Exploiting CLIP Self-Consistency to Automate Image Augmentation for Safety Critical Scenarios

Supplementary Material

A. Evaluation - SD Checkpoints

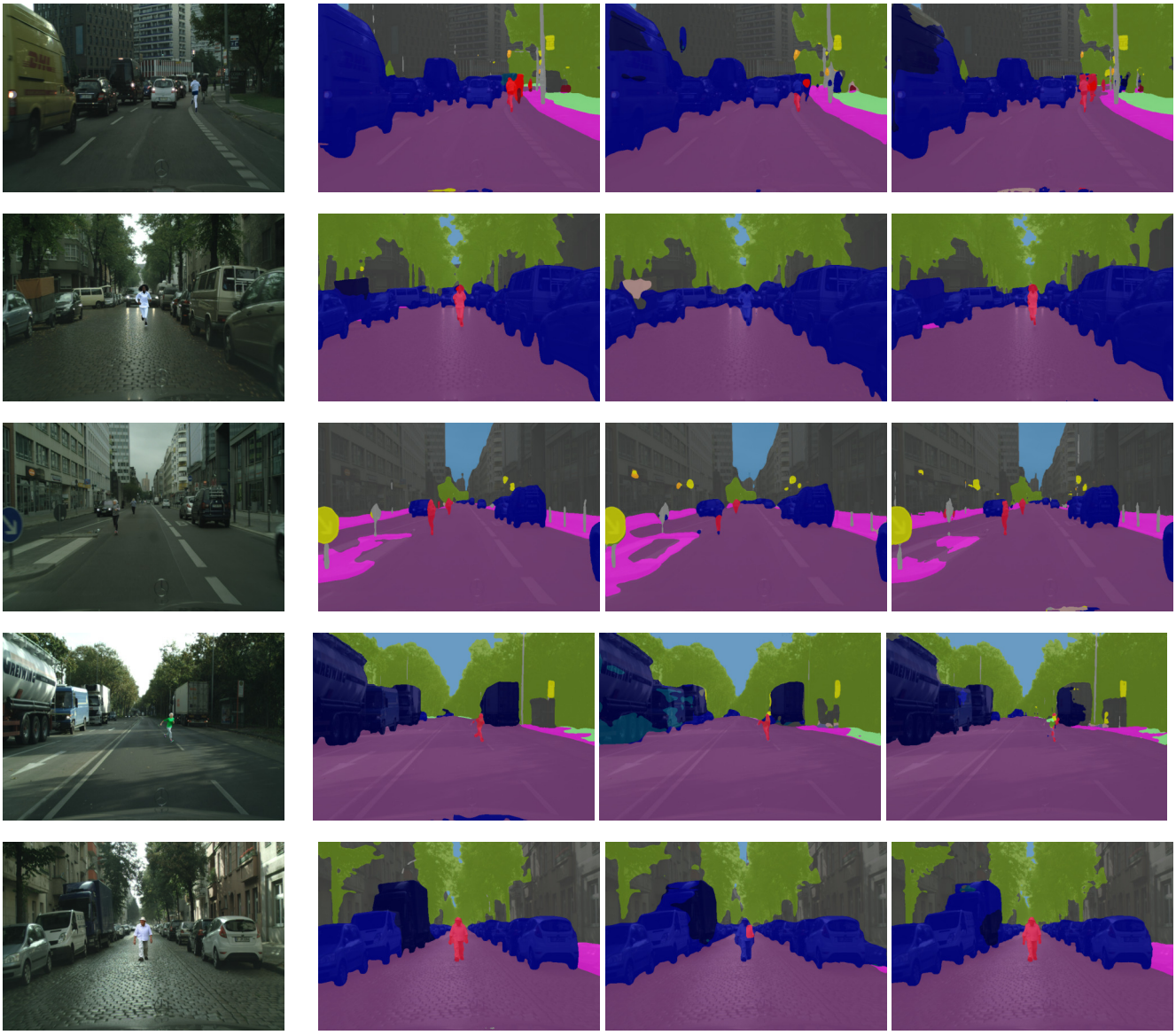
Within this supplemental material, we present further results on the comparison of the different model weights used for inpainting, namely *SD-v1.5*, *Reliberate-v2*, and *Deliberate-v5*. The qualitative results shown in Fig. 6 demonstrate the lack of realism of the *SD-v1.5* model as seen by the result in the second row. In comparison, there is no significant quality in inpatinng between *Reliberate-v2*, and *Deliberate-v5*

B. Inference on DNNs-under-test

In this section, we provide more segmentation inference results from sample images of the filtered augmented set with the three DNNs-under-test, *i.e.*, SETR (Strong), SETR (Weak), and ICNet. Qualitative evaluation, shown in Fig. 7, shows that while the strong model identifies the augmented pedestrians in most cases, the latter models have significant variations in the performance for just the augmented pedestrians while having reasonable performance over the other classes present in the image.

(a) Input Image (b) *SD-v1.5* (c) *Reliberate-v2* (d) *Deliberate-v5*

Figure 6. Samples of the augmented images with inpainted pedestrians using three different checkpoints using latent diffusion models [34]. Note that in some instances, *e.g.*, *SD-v1.5*, inpainting could completely fail.



(a) Input Image (b) *SETR Model (Strong)* (c) *SETR Model (weak)* (d) *ICNet*

Figure 7. Performance of three pre-trained models on images from the filtered augmented set.