Disharmony: Forensics using Reverse Lighting Harmonization

Supplementary Material

A. Results of Pretraining and Fine-Tuning

Fig. 6 illustrates the outputs generated by different training approaches: pretraining on DISK25K [24] only, pretraining on DISK25K followed by fine-tuning with the RealHM dataset [14], and pretraining on DISK25K followed by fine-tuning with both the RealHM and IH [4] datasets (the training approach used for our Disharmony).

When evaluating these models on a test set comprising various harmonization methods (DoveNet [8], Harmonizer [16], HT [13], Hi-Net [5], PCT-Net [11]) and composite images, it is evident that the model pretrained solely on DISK25K struggled to accurately detect the harmonized objects. However, when fine-tuned with the RealHM dataset, the model's performance improved, producing more refined results, though some false detections persisted.

In contrast, our proposed training approach, which involves pretraining on DISK25K and fine-tuning with both the RealHM and IH datasets, effectively eliminated these false detections and produced even more refined segmentation masks. This demonstrates the validity of our training methodology, showing that combining datasets that are neural network-based, physics-based, and handcrafted light-aware can yield superior segmentation results.

One might question why we did not train exclusively on the handcrafted (RealHM) and physics light-aware (IH) datasets. However, given their small sizes—216 and 60 images, respectively—compared to the 24,964 images in the DISK25K dataset, incorporating DISK25K was essential for achieving the comprehensive learning required for robust performance.

B. Ground Truth Comparison

To evaluate whether the forensic networks and Disharmony can correctly identify non-edited images, we tested them on the ground truth images from our generated test set. Our hypothesis was that if the images are not edited, the forensic networks should indicate the absence of edited regions. This experiment is significant, as previous studies have primarily focused on testing edited or generated image parts, leaving this aspect unexplored.

Fig. 7 presents examples of the output from various networks, including our forensic network, when ground truth images were used as input. As shown, all networks except HiFi-Net [12] incorrectly identified edited regions in the non-edited images, resulting in false positive masks. Although our network also produced some false positives, it demonstrated greater stability across the ground truth images compared to the other networks. HiFi-Net, however, failed to detect any regions, which we will further discuss in Sec. 4.1 and Sec. 4.2 , highlighting why this outcome is problematic.

C. Outputs from Text-Based Edits

As mentioned in Sec. 5, we generated test images using two text-based diffusion models, InstructPix2Pix [3] and Imagic [15], on the TEDBench dataset. At first glance, the edit instructions appear to be applied correctly, with the resulting images seemingly aligning with the provided prompts. However, upon closer examination, it becomes clear that the background, while retaining the general context of the original image, has undergone significant alterations, rather than just the specific parts intended for editing. This results in comprehensive changes to the image as a whole, rather than the targeted edits that were intended. Additionally, we evaluated the SSIM [27] and LPIPS [31] for these images, with the results presented in Tab. 3.

D. Outputs from Virtual Try edits

As mentioned in Sec. 5, we conducted preliminary tests of our network on virtual try-on edits, acknowledging that further improvements are needed. The results are illustrated in Fig. 9. In this experiment, StableVinton [17] was used to generate an edited image where the same woman is shown wearing different clothes. We also present the mask used for training the model. Visually, the network appears to detect the regions corresponding to the clothing, albeit with some noise. Additionally, the network erroneously identifies the human face as an edited part of the image, which we attribute to the fact that our model was not specifically trained on this dataset. Despite these limitations, the results suggest potential for future research, particularly in refining the model's accuracy for virtual try-on edits.

E. Outputs from Drag-Based edits

As discussed in Sec. 5, we conducted preliminary tests of our network on drag-based edits, recognizing that further improvements are necessary. In this experiment, DragDiffusion [23] was used to generate edited images based on user inputs. Upon running the network, we observed color shifting and blurring in some images, likely due to the diffusion process. However, our network was able to detect portions of the edited areas, although not consistently across the entire edited region. These findings suggest that there is potential for further exploration and refinement in this area. The results are illustrated in Fig. 10.

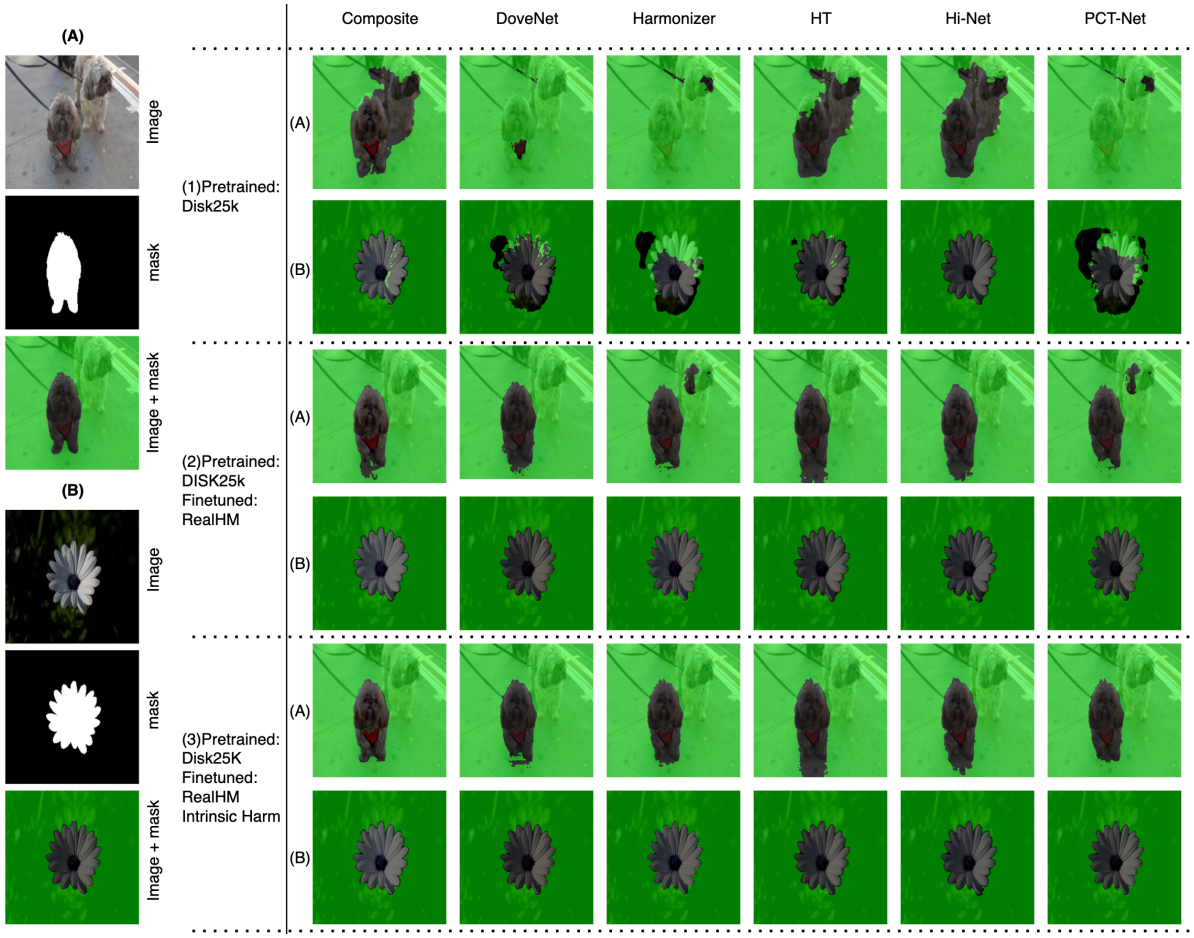

Figure 6. Evaluation of the aggregation method. We tested three different scenarios: (1) pretraining MaskFormer [6] with DISK25K [24], (2) pretraining with DISK25K followed by fine-tuning with the RealHM dataset [14], and (3) pretraining with DISK25K followed by fine-tuning with both the RealHM and IH dataset [4]. Based on these evaluations, we concluded that the most effective approach is to pretrain with DISK25K and fine-tune with both the RealHM and IH dataset.

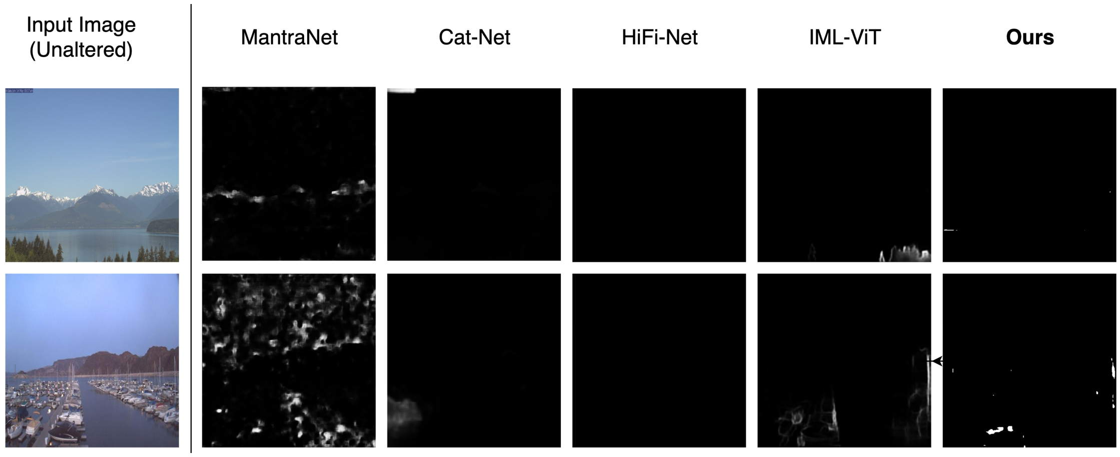

Figure 7. The resulting masks produced by various forensic models(MantraNet [28], Cat-Net [18], HiFi-Net [12], IML-ViT [20]) and Disharmony, given an unaltered input image

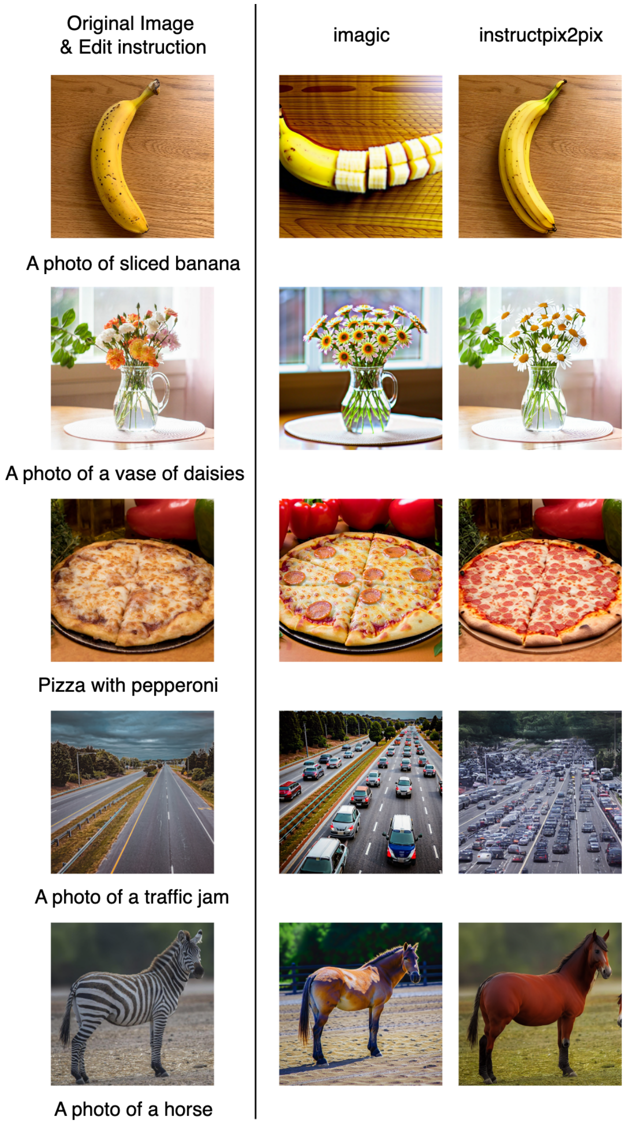

Figure 8. The resulting images produced by text edits using InstructPix2Pix [3] and Imagic [15] on the TEDBench dataset. Note that the backgrounds change even when the text instructions do not specify modifications to the background.

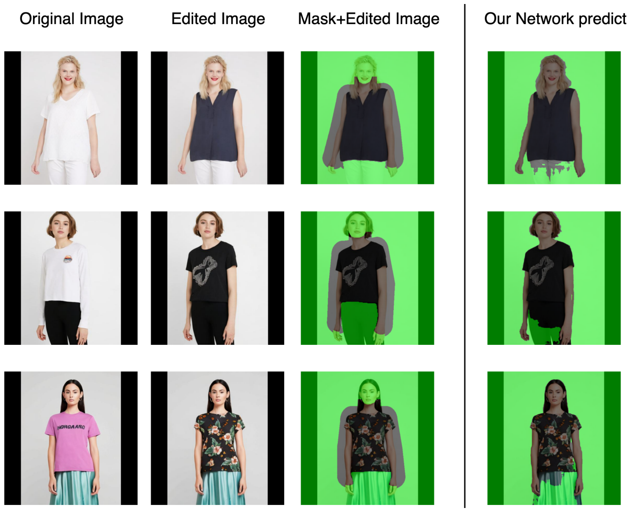

Figure 9. The resulting segmentation outcomes for images that have been edited using StableViton [17]

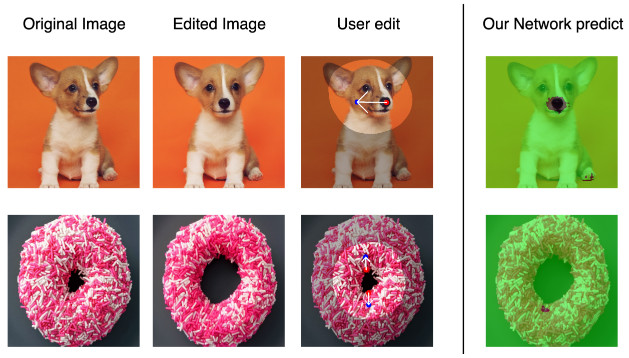

Figure 10. The resulting segmentation outcomes for images that have been edited using DragDiffusion [23]