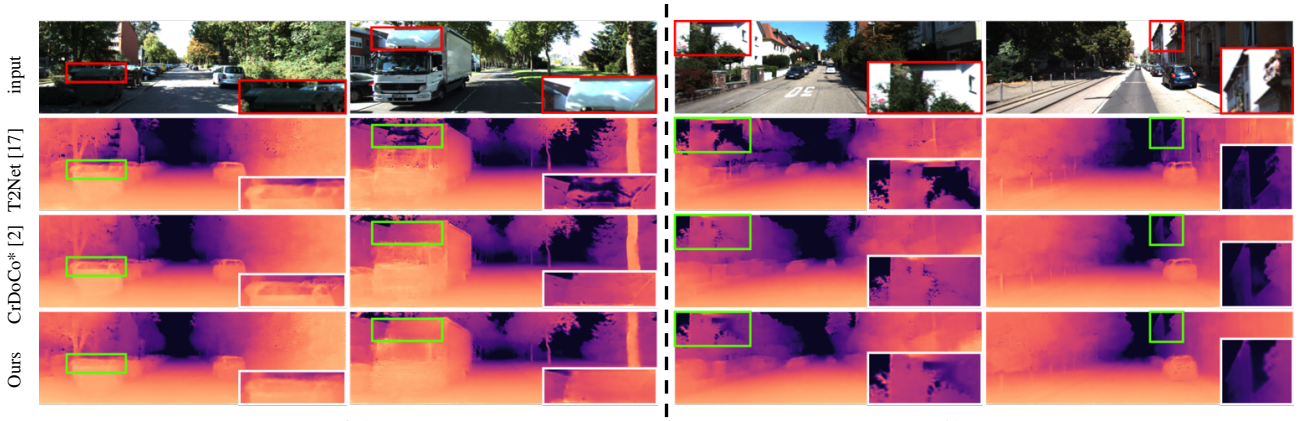

Figure 4: Qualitative results for successful and failure cases on KITTI [5].

ing objectives:

$$L_{depth\_estimation} = \lambda_{task} \cdot L_{task} + \lambda_{smooth} \cdot L_{smooth} + \lambda_{crdoco} \cdot L_{crdoco}, \tag{11}$$

where each λ controls the relative importance of each objective.

B. Training detail

In this section, we provide more details of our hyper-parameter setting and training strategy.

B.1. Hyper-parameters

As mentioned in Section 5 in our main paper, we set the relative weights of the different loss functions based on previous works and our experiments. Specifically, we follow [18] to set $\lambda_{cycle} = 10$ for our bidirectional image-to-image translation network. Also, similar to [17], we set $\lambda_{identity} = 100$, $\lambda_{task} = 100$, $\lambda_{smooth} = 0.1$. Lastly, we set $\lambda_{crdoco} = 1$ based on our ablation study as in Table 8.

B.2. Encoder with ImageNet initialization

As mentioned in Section 5 in our main paper, we leverage EfficientNet-B5 [14] pre-trained on ImageNet [3] as the encoder F_{enc} of the task network F. EfficientNet-B5 mainly consists of 9 stages and each stage yields feature maps with a different number of channel and resolution sizes. To build the encoder-decoder architecture, we remove its last dense layer. Note that we follow previous works [1, 10, 17, 2] to utilize skip connections [12].

During optimization, we utilize differential learning rates for the encoder F_{enc} based on the stage as shown in Table 9. More specifically. we use relatively lower learning rates for the initial few stages since these stages are already good at extracting general information, such as edges, through ImageNet initialization. By contrast, the last few stages are trained with relatively higher learning rates to enable F_{enc} to adopt to our depth estimation task. Note that we train our comparison methods using our task network F together with the ImageNet initialization and the differential learning rates for a fair comparison. Please refer to our implementation code for more details.

Stage	Output channels	Learning rate
1	48	lr / 10^3
2	24	lr / 10^3
3	40	lr / 10^3
4	64	lr / 10^3
5	128	lr / 10^2
6	176	lr / 10^2
7	304	lr / 10^1
8	512	lr / 10^1
9	2048	lr / 10^1

Table 9: Details of differential learning rates applied to EfficientNet-B5 [14] pre-trained on ImageNet [3] as the encoder F_{enc} of the task network F. Note that we set a base learning rate $lr = 0.0004$ for F as mentioned in Section 5 in our main paper.

C. Additional qualitative result

We provide additional qualitative results in Fig 4, highlighting relatively successful and failure cases. From the successful cases, our method is better at estimating consistent depth values on objects' surfaces than comparable methods [17, 2]. It is also worth analyzing the failure cases for future research. As a common trend, current UDA methods including our method fail to handle reflective surfaces (or overexposed white regions) in images. We suspect that this is because the larger surfaces with such a uniform color

do not provide useful information for depth estimation. One possible solution would be an introduction of an attention mechanism to utilize global information between pixels.

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins, 2020.

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

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[4] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pages 740–756. Springer, 2016.

[5] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[6] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[7] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[8] Philipp Heise, Sebastian Klose, Brian Jensen, and Alois Knoll. Pm-huber: Patchmatch with huber regularization for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2360–2367, 2013.

[9] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6647–6655, 2017.

[10] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[11] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, and Zhen Wang. Multi-class generative adversarial networks with the l2 loss function. *arXiv preprint arXiv:1611.04076*, 5:00102, 2016.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[13] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

[14] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[15] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2868–2876, 2017.

[16] Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9788–9798, 2019.

[17] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[18] 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.