Supplementary for LoopDA: Constructing Self-loops to Adapt Nighttime Semantic Segmentation

Fengyi Shen[1,2], Zador Pataki[2,4], Akhil Gurram[2], Ziyuan Liu[2], He Wang[3], Alois Knoll[1]
[1]Technical University of Munich, [2]Huawei Munich Research Center
[3]Peking University, [4]ETH Zurich
[1]fengyi.shen@tum.de,knoll@in.tum.de,[2]{first.last}@huawei.com,[3]hewang@pku.edu.cn

In this Supplementary, we provide additional experiments for evaluating our LoopDA framework as well as additional results obtained from our trained models.

1. Label fusion

As discussed in Sec.3.1.1 in the main paper, to ensure that p^d of the daytime domain is accurate enough for rendering, we fuse p^d and its ground-truth label y^d in each training for fine-grained rectification via linear combination,

$$p^d \leftarrow \alpha \cdot y^d + (1-\alpha) \cdot \mathcal{O}(\arg\max(p^d)) \qquad (1)$$

where $\mathcal{O}$ is the one-hot operator and α is a constant value between 0 and 1.

For better visualization, we show the fused results for different chosen α values at early warm-up phase (See Fig. 1). It is known to the community that a segmentation model cannot perform exactly the same as the ground-truth labels no matter how well they are trained. But in our case, to better condition the network for domain adaptation towards nighttime, we need good quality p^d in daytime domain for semantic rendering. Therefore, we use label fusion to combine y^d into p^d for refinement. In this process, we argue that the smallest possible value for α can be determined based on early warm-up training phase, where the segmentation head is only able to produce rough segmentation. Hence, larger weight for y_d is needed. We compare in Fig. 1 and observe that 0.5 is a suitable value for α. This ensures that the refined p^d perfectly overlaps with y^d also in class boundary regions. For the invalid labels (don't care regions) in y^d, we just assign p^d values to them.

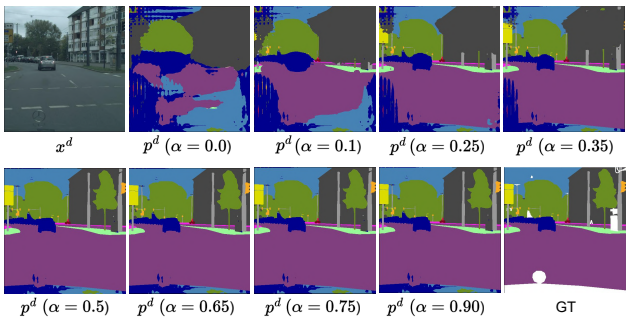

Figure 1. An illustration of the label fusion procedure to refine p^d. Different choices of α are visualized.

2. Analysis of image domain transfer

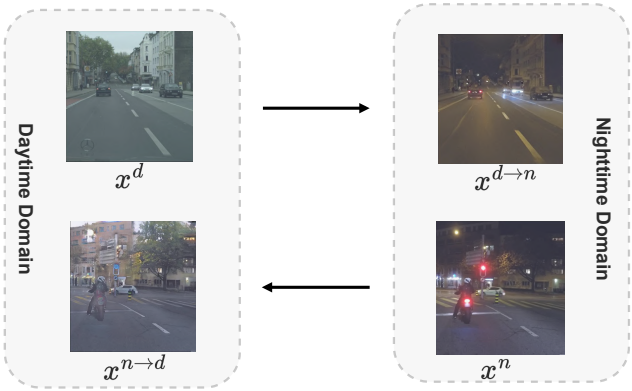

Figure 2. Examples of image domain translation in LoopDA taken from the outputs of our Dec^n and Dec^d.

In our ablation study of the main paper, we demonstrate that the inter-domain outer loop plays a crucial role for the performance improvement in the warm-up stage. Without the outer loop, the performance drops from 29.5 to 22.0 mIoU. The reason is that our image translation modules enforce the encoder to treat input images and their domain transferred version equally, thus learning domain agnostic representation from the input data. As Fig. 2 shows, this bi-directional alignment of image translation enriches the data diversity at input level. From the perspective of $x^{d\to n}$, it mimics nighttime appearance regarding the characteristics such as style and illumination, but maintains the structural contents of x^d. In terms of $x^{n\to d}$, it manages to recover low-light objects from x^n for better illumination condition

and resembles daytime appearances. Thus, the segmentation network is trained to be more robust across day and night domains.

Image translation between day and night is a challenging task, therefore, an image translator producing high quality outputs can better assist the downstream tasks such as semantic segmentation. For verification of our proposal, we also have experimented on the image translation results with and without semantic rendering layers, as well as an alternative variant of semantic rendering, SPADE layers [4], which are designed particularly for semantic image synthesis (See Fig. 3).

For day to night image translation, our proposed decoder

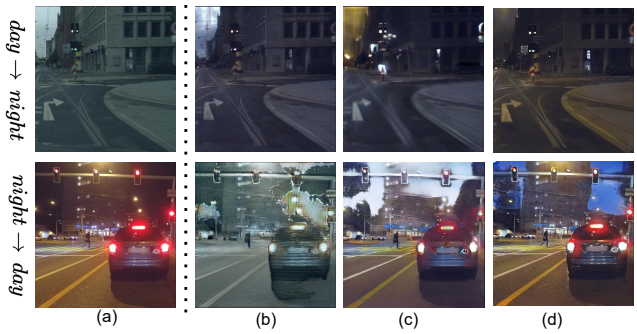

Figure 3. Examples of image domain translation. From (a)-(d): inputs, image domain translation using SPADE layers, LoopDA decoders w/o semantic rendering layers and LoopDA decoders w/ semantic rendering layers, respectively.

architecture with semantic rendering layers help preserve daytime contents during translation while producing images with higher fidelity. Translating from night to day in the second row of Fig. 3, SPADE is less effective dealing with night image translation (similar observation can be found in [5]), and decoder without rendering layers generate sky in building regions. However, the LoopDA decoder with rendering layers produces the best visual quality while recovering the building shape. This confirms the superiority of our decoder architecture design.

3. Additional ablative analysis

In addition to Table 3 in the main paper, in this Supplementary, we conduct three extra ablative experiments (marked in red) to provide more detailed information about our proposed LoopDA framework. Row(iv) indicates that sharing ground-truth between x^d and $x^{d\to n}$ substantially enhances the learning of domain agnostic features between daytime and nighttime domains. Excluding $\mathcal{L}_{seg}^{d\to n}$ from warm-up stage leads to a mIoU drop from 29.5 to 26.7. Row(v) suggests that it is meaningful to set additional constraints to maintain perceptual consistency during image translation, improving the generated image quality. If no

Table 1. Ablation study for Cityscapes → Dark Zurich adaptation results evaluated on Dark Zurich validation set.

Phase	Components	mIoU	
Warm-up stage	(i).baseline[6] on PSPNet (L_{seg}^{d})	20.6	+0.0
	(ii). no outer loop (w/o L_{percep}, L_{adv}, $L_{seg}^{d\to n}$, L_{outer})	22.0	+1.4
	(iii). no inner loop (w/o L_{inner})	26.3	+5.7
	(iv). no label sharing with $x^{d\to n}$ (w/o $L_{seg}^{d\to n}$)	26.7	+6.1
	(v). no perceptual consistency loss (w/o L_{percep})	27.1	+6.5
	(vi). no semantic rendering layers	28.6	+8.0
	(vii). LoopDA warm-up model	29.5	+8.9
ST stage	(viii). no offline pseudo-labels (without $\mathcal{G}_{off}$ in $\hat{\mathcal{L}}_{seg}^{n}$)	33.9	+13.3
	(ix). no 'DNA' (without $\mathcal{G}_{DNA}$ in $\hat{\mathcal{L}}_{seg}^{n}$)	35.7	+15.1
	(x). LoopDA full configuration ($\hat{\mathcal{L}}_{seg}^{n}$)	37.6	+17.0
	(xi). with extra distillation stage (LoopDA‡)	**38.7**	+18.1

$\mathcal{L}_{percep}$ is applied for warm-up training, the model performance decreases by 2.4 mIoU. Furthermore, in terms of the self-training stage, as can be observed in row(viii), if no offline pseudo-supervision is performed, the model segmentation for the dynamic objects is affected, thus bring the model accuracy from 37.6 down to 33.9 mIoU.

4. Additional details of LoopDA implementation

In supplement to the main paper, we provide additional details of implementing our proposed LoopDA framework. For the image decoders and discriminators, we adopt Adam [2] optimizers with default learning rate 1.0×10^{-3} but 1.0×10^{-4} for the discriminators. We apply polynomial decay policy to the learning rates. We set momentum between 0.9 and 0.99. We follow the discriminator architecture implemented in MUNIT [1], and as given in Equation 3 of the main paper, we apply LSGAN [3] adversarial loss for more stable GAN training.

For the image reconstruction losses, we set higher weights to the pixel locations which belongs to sobel gradients to prevent the output images from turning blurry. And for faster convergence, the LPIPS loss can be optionally applied to the image reconstruction process.

5. Discussion of parameter-free relighting

As [7] suggests, having a pre-processing step such as relighting before sending the input data for segmentation can help expose the challenging dark regions in nighttime data to certain degree, thus bringing benefits to the segmentation performance. Nevertheless, training a relighting network will lead to extra computational complexity. Most importantly, this means that the trained relighting network must be included into the inference phase all the time after training, which increases the inference time and is less practical. Therefore, we have explored to look for an alternative solution that also changes the illumination of the nighttime

inputs, but do not bring in trainable network parameters. To this end, we have experimented on a parameter-free daytime instance adaptive relighting mechanism. For incoming training samples x^d and x^n ($x^d, x^n \in \mathbb{R}^{3\ \ h\ \ w}$) in each iteration, we compute their channel-wise means μ^d and μ^n ($\mu^d, \mu^n \in \mathbb{R}^{3\ 1\ 1}$), and calculate the absolute value of the illumination differences pixel-wisely between them, i.e., $\Delta = |\mu^d - \mu^n|$. Then, we use this illumination difference to relight x^n element-wisely according to the following condition, $x^n_{i,j,k} = \begin{cases} x^n_{i,j,k} + \Delta_i, & x^n_{i,j,k} < \Delta_i \\ x^n_{i,j,k}, & x^n_{i,j,k} \geq \Delta_i \end{cases}$. Moreover, we also update a global Δ^0 using exponential moving average, $\Delta^0 \leftarrow \beta \cdot \Delta + (1 - \beta) \cdot \Delta^0$ where β = 0.001, which prepares the relighting values at inference time. As shown in Fig. 4, this parameter-free relighting procedure is able to improve the visual visibility of the nighttime inputs, recovering some low-light objects. Nevertheless, after including this step into our LoopDA training, we have not yet observed substantial improvement to our segmentation task. However, we still would like to share these interesting visual observations with the community and leave them for future investigation.

Figure 4. Examples of daytime instance adaptive relighting for x^n.

6. Additional qualitative comparison

In this section we provide more qualitative comparison between LoopDA and state-of-the-art methods on Dark Zurich validation set in Fig. 5.

References

[1] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[3] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[4] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[5] Stephan R Richter, Hassan Abu Al Haija, and Vladlen Koltun. Enhancing photorealism enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[6] 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.

[7] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15769–15778, 2021.

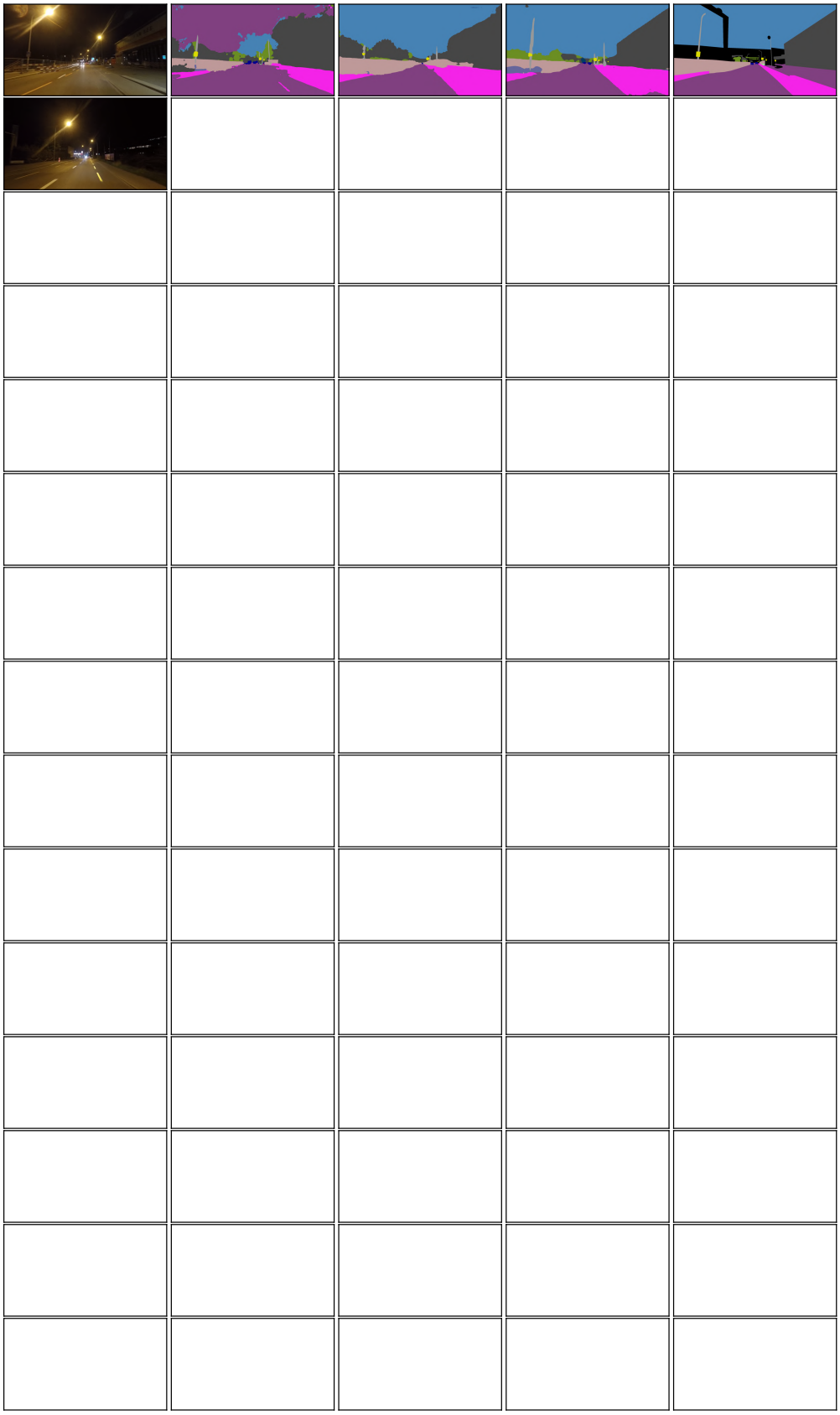

Figure 5. Qualitative comparison with state-of-the-art methods for Cityscapes → Dark Zurich adaptation on Dark Zurich validation set.