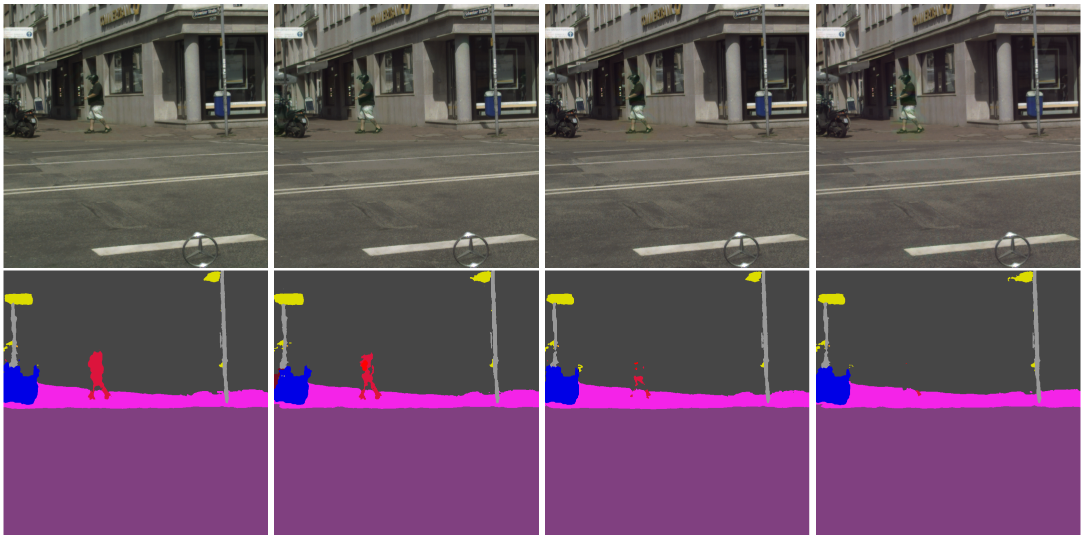

Figure 7. The relation between infinity norm and the corresponding type#1 performance. From left to right, the infinity norm is 4, 6, 8, 10 separately.

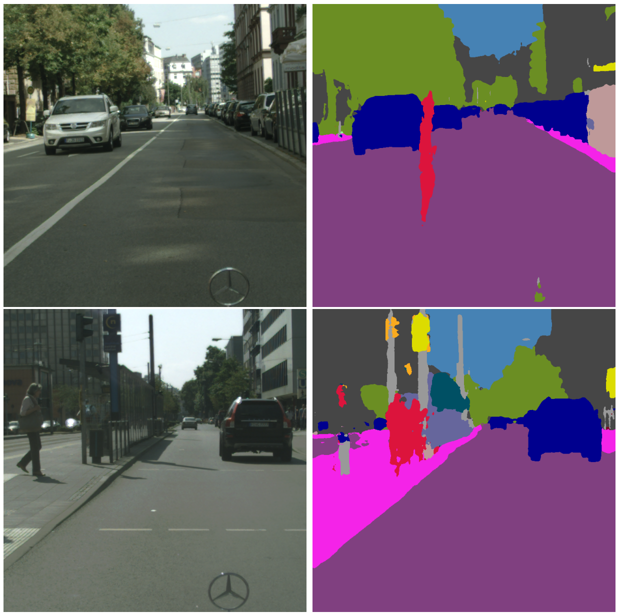

Figure 8. CityScapes Type#2 and Type#3 attack on DeepLabV3Plus model with MobileNet backbone. For Type#2 attack, the model learns to add a new set of 'person' labels in the prediction. Type#3 attack combines both Type#1 and Type#2 by vanishing 'person' and 'rider' labels into the background, and adding a new set of 'person' labels in the prediction. Top left: The input image. Top right: The input image + perturbations. Lower left: Predictions before attacks. Lower middle: Manipulated label mask. Lower right: Predictions after attacks.

6. Supplementary Material

6.1. Infinity norm of the perturbations

Although there have been lots of discussions on how the infinity norm affects the performance of adversarial attacks statistically, few of them are visualized. Here we select four infinity norm thresholds and visualize the corresponding performance on the same testing image. We choose to use DeepLabV3Plus-MobileNet as the target model. Figure 7 shows how different infinity norms, 4, 6, 8, 10 affect the performance of 'person' label vanishment. The results suggest that as the infinity norm increases, our model yields better performance against the target model.

6.2. Results of DeepLabV3Plus on CityScapes

In this session, we visualize samples of the ResNet-based models attack against the DeepLabV3Plus-MobileNet-based target for different attack types. Figure 9, Figure 11, Figure 8 show the result for type#1, type#2, type#3, respectively.

We also visualize a sample of type#1 cross-modal attacks on Cityscape. We have two models here. The first one is trained to attack a DeepLabV3Plus-MobileNet model. The second one is trained to attack a DeepLabV3Plus-ResNet model. Then these two target models are switched. Figure 12 shows the first model attacks against the second model's target while Figure 10 shows the second model attacks against the first model's target. The results are consistent with the success rate showed in Table 3.

Finally, we also give the result of attacking DeepLabV3Plus-MobileNet for type#2 and type#3 on Cityscapes, as Figure 8 shows.

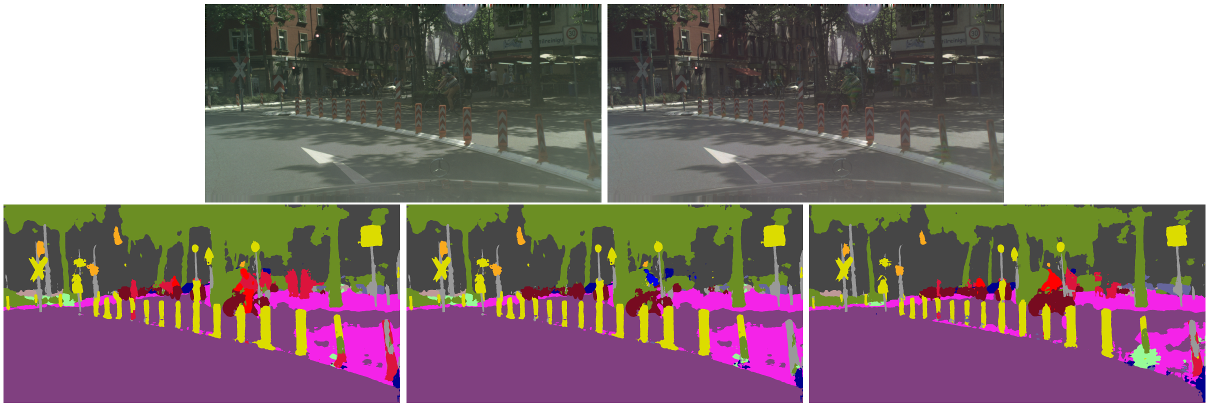

Figure 9. Type#1 attack against DeepLabV3Plus-MobileNet model on Cityscapes. The ‘person’ and ’rider’ labels vanish into the background. Top left: The input image. Top right: The input image + perturbations. Lower left: Predictions before attacks. Lower middle: Manipulated label mask. Lower right: Predictions after attacks.

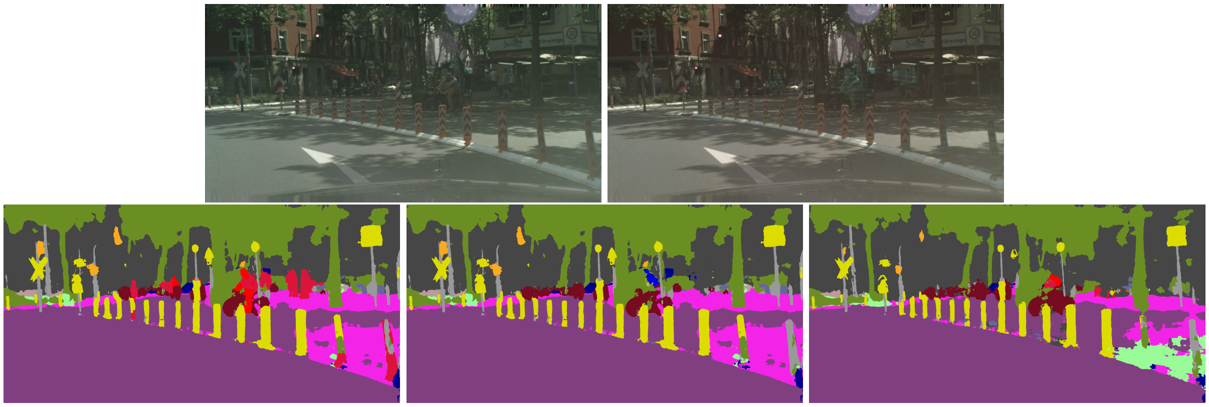

Figure 10. Type#1 attack against DeepLabV3Plus-MobileNet on Cityscapes. The ‘person’ and ‘rider’ labels vanish into the background. Our model is trained to attack DeepLabV3Plus-Resnet and evaluate on DeepLabV3Plus-MobileNet. Top left: The input image. Top right: The input image + perturbations. Lower left: Predictions before attacks. Lower middle: Manipulated label mask. Lower right: Predictions after attacks.

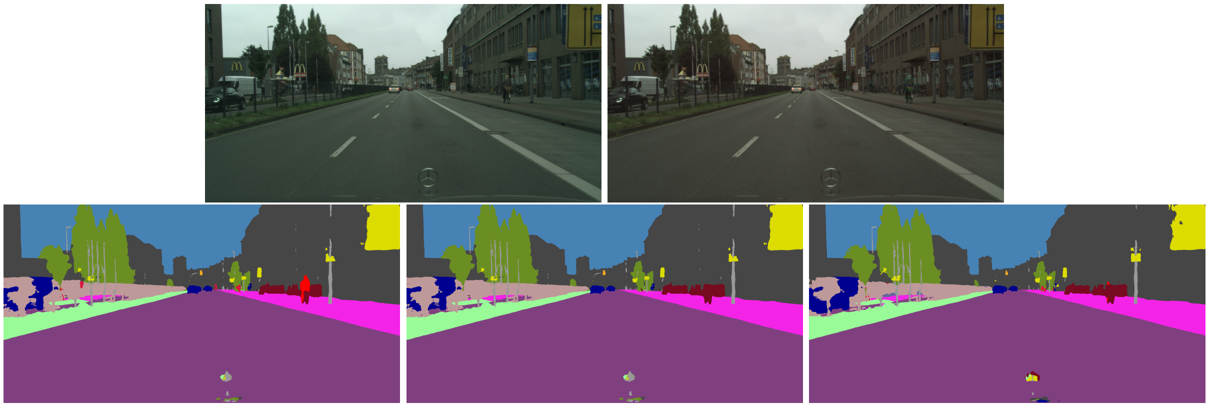

Figure 11. Type#1 attack against DeepLabV3Plus-ResNet on Cityscapes. The ‘person’ and ‘rider’ labels vanish into the background. Top left: The input image. Top right: The input image + perturbations. Lower left: Predictions before attacks. Lower middle: Manipulated label mask. Lower right: Predictions after attacks.

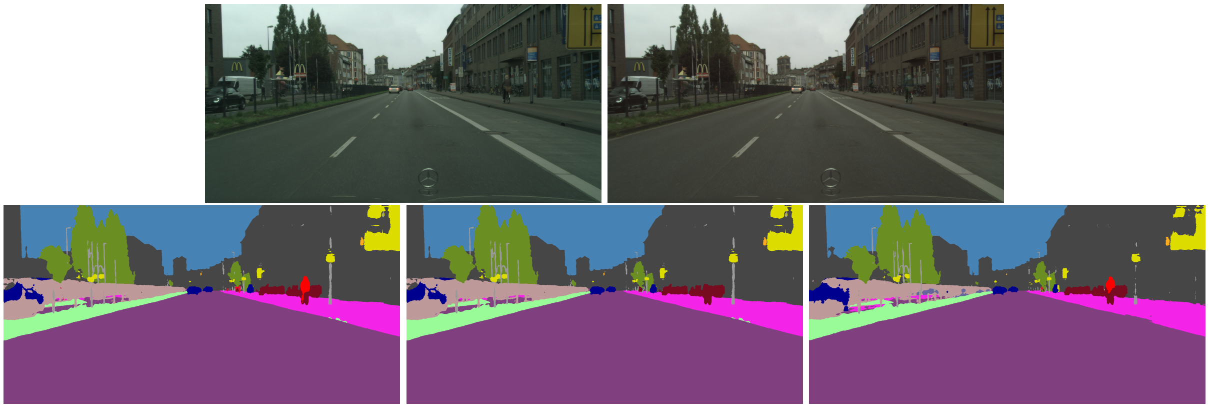

Figure 12. .Type#1 attack against DeepLabV3Plus-ResNet on Cityscapes The ‘person’ and ‘rider’ labels vanish into the background. Our model is trained to attack DeepLabV3Plus model with MobileNet backbone, and evaluate on DeepLabV3Plus model with ResNet backbone. Top left: The input image. Top right: The input image + perturbations. Lower left: Predictions before attacks. Lower middle: Manipulated label mask. Lower right: Predictions after attacks.

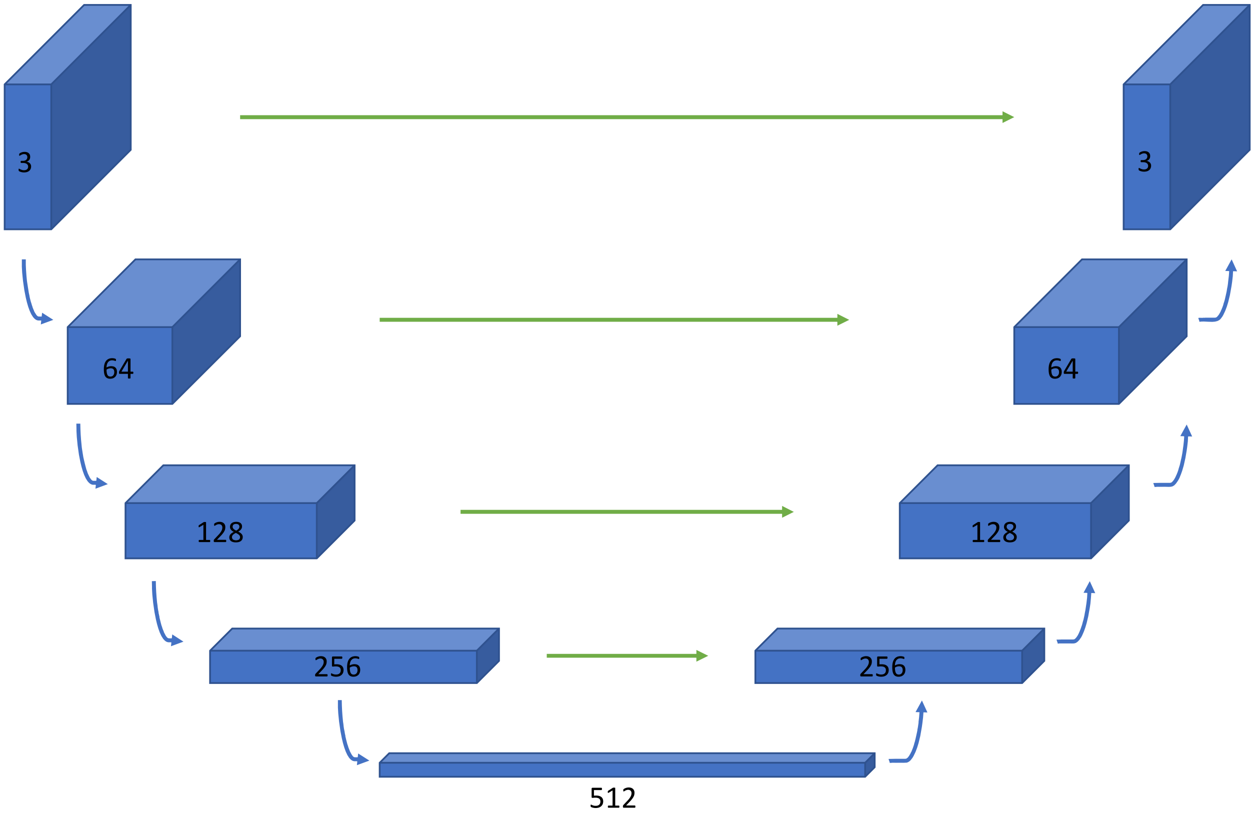

Figure 13. The detailed structure of the generator. 3, 64, 128, 256 are number of channels.