Supplementary: Diffusion-based Image Translation with Label Guidance for Domain Adaptive Semantic Segmentation

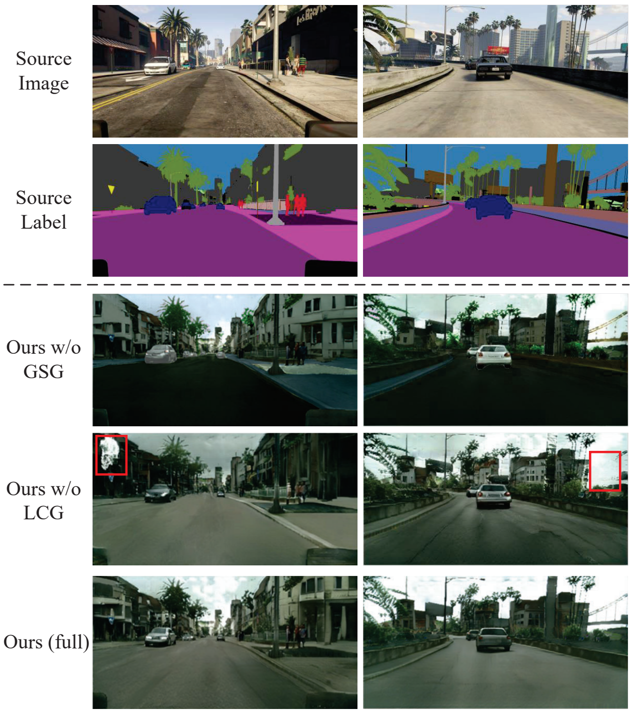
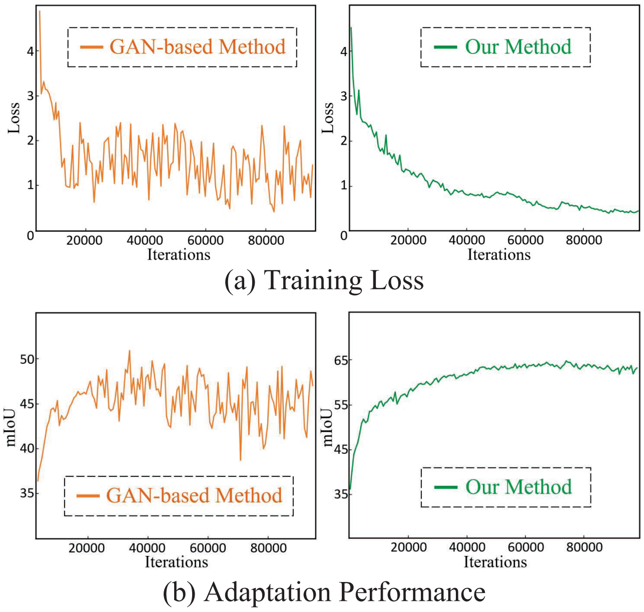


Figure 1: The visualization of translated images when removing the LCG and GSG modules, respectively.

A. Effect of LCG and GSG modules

In the main paper, we already demonstrate the effectiveness of LCG and GSG quantitatively. To provide a more comprehensive ablation study, we present additional qualitative results to visually illustrate the effect of each module. As shown in the third row of Fig. 1, the model without GSG produces translated images lacking global harmony, where objects such as cars and trees appear independent from their surroundings, indicating that GSG plays a key role in harmonizing the entire scene. In contrast, the model without LCG can ensure global harmony but cannot preserve the details well. As shown in the fourth row of Fig. 1, parts of the building and bridge are missing (marked with red boxes). This observation demonstrates the importance of LCG in preserving local details. By incorporating both modules, our approach achieves both global-harmony and local-precision image translation, as shown in the last row of Fig. 1. These qualitative results demonstrate the complementary nature of LCG and GSG and show their effects in achieving high-quality image translation results.



(a) Training Loss

(b) Adaptation Performance

Figure 2: The loss and performance curves of the GAN-based image translation method [1] and ours.

B. Training stability analysis

In Fig. 2 (a), we show the training loss curves of the state-of-the-art GAN-based image translation method [1] and our diffusion-based method, respectively. We can observe that our method exhibits a more stable training process. During the training process, we also use the translated images to train a target-domain segmentation model and then evaluate its adaptation performance on the target domain. As shown in Fig. 2 (b), we can see that our method achieves a more stable and higher adaptation performance than the GAN-based method, suggesting that our method is not only stable but also effective.

C. Ablation on data augmentation

As discussed in Section 4.2 of the main paper, our image translation framework is designed to handle noisy and masked images. To improve the model's robustness, we generate additional noisy and masked images for data augmentation. Specifically, we follow the Eqn. 2 in the main paper to add noise to the training data, and then mask the noisy images using binary masks provided by source labels. We randomly select 10% of the training data to execute the data augmentation. As shown in Tab. 1, the augmentation can bring a slight improvement of 0.2%∼0.3%.

Table 1: Ablation study on data augmentation.

	Model	G→C	S→C
(a)	Ours w/o augmentation	61.7	60.7
(b)	Ours	61.9	61.0

References

[1] Li Gao, Lefei Zhang, and Qian Zhang. Addressing domain gap via content invariant representation for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 7528–7536, 2021. 1