1. Limitations

The training data we used is class-agnostic, and as a result no semantic information was considered explicitly (though maybe implicitly), which may restrict the potential of the proposed framework to be applied to other semantic-related tasks.

2. Comparison with the Dense-CL Method

To verify our effectiveness, we compare our DPT with DenseCL [2] based on GCA [1]. Since the backbone of GCA is different from that of DenseCL, for a fair comparison, we replace the backbone with the ResNet50, and perform fine-turning after loading pre-trained weights from DenseCL. As shown in Table 1, our DPT achieves better performance with fewer parameters.

3. Ablation on losses

In our DPT, we adopt three kinds of losses, L1 regression loss, Composition loss and Laplacian loss. The impact of these loss functions ($L_{l1}$, $L_{lap}$ and $L_{comp}$) on the final performance is shown in Table 2.

4. More fine-tuning qualitative results on Composition-1k

We fine-tune on Composition-1k with DPT initialization. As Fig. 1, Fig. 2, Fig. 3, Fig. 4 show, we provide more qualitative results of our DPT on the Composition-1k test set. From the results, it can be seen that our method performs well in various detailed foreground objects.

5. More fine-tuning qualitative results on Distinct-646

As Fig. 5 – Fig. 8 show, we further present more qualitative results of our DPT on Distinct-646 [1]. We first generate trimap by randomly dilating alpha mattes from the ground truth alpha matte with a threshold of 20 (following [3]). Then we use the fine-tuning model on Composition-1k to test directly on Distinct-646. We can see that our DPT achieves good performance, especially in detailed regions.

6. More fine-tuning qualitative results on Natural human images

To show the robustness of our method, we use a fine-tuning model on Composition-1k to test on 2k resolution natural human images as shown in Fig. 9. From left to right, we provide the image, generated alpha matte $\alpha$, and the synthetic foreground generated by $\alpha$ and image. We can see that our method performs well at extracting the boundary of the human body, such as the hair part.

References


Figure 1. Qualitative results of our method on the Composition-1k test set.
Figure 2. Qualitative results of our method on the Composition-1k test set.
Figure 3. Qualitative results of our method on the Composition-1k test set.
Figure 4. Qualitative results of our method on the Composition-1k test set.

pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3024–3033,
Figure 5. Qualitative results of our method on the Distinct-646 test set.
Figure 6. Qualitative results of our method on the Distinct-646 test set.
Figure 7. Qualitative results of our method on the Distinct-646 test set.

Figure 8. Qualitative results of our method on the Distinct-646 test set.
Figure 9. Qualitative results of our method on natural human images.