Training Matting Models Without Alpha Labels

Wenze Liu[1], **Zixuan Ye**[2], **Hao Lu**[2*], **Zhiguo Cao**[2], **Xiangyu Yue**[1*]

[1] MMLab, The Chinese University of Hong Kong
[2] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology

Abstract

The labeling difficulty has been a longstanding problem in deep image matting. To escape from fine labels, this work explores using rough annotations such as trimaps coarsely indicating the foreground/background as supervision. We present that the cooperation between learned semantics from indicated known regions and proper assumed matting rules can help infer alpha values at transition areas. Inspired by the nonlocal principle in traditional image matting, we build a directional distance consistency loss (DDC loss) at each pixel neighborhood to constrain the alpha values conditioned on the input image. DDC loss forces the distance of similar pairs on the alpha matte and on its corresponding image to be consistent. In this way, the alpha values can be propagated from learned known regions to unknown transition areas. With only images and trimaps, a matting model can be trained under the supervision of a known loss and the proposed DDC loss. Experiments on AM-2K and P3M-10K dataset show that our paradigm achieves comparable performance with the fine-label-supervised baseline, while sometimes offers even more satisfying results than human-labeled ground truth.

Code — https://github.com/poppuppy/alpha-free-matting

Introduction

Image matting, a fundamental task in image editing, aims to decompose an input image $\boldsymbol{I}$ into two layers, *i.e.*, the foreground $\boldsymbol{F}$ and the background $\boldsymbol{B}$. Specifically, it estimates the foreground opacity $\boldsymbol{\alpha}$ a.k.a. alpha matte as

$$\boldsymbol{I} = \boldsymbol{\alpha}\boldsymbol{F} + (1-\boldsymbol{\alpha})\boldsymbol{B}\,. \tag{1}$$

Recent years have witnessed dramatic improvement in image matting techniques (Hou and Liu 2019; Li and Lu 2020; Liu et al. 2021a,b; Tang et al. 2019; Park et al. 2022; Yao et al. 2023; Yu et al. 2021), particularly since Deep Image Matting (DIM) (Xu et al. 2017) pioneered an end-to-end training paradigm. Despite the prosperity of the labeling-and-training paradigm, large-scale datasets are difficult to collect for the subtle details required. Existing datasets either composite (Xu et al. 2017) few annotated foreground objects with thousands of backgrounds, or make small special-purpose (Li et al. 2022, 2021) datasets. Although the deficiency of datasets is regarded as a bottleneck of deep image matting, the contradiction between the need of large-scale datasets and the labeling difficulty has yet to be handled.

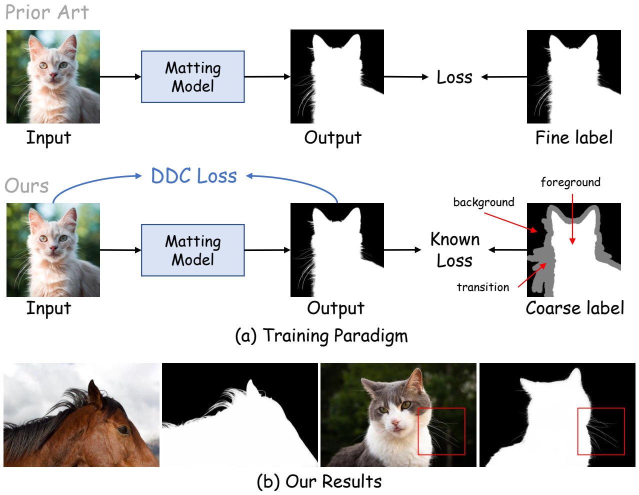

Figure 1: **Our training paradigm.** (a) Compared to prior art with alpha matte labels, we use only coarse trimap labels. We use an L1 loss to supervise the known regions indicated by the trimap, and devise a DDC loss to infer the alpha values at transition areas. (b) Even without fine annotated labels, our trained model predicts accurate alpha mattes.

Solving $\boldsymbol{\alpha}$ from Eq. (1) has not always been relying on labels. Aided by user prompts, *e.g.*, trimaps indicating the known (1 for foreground and 0 for background), and the unknown (0.5 for transition) regions as shown in Fig. 1, many works (Chuang et al. 2001; Sun et al. 2004; Levin, Lischinski, and Weiss 2007; He, Sun, and Tang 2010; Lee and Wu 2011; Chen, Li, and Tang 2013) try to trace the color cues to infer alpha values from known to unknown. For example, Closed Form Matting (Levin, Lischinski, and Weiss 2007) assumes the color line model in each neighborhood to figure out the alpha matte. Nonlocal Matting (Lee and Wu 2011) borrows the nonlocal principle from image denoising (Buades, Coll, and Morel 2005, 2008) to solve the alpha value with affinity matrices. These traditional methods tend to be impaired in the face of complex textures or similar colors on the junction of foreground and background (Xu et al.

[*] Corresponding authors.

2017), because they primarily consider low-level color or texture information. In contrast, deep learning based methods inject semantic understanding to adapt the models to complex scenarios, resulting in a significant performance improvement. Along the torrent, recent studies (Li, Zhang, and Tao 2021; Li et al. 2022; Qiao et al. 2020; Zhang et al. 2019; Li et al. 2021; Wei et al. 2021; Yang et al. 2022) prove that strong semantics introduced by deep models can reduce the usage costs, either by simplifying user prompts (Wei et al. 2021; Yang et al. 2022) or exploring automatic image matting (Li, Zhang, and Tao 2021; Li et al. 2022; Qiao et al. 2020; Zhang et al. 2019; Li et al. 2021) with no user prompt required. Different from research efforts devoted on the user end, we would like to study whether a collaborative approach involving robust semantic understanding and well-defined matting rules can mitigate the need for fine annotations during the training phase.

In this work, we demonstrate that, with proper supervision, deep image matting models can be effectively trained without any fine annotations. For one thing, deep models are validated (Long, Shelhamer, and Darrell 2015; Xiao et al. 2018; Kirillov et al. 2019) to be expert at learning the semantics from coarse labels. For another, traditional solutions (Levin, Lischinski, and Weiss 2007; Chen, Li, and Tang 2013) have proven that proper assumed rules can propagate the matting constraints from known to unknown regions. Based on these existing experiences, we conjecture that the labels only need to provide rough semantics, while not necessary to indicate the transparency at transition areas. We thus build a preliminary matting supervision, where an L1 loss termed known loss supervises the known areas indicated by the trimap, and the unknown ones are supervised using the nonlocal principle. The preliminary experiments suggests that i) the deep model can utilize the information learned from known regions to infer unknown ones, but in the sense of hard segmentation, and ii) the additional nonlocal assumption helps produce some details at transition areas. Based on further observation and analysis, we find that the embodiment of nonlocal principle in (Lee and Wu 2011; Chen, Li, and Tang 2013) is unsuited for the deep learning process, due to the 'braking effect' and 'hard segmentation effect'. The former impedes the refinement of details with long-range dependency, while the latter hurts smooth transition of alpha values at boundaries. Aiming at the two problems, we introduce a novel expression of nonlocal principle as the loss function, called *distance consistency loss* (DC loss). DC loss forces the euclidean distance between each pixel and several similar neighbors in the predicted alpha matte to be equal with that in the image. DC loss well addresses the two issues above, but introduces undesired texture noise in interior regions, where it is incompatible with known loss. With a detailed analysis, we update DC loss with *directional distance consistency loss* (DDC loss), solving the conflict between the two losses.

In summary, we establish a novel training paradigm for image matting based on the proposed DDC loss as shown in Fig. 1. With only trimaps as labels, the model trained under our proposed paradigm can predict fine details, *e.g.*, the cat beard as shown on the right of Fig. 1. Experiment results on animal (Li et al. 2022) and portrait (Li et al. 2021) matting datasets show that our models performs comparably with the baseline supervised by fine alpha labels, which verifies the feasibility and effectiveness of the proposed paradigm. Further more, we provide detailed illustration and analysis to shed light on the working principle of our approach.

Related Work

Traditional Image Matting (Chuang et al. 2001; Levin, Lischinski, and Weiss 2007; Lee and Wu 2011; He, Sun, and Tang 2010; Chen, Li, and Tang 2013; Aksoy, Ozan Aydin, and Pollefeys 2017) utilize prior knowledge into Eq. (1) to propagate the user specified constraints towards the alpha matte solution. Bayesian Matting (Chuang et al. 2001) achieves this by color sampling. Closed Form Matting (Levin, Lischinski, and Weiss 2007) assumes certain smoothness, *i.e.*, the color line model, based on which a matting Laplacian matrix is derived. It yields impressive results when the color line model assumption holds. In order to relax the conditions, Nonlocal Matting (Lee and Wu 2011) introduces nonlocal principle, a conception originally adopted in image denosing (Buades, Coll, and Morel 2005, 2008) tasks. Nonlocal Matting expresses the nonlocal principle via an affinity matrix to form a constraint condition, which is then combined with the user constraint to solve the alpha matte. Follow up work such as Fast Matting (He, Sun, and Tang 2010) and KNN Matting (Chen, Li, and Tang 2013) studies the construction of the affinity matrix. Our work takes in matting priors, but for building a feasible training mode without alpha labels in the deep matting era.

Deep Image Matting, first presented in (Shen et al. 2016; Cho, Tai, and Kweon 2016; Xu et al. 2017), has opened a new era for modern image matting methods, with labeled data involved. Early work (Lu et al. 2019; Hou and Liu 2019; Li and Lu 2020; Liu et al. 2021a,b; Tang et al. 2019; Park et al. 2022; Yao et al. 2023; Yu et al. 2021; Cai et al. 2022) takes both the image and a trimap indicating the smooth/-transition areas as the input, and mainly studies the network architecture to better adapt the high-level semantics and low-level details for the matting task. In the interests of reducing the user efforts, there recently appears some attempts that replace the auxiliary trimap input with simpler prompts such as scribble (Xiao et al. 2018) and click (Wei et al. 2021). Another branch of work (Li, Zhang, and Tao 2021; Li et al. 2022; Qiao et al. 2020; Zhang et al. 2019; Li et al. 2021; Ma et al. 2023; Ke et al. 2022; Chen et al. 2018) studies to fully get rid of auxiliary inputs, leading to an automatic matting manner, where the model only takes in the image and needs to find salient objects itself. Besides simplifying the inputs from the user side, several works improve the deficiency of datasets, but still face problems of unrealistic content (Xu et al. 2017) and few categories (Li et al. 2022, 2021). There is also attempt (Liu et al. 2020) to add coarse training labels to lift the matting performance. With the same spirit of alleviating the data deficiency, we instead set out from lightening of annotation burdens–explore to supervise the model with coarse trimaps. To this end, our work pays special attention to loss design. There has been L1 loss, compositional loss (Xu et al. 2017), Laplacian loss (Niklaus and Liu 2018),

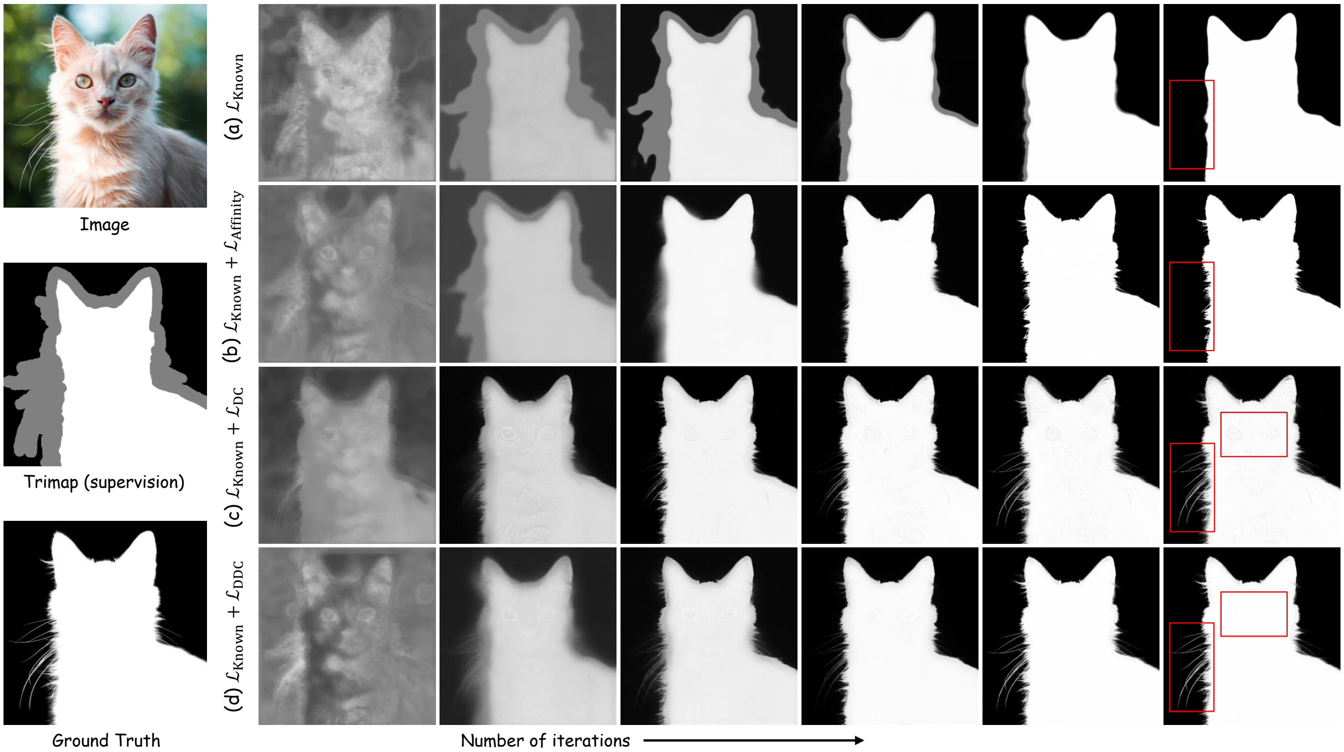
Image
Trimap (supervision)
Ground Truth
(a) $\mathcal{L}_{\text{Known}}$
(b) $\mathcal{L}_{\text{Known}} + \mathcal{L}_{\text{Affinity}}$
(c) $\mathcal{L}_{\text{Known}} + \mathcal{L}_{\text{DC}}$
(d) $\mathcal{L}_{\text{Known}} + \mathcal{L}_{\text{DDC}}$
Number of iterations

Figure 2: **The outputs during training of four supervision policies.** (a) The model can learn to extend the semantics from unknown to known with known loss. (b) The cooperation between known loss and affinity loss helps predict details, but fails to delineate long hair and causes hard segmentation. (c) The proposed DC loss well fits long hair and smooth transition on boundaries, but introduces texture noise on the foreground. (d) DDC loss eliminates the interior noise with no side effects.

etc. used for deep image matting with alpha labels, while we will introduce ones customized for coarse trimap labels.

Method

We first study the possibility of using trimaps as labels during training deep matting models. In what follows, we analyze the deficiency of the form of nonlocal principle in prior art for deep learning process. Then, we present our novel expression of nonlocal principle, and introduce our DC loss. At last, we solve the conflict problem between known loss and DC loss by updating DC loss to DDC loss.

Exploration and Analysis

We adopt the recent image matting model ViTMatte-S (Yao et al. 2023) and the AM-2K dataset (Li et al. 2022) for preliminary experiments. Detailed training configurations can be found in the experiment section.

Learning semantics from trimaps. We use an L1 loss termed known loss to supervise only the known regions indicated by 0 and 1 in trimaps. An instance on the evolution process of the output during training is shown in Fig. 2 (a). The model quickly finds the foreground object in early stage, then learns to fit the shape of the trimap, and finally the unknown region shrinks to disappearance. It explains that the learned semantics can be extended to unsupervised regions. However, known loss can only provide low-quality segmentation results, and can not generate details.

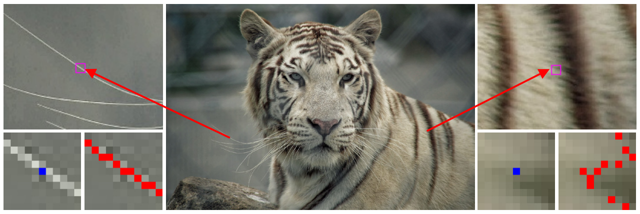

Figure 3: **Similar pixels in a local window.** Centered at a pixel (blue) in the image, the top K similar pixels (red) are selected in each $K \times K$ local window (pink).

Learning alpha values at unknown regions. We tried to add constraints with nonlocal principle (Lee and Wu 2011). In image matting, the nonlocal principle claims that the alpha value α_i of a pixel $\boldsymbol{I}_i$ is a weighted sum of alpha values of pixels similar to $\boldsymbol{I}_i$, formulated by $\boldsymbol{A}\boldsymbol{\alpha} = \boldsymbol{\alpha}$. Note that $\boldsymbol{\alpha}$ is flattened to $N \times 1$, where N indicates the number of pixels, and $\boldsymbol{A}$ is the affinity matrix of size $N \times N$. In KNN Matting (Chen, Li, and Tang 2013), $\boldsymbol{A}$ calculates

$$\boldsymbol{A}(i,j) = 1 - \frac{\|\boldsymbol{I}_i - \boldsymbol{I}_j\|_2}{C}, j \in \text{argtopk}\{-f(i,j)\}, \quad (2)$$

and then row-normalized, where C is a constant to make

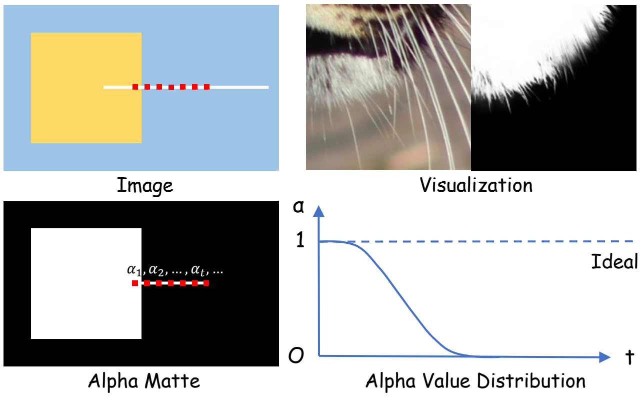

Figure 4: **The growth arrest problem of long hair supervised by affinity loss.** Suppose there is a yellow tiger with a hair whose intensity is a constant in the image. We derive that under such conditions alpha values of approximate linear or quadratic variation produce small loss. Considering that the already generated coarse mask provides the initial starting and ending value as 1 and 0, the distribution of alpha values are stuck at a state shown in the bottom right plot.

$\boldsymbol{A}(i,j) \in [0,1]$, f is set as $f(i,j) = \|\boldsymbol{I}_i - \boldsymbol{I}_j\|_2 + d_{ij}^2$, and d_{ij} denotes the spatial distance between i and j. Though Eq. (2) takes the whole image into account, the item d_{ij} limits the high response pixels within a small neighborhood for each pixel. Meanwhile, d_{ij} introduces noise into the similarity calculation of the pixel values (see also supplementary materials). In order to reduce both the computational workload and the noise introduced, we choose to calculate the affinity matrix $\boldsymbol{A}$ within a fixed $K \times K$ window centered at each pixel, where $K \geq 3$ is an odd number. Specifically, $\boldsymbol{A}$ is initialized as all zeros, and we calculate the similarity score between the central pixel and all pixels in the window with $f(i,j) = \|\boldsymbol{I}_i - \boldsymbol{I}_j\|_2$, and fill only top K scores into $\boldsymbol{A}$ according to the indices. An instance for similar pixel selection is exhibited in Fig. 3. On the basis of supervising the known regions with L1 loss, the affinity loss

$$\mathcal{L}_{\text{affinity}} = \frac{1}{N}\|\boldsymbol{A\alpha} - \boldsymbol{\alpha}\|_1 \tag{3}$$

is added onto the whole alpha matte. As shown in Fig. 2 (b), the added loss invites subtle details to the alpha matte.

Though affinity loss $\mathcal{L}_{\text{affinity}}$ helps predict subtle details, it encounters two problems: i) it cannot predict details with long-range dependency, *e.g.*, long hair, and ii) it does dot encourage smooth transition of alpha values at boundaries. The problems will be elaborated as follows.

Braking effect. We have a close look at the failure of predicting details with long-range dependency supervised by affinity loss. Fig. 4 provides a visualization and a simulated example on long hair prediction. Let the blue color denote the background, and the yellow be a tiger with a hair of one pixel wide. For convenience we assume that the hair color is homogeneous, so that given a pixel on the hair, the selected similar pixels evenly distribute in both sides and the normalized weights are all $\frac{1}{K}$ in $\boldsymbol{A}$. Since the model first generates coarse mask for the foreground object and then starts to delineate the fine details (cf. Fig. 2 (b), from left to right), here we suppose the tiger body has been correctly predicted, and use a sequence $\{\alpha_t\}$ to denote the alpha values corresponding to the hair, from left to right. Then the condition $\boldsymbol{A\alpha} = \boldsymbol{\alpha}$ is equivalent to

$$\frac{1}{K}(S_{t-1} - S_{t-K} + \alpha_t) = \alpha_{t-\frac{K-1}{2}}\,, \tag{4}$$

where $S_t = \sum_{i=1}^{t} \alpha_i$. Then $\{\alpha_t\}$ has a recursion

$$\alpha_t = \alpha_{t-K} + K(\alpha_{t-\frac{K-1}{2}} - \alpha_{t-\frac{K+1}{2}})\,. \tag{5}$$

Its characteristic equation $x^K - K(x^{\frac{K+1}{2}} - x^{\frac{K-1}{2}}) - 1 = (x-1)^3(\frac{K^2-1}{8}x^{\frac{K-3}{2}} + \sum_{i=1}^{\frac{K-3}{2}} \frac{i(i+1)}{2}(x^{i-1} + x^{K-2-i})) = 0$ has three multiple real roots 1 and $K-3$ complex roots. Given the monotonically non-increasing feature, $\{\alpha_t\}$ is a form of $\alpha_t = C_1 + C_2 t + C_3 t^2$, where C_1, C_2, C_3 are constants. Besides being always 1 as expected, $\{\alpha_t\}$ can also be approximate linear or quadratic variation over t to produce small loss. Meanwhile, under the power of the known loss, the starting and ending value are initialized as 1 and 0 respectively. Under their influence $\{\alpha_t\}$ tends to distribute as in the bottom right subplot of Fig. 4, implying that the hair growth gets stuck early. Enlarging the window size K helps a little when the hair is similar to the body in color, because more body values are taken into account in windows responsible for calculating the first few α_t's. However, it is unhelpful for isolated hair according to the reasoning of Fig. 4. Therefore the condition $\mathcal{L}_{\text{affinity}} = 0$ does not satisfy our optimization objective.

Hard segmentation effect. $\mathcal{L}_{\text{affinity}}$ does not ensure smooth transitions at boundaries. Consider a certain part on the edge where the color change obeys linear variation horizontally in the image. Then its cross section can be described as $\boldsymbol{I}_i = \boldsymbol{a}x_i + \boldsymbol{b}$, where x_i is the horizontal coordinate of pixel i, and $\boldsymbol{a}$, $\boldsymbol{b}$ are constant vectors. Given a pixel on the hair, the top similar pixels are evenly distributed on both sides, whose similarity scores can be calculated as

$$\boldsymbol{A}(i,j) = 1 - \frac{\|\boldsymbol{I}_i - \boldsymbol{I}_j\|_2}{C} = 1 - \frac{|x_i - x_j|\|\boldsymbol{a}\|_2}{C}\,. \tag{6}$$

Therefore symmetrical pixels correspond to equal similarity scores. Let $w_i, i = 1, 2, ..., K$ denote the row-normalized weights, where $w_i = w_{K-i}$. Now $\boldsymbol{A\alpha} = \boldsymbol{\alpha}$ becomes $\sum_{j=1}^{K} w_j \alpha_j = \alpha_i$. Then multiple situations satisfy $\boldsymbol{A\alpha} = \boldsymbol{\alpha}$, *e.g.*, $\alpha_i = m x_i + n$, where m is an arbitrary number and n can be chosen to meet other constraints. Because known loss encourages hard segmentation, the linear variation is squeezed to be more sharp similar to the bottom right subplot in Fig. 4. In other words, the combination of the two losses contributes to hard segmentation.

Distance Consistency Loss

We probe to add proper constraints to adapt nonlocal principle to the deep learning process. As discussed on Fig. 4,

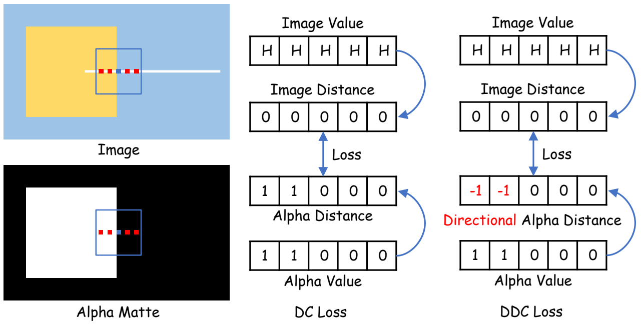

Figure 5: **A calculation instance of DC loss and DDC loss.** On a hair whose pixel value is H, K similar pixels are selected in the $K \times K$ window centered at a certain pixel. With the selected indices, the corresponding alpha values are gathered. DC loss first calculates the euclidean distances between the center pixel and the selected similar pixels both in the image and in the alpha matte, and then forces the two distance to be equal. Based on DC loss, DDC loss eliminates the interior noise by preserving the sign of alpha distance.

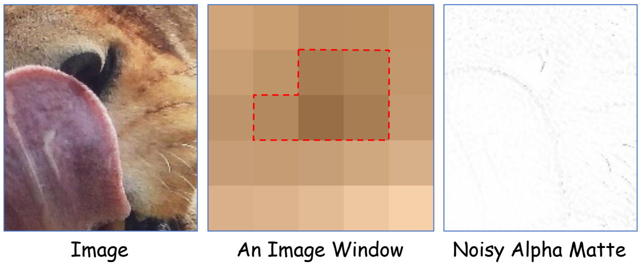

Figure 6: **Noise introduced by DC loss.** Owing to the body textures, the selected similar pixels (in the red dotted box) often have different values. DC loss forces the different values to exist also in the alpha matte. Such conflicts between DC loss and known loss introduces noise, appearing as undesired image textures in the alpha matte.

affinity loss only stipulates linear or quadratic variation as a whole, but does not control the local trend. Therefore the affinity loss fails to provide enough punishment when alpha values tend to distribute as in the bottom right subplot. We address this by presenting a new assumption: if some pixels in the image are similar, then their corresponding alpha values are *similar of the same degree*. Different from the affinity loss, we constrain pair-wise distance in the alpha matte and in the image to be equal. The new assumption brings about the Distance Consistency (DC) loss shown in Fig. 8:

$$\mathcal{L}_{\text{DC}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j}\left|\|\alpha_i - \alpha_j\|_2 - \|\boldsymbol{I}_i - \boldsymbol{I}_j\|_2\right| , \quad j \in \operatorname{argtopk}\{-\|\boldsymbol{I}_i - \boldsymbol{I}_j\|_2\} . \tag{7}$$

DC loss forces the variation of the alpha values to be the same as that in the image, so the smooth transition property at boundaries can be naturally satisfied. Moreover, DC loss governs the difference between local alpha pairs according to the pixel distance, which avoids the situation in the bottom right subplot in Fig. 4. Fig. 2 (c) suggests the power of DC loss to fit subtle details.

Directional Distance Consistency Loss

Though DC loss offers subtle details, it causes conflicts against known loss. An instance of noisy output is shown in Fig. 6. On the one hand, known loss forces '1' values on the foreground; on the other, DC loss expects certain variation because of the body textures. The loss confrontation introduces texture noise. To make the two losses compatible, we update Eq. (7) by preserving the sign of alpha distance:

$$\mathcal{L}_{\text{DDC}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j}\left|\alpha_i - \alpha_j - \|\boldsymbol{I}_i - \boldsymbol{I}_j\|_2\right| , \quad j \in \operatorname{argtopk}\{-\|\boldsymbol{I}_i - \boldsymbol{I}_j\|_2\} , \tag{8}$$

which is termed as the Directional Distance Consistency (DDC) loss. The calculation process of DDC loss is shown on the right of Fig. 5, and the code can be found in the supplementary. Consider a similar pair of pixel i and j. If j is selected as the similar pixel at the i-centered window, then i will likely be selected at the window centered at j. Then DDC loss has a minimum of $2\|\boldsymbol{I}_i - \boldsymbol{I}_j\|$ at each similar pair, as deduced in the following absolute value inequality:

$$\begin{aligned} &\left|\alpha_i - \alpha_j - \|\boldsymbol{I}_i - \boldsymbol{I}_j\|\right| + \left|\alpha_j - \alpha_i - \|\boldsymbol{I}_i - \boldsymbol{I}_j\|\right| \\ \geq &\left|\alpha_i - \alpha_j - \|\boldsymbol{I}_i - \boldsymbol{I}_j\| + \alpha_j - \alpha_i - \|\boldsymbol{I}_i - \boldsymbol{I}_j\|\right| \\ = &2\|\boldsymbol{I}_i - \boldsymbol{I}_j\| . \end{aligned} \tag{9}$$

Note the condition of equality meets when $0 \leq |\alpha_i - \alpha_j| \leq \|\boldsymbol{I}_i - \boldsymbol{I}_j\|$. When the pair are in the absolute foreground/background, known loss forces $\alpha_i = \alpha_j$, satisfying the condition of equality. At a cross section of the boundary, alpha values change monotonously. Let us just take $\alpha_i \geq \alpha_j$, then the condition of equality $0 \leq \alpha_i - \alpha_j \leq \|\boldsymbol{I}_i - \boldsymbol{I}_j\|$ illustrates that the change rate of α is bounded by $\|\boldsymbol{I}_i - \boldsymbol{I}_j\|$, so the smooth transition property still holds. In short, we set a positive minimum for DC loss, where the know loss also reaches its minimum 0.

Total loss. The total loss is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{known}} + \lambda \mathcal{L}_{\text{DDC}} , \tag{10}$$

where λ is a positive number to balance two losses.

Computational complexity. Since the window size K is small, DDC loss adds negligible additional computation.

Experiment

Here we verify the effectiveness of our training paradigm. More results can be found in the supplementary materials.

Implementation Details

Training setting. We choose the ViTMatte (Yao et al. 2023) as the deep matting model. In Eq. (8), the window size K is set as 11 by default. For the total loss Eq. (10), λ is set as 10. Other details can be found in the supplementary.

Dataset. Affected by domain gap, models trained on synthetic data (Composition-1K (Xu et al. 2017), Distinct 646 (Qiao et al. 2020), etc.) often work poorer in reality. Without requiring fine labels, the proposed method does not

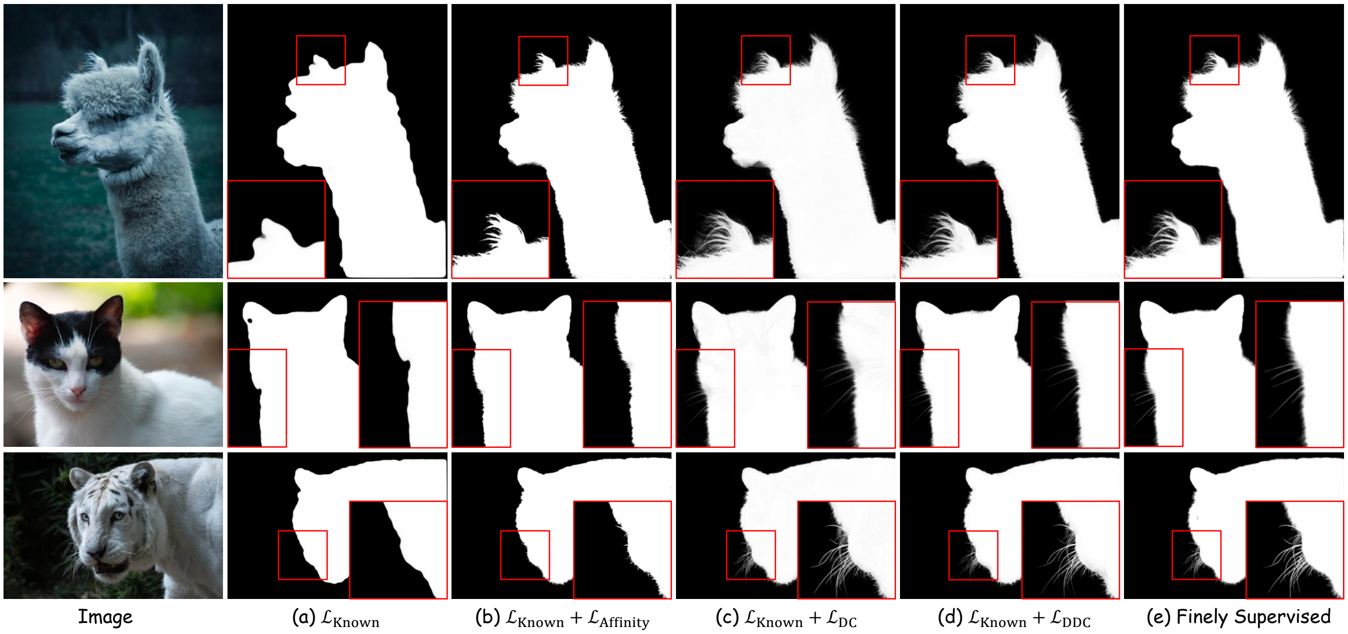
Image
(a) $\mathcal{L}_{\text{Known}}$
(b) $\mathcal{L}_{\text{Known}} + \mathcal{L}_{\text{Affinity}}$
(c) $\mathcal{L}_{\text{Known}} + \mathcal{L}_{\text{DC}}$
(d) $\mathcal{L}_{\text{Known}} + \mathcal{L}_{\text{DDC}}$
(e) Finely Supervised

Figure 7: **Visual results of trimap-supervised baselines and the finely supervised baseline.** The examples are chosen from the AM-2K (Li et al. 2022) test set and the P3M-NP-500 test set of P3M-10K (Li et al. 2021).

Label	SAD	MAD	MSE	Grad	Conn	SAD-T	MSE-T
Matte	26.00	0.0149	0.0101	16.04	10.05	10.82	0.0268
Trimap	30.26	0.0175	**0.0099**	**14.98**	14.33	16.07	0.0356

Table 1: Comparison on the training paradigm.

Method	Label	SAD	MAD	MSE	Grad	Conn	SAD-T
SHM (Chen et al. 2018)	Matte	17.81	0.0102	0.0068	12.54	17.02	10.26
LF (Zhang et al. 2019)	Matte	36.12	0.0210	0.0116	21.06	33.62	19.68
SSS (Aksoy et al. 2018)	Matte	552.88	0.3225	0.2742	60.81	555.97	88.23
HATT (Qiao et al. 2020)	Matte	28.01	0.0161	0.0055	18.29	17.76	13.36
GFM (Li et al. 2022)	Matte	10.26	0.0059	0.0029	8.82	9.57	8.24
Ours	Trimap	30.26	0.0175	0.0099	14.98	14.33	16.07

Table 2: Automatic animal matting results on AM-2K (Li et al. 2022) test set.

rely on data synthesis to produce labels. Hence, we verify the effectiveness of our method directly on natural datasets AM-2K (Li et al. 2022) and P3M-10K (Li et al. 2021).

Evaluation. We report the common used metrics of Sum of Absolute Differences (SAD), Mean Squared Error (MSE), Gradient (Grad) and Connectivity (Conn) proposed by (Rhemann et al. 2009) and Mean Absolute Difference (MAD) following (Li et al. 2022) for evaluation.

Main Results

Validation of the effectiveness. The effectiveness of our proposed paradigm should be assessed by comparison against the fine-label-supervised counterpart as illustrated in Fig. 1. The quantitative results of the two baselines correspond to Row 1 and Row 2 in Table 1 respectively. It demonstrates that our paradigm has comparable performance with the finely supervised baseline, and even better on the Grad and MSE metric. The visual comparison can be referred to Fig. 7 (d) and (e), where the model trained under our paradigm by only coarse trimap labels predicts similar alpha mattes with the counterpart trained by alpha matte labels.

Method	Label	SAD	MAD	MSE	Grad	Conn	SAD-T	MSE-T
LF (Zhang et al. 2019)	Matte	32.59	0.0188	0.0131	31.93	19.50	14.53	0.0420
HATT (Qiao et al. 2020)	Matte	30.53	0.0176	0.0072	19.88	27.42	13.48	0.0403
SHM (Chen et al. 2018)	Matte	20.77	0.0122	0.0093	20.30	17.09	9.14	0.0255
MODNet (Ke et al. 2022)	Matte	16.70	0.0097	0.0051	15.29	13.81	9.13	0.0237
GFM (Li et al. 2022)	Matte	15.50	0.0091	0.0056	14.82	18.03	10.16	0.0268
P3M (Li et al. 2021)	Matte	11.23	0.0065	0.0035	10.35	12.51	5.32	0.0094
Ours	Trimap	31.66	0.0182	0.0126	15.41	13.66	12.03	0.0371

Table 3: Automatic portrait matting results on P3M-NP-500 (Li et al. 2021).

Loss policy	SAD	MAD	MSE	Grad	Conn	SAD-T	MSE-T
$\mathcal{L}_{\text{known}}$	45.94	0.0266	0.0203	77.70	12.35	29.89	0.1096
$\mathcal{L}_{\text{known}} + \mathcal{L}_{\text{affinity}}$	32.54	0.0188	0.0132	23.41	**11.00**	17.25	0.0496
$\mathcal{L}_{\text{known}} + \mathcal{L}_{\text{DC}}$	42.24	0.0247	0.0117	16.07	21.59	18.82	0.0396
$\mathcal{L}_{\text{known}} + \mathcal{L}_{\text{DDC}}$	**30.26**	**0.0175**	**0.0099**	**14.98**	14.33	**16.07**	**0.0356**

Table 4: Comparison among four training policies.

Comparison among different loss policies. Row 1-4 of Table 4 provides the quantitative results of each loss policy corresponds to the visualizations in Fig. 2 (a)-(d). While all supervised by trimaps, Row 1 uses only the known loss,

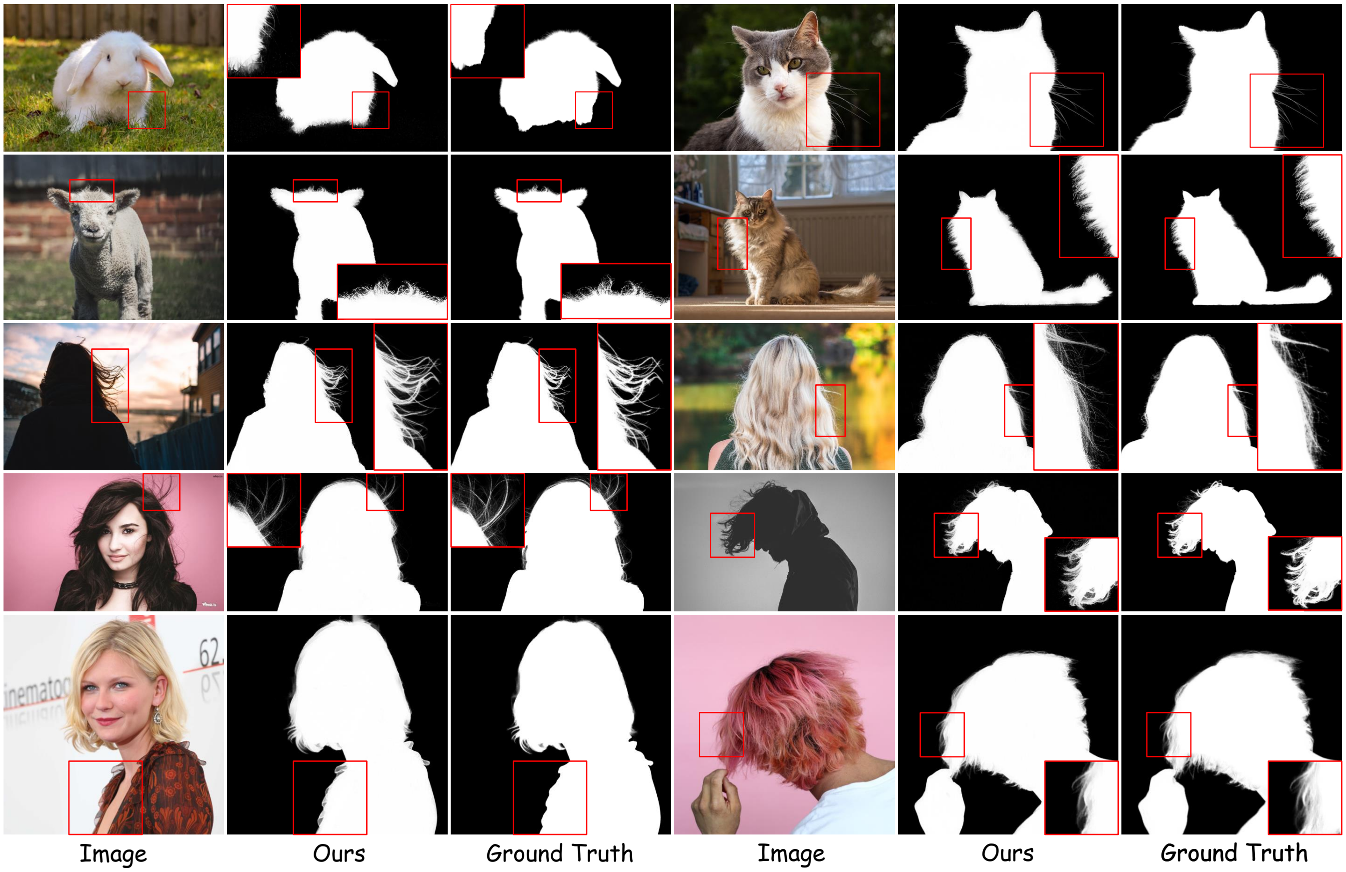

Figure 8: **Visualization on the AM-2K (Li et al. 2022) test set and the P3M-NP-500 test set of P3M-10K (Li et al. 2021).**

Row 2 adds affinity loss described by Eq. (3), Row 3 instead adopts known loss and DC loss in Eq. (7), and Row 4 replaces DC loss with DDC loss based on Row 3. Fig. 2 shows that known loss can only provide a coarse mask, in accord with the large SAD error in Table 4. Affinity loss can help fit a few details and reduce the errors, while DC loss invites more subtle details and largely reduces the Grad error by more than 7 points. The visual effect can be verified by the cat beard in the red box. However, due to the texture noise introduced (Fig. 2 (c) and Fig. 6), DC loss even increases the SAD error. The updated version DDC loss fits the details well and preserves the constant values in known areas, contributing high-quality alpha matte predictions. Quantitatively, DDC loss reduces the Grad error by 8.43 points while remains other metrics comparable with affinity loss.

Comparison with other methods. We compare our trained model with several recent automatic matting methods. As shown in Table 2 and Table 3, our model obtains comparable performance with recent automatic matting approaches. We provide several examples in Fig. 8. One can see that ours predicts subtle details for both animals and humans. Further more, because ours is directly trained under matting priors, it can sometimes produce results more correct than the human-labelled ground truth. For example, in the left subfigure of Row 1, ours produces more detailed alpha values on the conjunction of the rabbit and the grass; in the left subfigure of the last row, ours well delineate the chiffon on the clothes, while it is considered as opaque in the ground truth.

Conclusion

We present a new training paradigm to rid deep image matting of fine labels. In the proposed method, coarse trimap labels indicate the foreground/background, and a novel directional distance consistency loss (DDC loss) controls the alpha values at transition areas conditioned on the image. The design idea of DDC loss is to force the local distance between similar neighbors on the alpha matte and on the corresponding image to be equal. Experiments prove that the proposed paradigm yields high-quality matting predictions comparable to the finely supervised baseline. For future work, we will explore transparent objects matting and other image matting tasks such as interactive matting.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (No. 62306261), The Shun Hing Institute of Advanced Engineering (No. 8115074), CUHK Direct Grants (No. 4055190), and Hubei Provincial Natural Science Foundation of China under Grant No. 2024AFB566.

References

Aksoy, Y.; Oh, T.-H.; Paris, S.; Pollefeys, M.; and Matusik, W. 2018. Semantic soft segmentation. *ACM Transactions on Graphics (TOG)*, 37(4): 1–13.

Aksoy, Y.; Ozan Aydin, T.; and Pollefeys, M. 2017. Designing effective inter-pixel information flow for natural image matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 29–37.

Buades, A.; Coll, B.; and Morel, J.-M. 2005. A review of image denoising algorithms, with a new one. *Multiscale modeling & simulation*, 4(2): 490–530.

Buades, A.; Coll, B.; and Morel, J.-M. 2008. Nonlocal image and movie denoising. *International journal of computer vision*, 76: 123–139.

Cai, H.; Xue, F.; Xu, L.; and Guo, L. 2022. Transmatting: Enhancing transparent objects matting with transformers. In *European Conference on Computer Vision*, 253–269. Springer.

Chen, Q.; Ge, T.; Xu, Y.; Zhang, Z.; Yang, X.; and Gai, K. 2018. Semantic human matting. In *Proceedings of the 26th ACM international conference on Multimedia*, 618–626.

Chen, Q.; Li, D.; and Tang, C.-K. 2013. KNN matting. *IEEE transactions on pattern analysis and machine intelligence*, 35(9): 2175–2188.

Cho, D.; Tai, Y.-W.; and Kweon, I. 2016. Natural image matting using deep convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 626–643. Springer.

Chuang, Y.-Y.; Curless, B.; Salesin, D. H.; and Szeliski, R. 2001. A bayesian approach to digital matting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, II–II. IEEE.

He, K.; Sun, J.; and Tang, X. 2010. Fast matting using large kernel matting laplacian matrices. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2165–2172. IEEE.

Hou, Q.; and Liu, F. 2019. Context-aware image matting for simultaneous foreground and alpha estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4130–4139.

Ke, Z.; Sun, J.; Li, K.; Yan, Q.; and Lau, R. W. 2022. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1140–1147.

Kirillov, A.; Girshick, R.; He, K.; and Dollár, P. 2019. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6399–6408.

Lee, P.; and Wu, Y. 2011. Nonlocal matting. In *CVPR 2011*, 2193–2200. IEEE.

Levin, A.; Lischinski, D.; and Weiss, Y. 2007. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2): 228–242.

Li, J.; Ma, S.; Zhang, J.; and Tao, D. 2021. Privacy-preserving portrait matting. In *Proceedings of the 29th ACM international conference on multimedia*, 3501–3509.

Li, J.; Zhang, J.; Maybank, S. J.; and Tao, D. 2022. Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision*, 130(2): 246–266.

Li, J.; Zhang, J.; and Tao, D. 2021. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235*.

Li, Y.; and Lu, H. 2020. Natural image matting via guided contextual attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11450–11457.

Liu, J.; Yao, Y.; Hou, W.; Cui, M.; Xie, X.; Zhang, C.; and Hua, X.-s. 2020. Boosting semantic human matting with coarse annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8563–8572.

Liu, Q.; Xie, H.; Zhang, S.; Zhong, B.; and Ji, R. 2021a. Long-range feature propagating for natural image matting. In *Proceedings of the 29th ACM International Conference on Multimedia*, 526–534.

Liu, Y.; Xie, J.; Shi, X.; Qiao, Y.; Huang, Y.; Tang, Y.; and Yang, X. 2021b. Tripartite information mining and integration for image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7555–7564.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Lu, H.; Dai, Y.; Shen, C.; and Xu, S. 2019. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3266–3275.

Ma, S.; Li, J.; Zhang, J.; Zhang, H.; and Tao, D. 2023. Rethinking portrait matting with privacy preserving. *International journal of computer vision*, 1–26.

Niklaus, S.; and Liu, F. 2018. Context-aware synthesis for video frame interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1701–1710.

Park, G.; Son, S.; Yoo, J.; Kim, S.; and Kwak, N. 2022. Matteformer: Transformer-based image matting via prior-tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11696–11706.

Qiao, Y.; Liu, Y.; Yang, X.; Zhou, D.; Xu, M.; Zhang, Q.; and Wei, X. 2020. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13676–13685.

Rhemann, C.; Rother, C.; Wang, J.; Gelautz, M.; Kohli, P.; and Rott, P. 2009. A perceptually motivated online benchmark for image matting. In *2009 IEEE conference on computer vision and pattern recognition*, 1826–1833. IEEE.

Shen, X.; Tao, X.; Gao, H.; Zhou, C.; and Jia, J. 2016. Deep automatic portrait matting. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 92–107. Springer.

Sun, J.; Jia, J.; Tang, C.-K.; and Shum, H.-Y. 2004. Poisson matting. In *ACM SIGGRAPH 2004 Papers*, 315–321.

Tang, J.; Aksoy, Y.; Oztireli, C.; Gross, M.; and Aydin, T. O. 2019. Learning-based sampling for natural image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3055–3063.

Wei, T.; Chen, D.; Zhou, W.; Liao, J.; Zhao, H.; Zhang, W.; and Yu, N. 2021. Improved image matting via real-time user clicks and uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15374–15383.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, 418–434.

Xu, N.; Price, B.; Cohen, S.; and Huang, T. 2017. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2970–2979.

Yang, S. D.; Wang, B.; Li, W.; Lin, Y.; and He, C. 2022. Unified interactive image matting. *arXiv preprint arXiv:2205.08324*.

Yao, J.; Wang, X.; Yang, S.; and Wang, B. 2023. ViTMatte: Boosting image matting with pre-trained plain vision transformers. *Information Fusion*, 102091.

Yu, Q.; Zhang, J.; Zhang, H.; Wang, Y.; Lin, Z.; Xu, N.; Bai, Y.; and Yuille, A. 2021. Mask guided matting via progressive refinement network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1154–1163.

Zhang, Y.; Gong, L.; Fan, L.; Ren, P.; Huang, Q.; Bao, H.; and Xu, W. 2019. A late fusion cnn for digital matting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7469–7478.