In-Context Matting

Supplementary Material

1. Overview

The supplementary material includes the following sections:

- Implementation details of IconMatting.
- Additional experiments.
- Dataset.
- Implementation of a baseline-In-Context Pipeline.

2. Implementation Details of IconMatting

For hyperparameters and detailed information on the model, please refer to the code in the supplementary material.

Here, we supplement schematic diagrams for the inter- and intra-similarity modules. The inter-similarity computes the similarity between features extracted from the target image and the in-context query derived from the reference image, generating an average similarity map, S. The intra-similarity combines the self-attention maps representing intra-image similarities within the target image with the similarity map S obtained from the inter-similarity module. This fusion employs elements of S as weights assigned to the self-attention maps, thereby providing guidance information for the matting target.

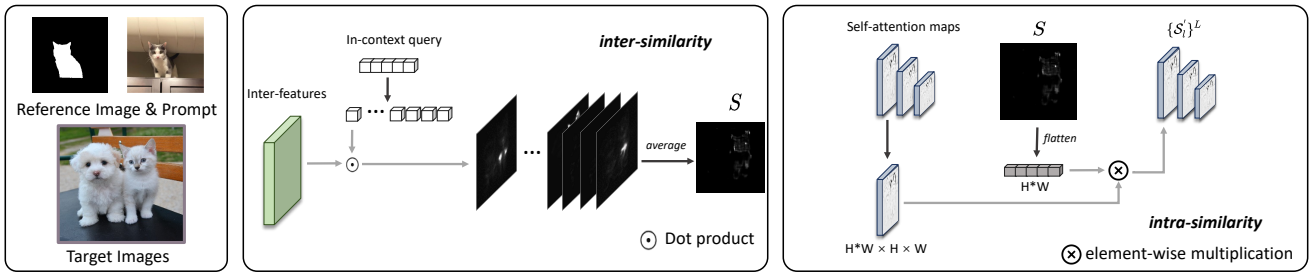

Figure 1. **Illustration of the inter- and intra-similarity modules.** For simplicity, the resize operation is omitted, only the calculation of one element of the in-context query is depicted, and the fusion process of self-attention maps from a single scale is shown.

3. Additional Experiments

3.1. More Qualitative Results

Here, we visualize more qualitative results of IconMatting in real-world scenarios, as shown in Fig. 2.

3.2. Why Composited Dataset is Not Used?

Our IconMatting, developed as a model for image matting, underwent exclusive training solely on real-world datasets, omitting composited data—a practice uncommon within the domain of image matting. This methodology was adopted due to the substantial discrepancy observed in model performance between composited and real-world datasets when employing Stable Diffusion as the backbone. To substantiate this assertion, we conducted experiments on Composition-1k and RM-1k datasets. For IconMatting, we modified it to be a trimap-based image matting model. First, we removed the in-context similarity module of IconMatting. Second, we concatenated the trimap with the original image and feed them together into the matting head. The training and test sets of Composition-1k were pre-defined, while RM-1k underwent training and test set partitioning in an 8:2 ratio.

The experimental findings, as depicted in Tab. 1, reveal a significant performance gap between RM-1k and Composition-1k datasets. Despite the considerably smaller sample size in RM-1k compared to Composition-1k, the former demonstrated notably superior test results. This discrepancy highlights that, Stable Diffusion, pre-trained for generative tasks, does not perform optimally for image matting tasks on composited datasets within the context of our study.

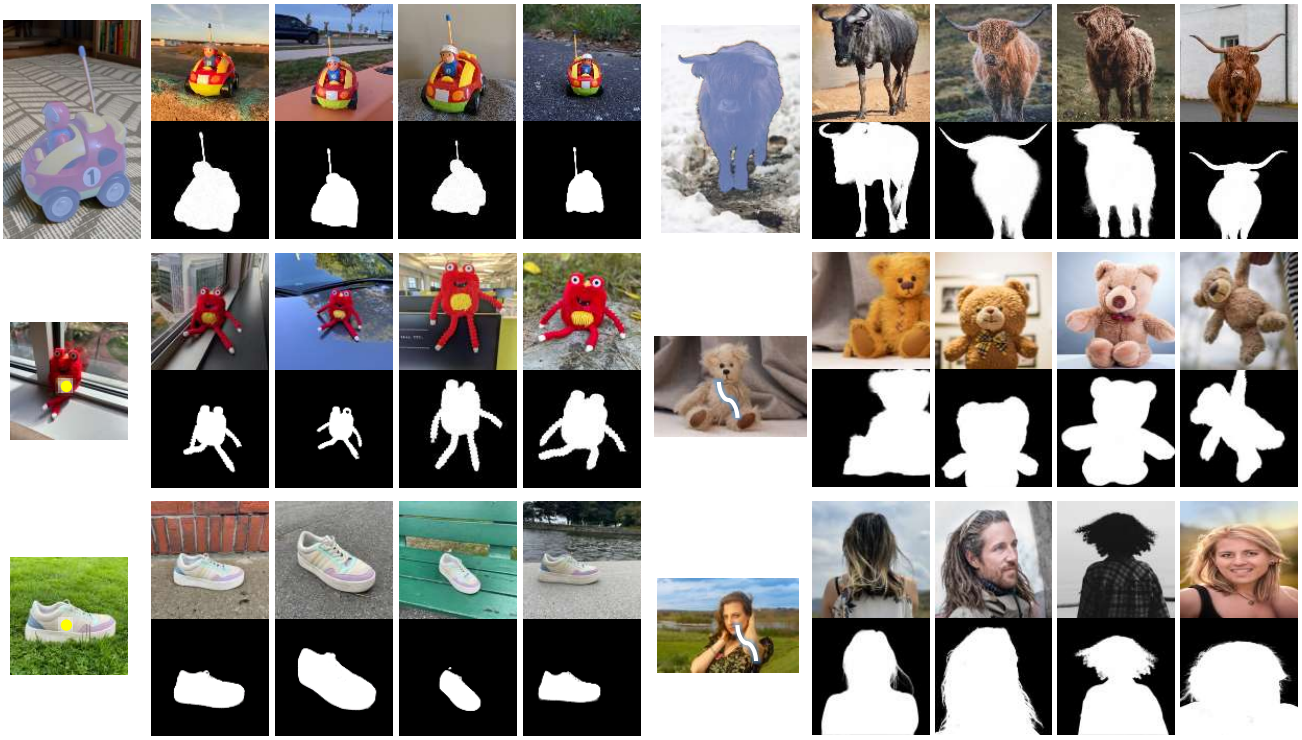

Figure 2. **Qualitative results of IconMatting.** The first column shows the reference input, while the remaining columns display target images and their predicted alpha mattes. Given a single reference input, our method can automatically process the same instance or category of foreground.

Dataset	MSE	SAD	GRAD	CONN
Composited dataset	0.0227	110.12	68.84	59.02
Real-world dataset	0.0029	14.11	16.32	13.56

Table 1. **Experiment on composited dataset and real-world dataset.**

3.3. Ablation on Diffusion Time Steps

When using Stable Diffusion as a feature extractor, the choice of the time step t during the diffusion process is crucial. Generally, a small t corresponds to features that represent the image more closely, with minimal noise added, which is why methods like ODISE [4] for image segmentation use $t = 0$. Conversely, a large t captures more abstract semantic features, as seen in DIFT [3] for semantic correspondence using $t = 261$.

For IconMatting, we aim to extract features that both express abstract semantics in the inter-similarity module and retain detailed characteristics of the image for predicting the alpha matte with the matting head. Therefore, selecting the appropriate t is a

trade-off. As shown in Table 2, our experiments show that performance is suboptimal for rather small t values (*e.g.*, 0) or too large t values (*e.g.*, over 300). The optimal performance is achieved when t is set to 200.

3.4. Visualization on In-Context Similarity

To further illustrate the effectiveness of our core module, the in-context similarity, we visualize both inter- and intra-similarity modules. For inter-similarity, we directly visualize S; for intra-similarity, we resize multi-scale $\{\mathcal{S}_l^{'}\}^L$ to a uniform scale, averaged it, and then visualized the result. This is demonstrated in the Fig. 3.

In the case of an alarm clock and a monkey, due to some differences between the reference input and the matting target, the lower left part of the matting target is lost in the results of inter-similarity. However, by relying on the intra-similarity, the

Timesteps	MSE	SAD	GRAD	CONN
0	0.0091	22.12	20.18	10.98
100	0.0094	21.01	19.54	10.94
200	0.0081	19.12	18.65	11.21
300	0.089	23.84	20.11	12.46
400	0.098	31.39	24.57	14.28

Table 2. **Ablation study on the choice of diffusion timesteps.**

results of intra-similarity complement and complete the matting target, thus predicting a complete alpha matte.

This analysis underscores the significance of considering both inter- and intra-similarity in our approach.

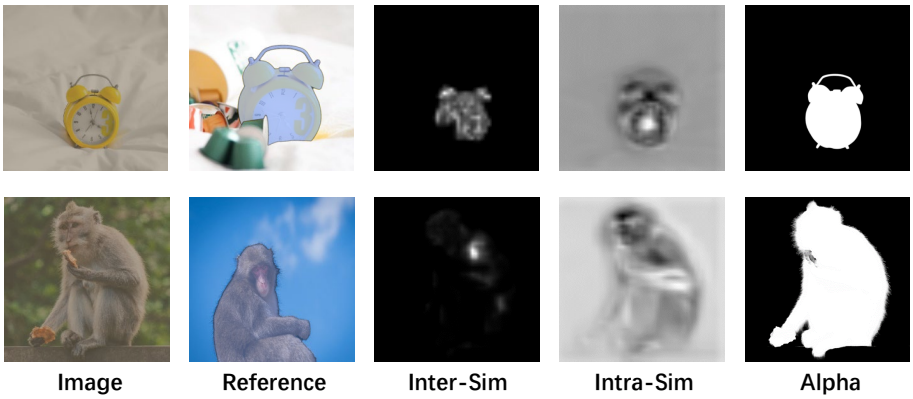

Figure 3. **Visualization on in-context similarity.**

4. Dataset

Figure 4. **ICM-**57 encompasses foregrounds of both the same category and same instance.

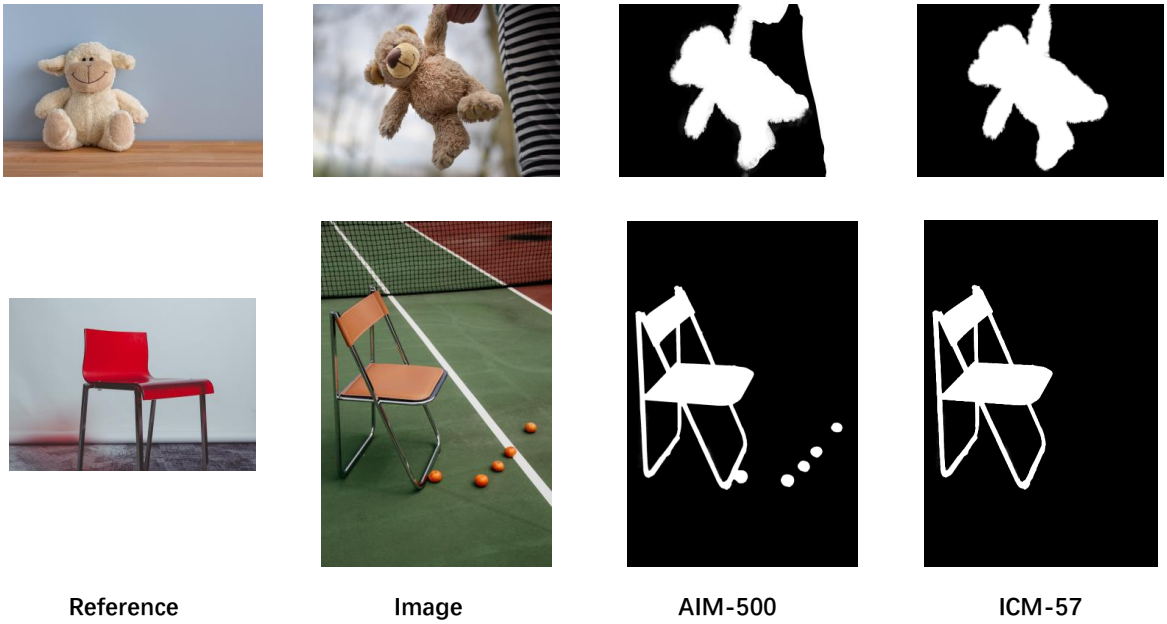

Figure 5. **Modifications made to the annotations.** In the second row of pictures, only the chair is preserved in the alpha matte, which meets the needs of in-context matting.

To complete the training and testing of the model, we constructed a mixing training set and a test set.

For our test set ICM-57, as described in the main text, it encompasses foregrounds of the same category and same instance, fulfilling the essence of in-context matting. Examples of such instances are depicted in Fig 4. In order to utilize a portion of images from AIM-500, we modified their annotations to align with the requirements of in-context matting, as illustrated in Fig. 5.

5. In-context Matting Pipeline

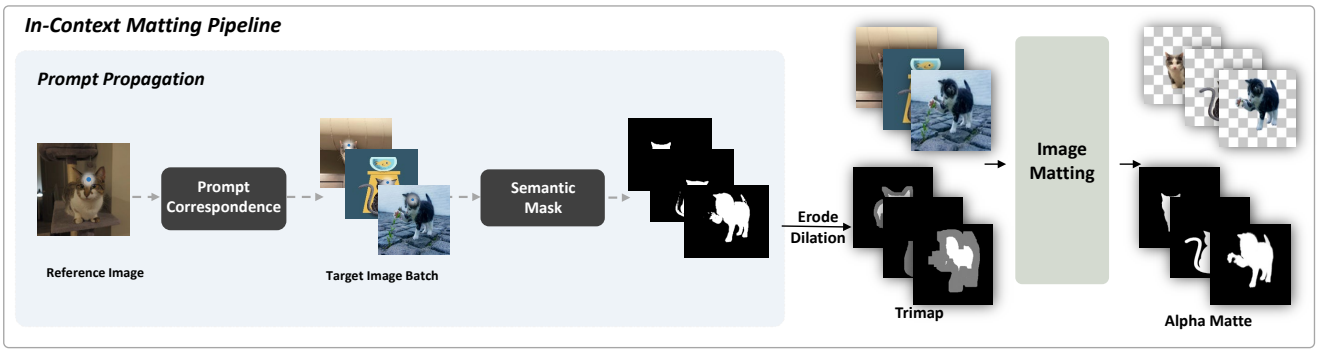

Figure 6. **In-Context Matting Pipeline (ICMP)** consists of two parts: prompt propagation module and interactive image matting module. Prompt propagation module generates prompts for target images based on the prompts on reference image through semantic correspondence, then interactive image matting module predicts alpha matte with images and prompts.

In our novel task setting for image matting, there wasn't an existing baseline available for direct comparison with our IconMatting. Therefore, leveraging some off-the-shelf models, we established a pipeline aimed at achieving in-context matting, serving as a baseline in the experiment in the main text.

In-Context Matting Pipeline (ICMP) consists of two modules: a prompt propagation module and an image matting module, as shown in Fig. 6. With ICMP, users provide points as prompts on reference images to indicate matting targets, then the

prompt propagation module extends these prompts to all input images and the semantic masks of the corresponding matting targets are obtained. Subsequently, the image matting module processes the input images and corresponding semantic masks, generating alpha mattes for the specified matting targets across all images.

We utilize the semantic correspondence property of features provided by DINOv2 [2] to achieve prompt correspondence, and use SAM [1] to extract coarse semantic masks corresponding to the prompt points, thus realizing our designed prompt propagation, and thus ICMP.

5.1. Prompt Propagation Module

While auxiliary input-based image matting can yield the alpha matte for a user-specified matting target, it requires manual prompting for each input image, even when the matting target is the same. Prompt propagation addresses this issue by disseminating the prompts provided on the example image to all input images, resulting in a set of semantic masks.

In essence, prompt propagation can be likened to semantic correspondence, wherein prompts from the example image are matched to corresponding prompts in other images. By prompt propagation, the user's prompt for the example image can be propagated to the other images, eliminating the need to manually provide a prompt for each image.

Considering that features extracted by DINOv2, a model pretrained on large-scale datasets, exhibit strong generalization to real-world data for semantic correspondence without further training, we employ DINOv2 to propagate prompt points for the example image to other images. Given an input image and prompt points indicating the matting target, SAM can be used to generate a semantic mask corresponding to the prompt.

5.2. Image Matting Module

This module enables the extraction of the alpha matte for the matting target based on semantic masks from prompt propagation.

By applying morphological operations to the mask, we transform a semantic mask from prompt propagation into a pseudo-trimap. Then, any trimap-based image matting model can be used to obtain a alpha matte for the matting target, and our choice is VitMatte [5].

References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 5

[2] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5

[3] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 2

[4] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2

[5] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pre-trained plain vision transformers. *Information Fusion*, page 102091, 2023. 5