Human Instance Matting via Mutual Guidance and Multi-Instance Refinement

Yanan Sun[1,2] Chi-Keung Tang[1] Yu-Wing Tai[2]
[1]HKUST [2]Kuaishou Technology
{now.syn, yuwing}@gmail.com cktang@cs.ust.hk

Abstract

This paper introduces a new matting task called human instance matting (HIM), which requires the pertinent model to automatically predict a precise alpha matte for each human instance. Straightforward combination of closely related techniques, namely, instance segmentation, soft segmentation and human/conventional matting, will easily fail in complex cases requiring disentangling mingled colors belonging to multiple instances along hairy and thin boundary structures. To tackle these technical challenges, we propose a human instance matting framework, called InstMatt, where a novel mutual guidance strategy working in tandem with a multi-instance refinement module is used, for delineating multi-instance relationship among humans with complex and overlapping boundaries if present. A new instance matting metric called instance matting quality (IMQ) is proposed, which addresses the absence of a unified and fair means of evaluation emphasizing both instance recognition and matting quality. Finally, we construct a HIM benchmark for evaluation, which comprises of both synthetic and natural benchmark images. In addition to thorough experimental results on complex cases with multiple and overlapping human instances each has intricate boundaries, preliminary results are presented on general instance matting. Code and benchmark are available in https://github.com/nowsyn/InstMatt.

1. Introduction

Fast development of mobile internet technology has triggered the rapid growth of multimedia industry especially we-media, where users are heavily engaged in editing tools to beautify or re-create their image and video contents. As one of the primary techniques for efficient image editing, image matting has achieved significant improvement with the wide adoption of deep neural networks in the task. However, existing matting methods still fail or else are not easy to use in many scenarios, such as extracting the foreground human while removing background humans, or instance-level editing as shown in Figure 1: what if we want to independently extract and edit each human instance?

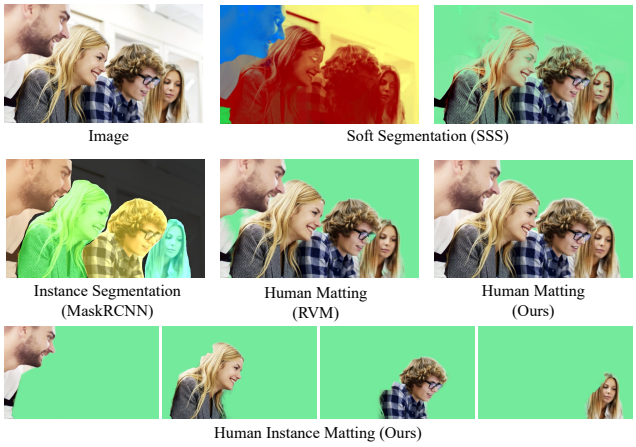

Figure 1. Comparisons with related tasks, including soft segmentation results from SSS [5], instance segmentation results from MaskRCNN [22], human matting results from RVM [36] as well as ours, and human instance matting results from ours.

Similar to semantic versus instance segmentation, existing matting methods, which focus on a region based on a given trimap or a known object class, are unable to differentiate instances. To address this issue, we propose a new task called human instance matting (HIM), which aims to automatically extract precise alpha matte for each human instance in a given image. HIM shares similarities to the following conventional tasks while embodying fundamental differences making it a problem on its own: 1) instance segmentation aims at distinguishing instances, but it can only produce sharp object boundary without semi-transparency consideration; 2) recent soft segmentation [5] is capable of generating soft segments for multiple instances of different classes with instance-aware features, but cannot deal with instances of the same class; 3) conventional matting aims at extracting precise alpha matte, but it lacks instance awareness. Overall, human instance matting is a unified task encompassing the characteristics of the aforementioned related tasks while introducing new technical challenges.

Conventional matting is based on the image compositing

This work was done when Yanan Sun was a student intern at Kuaishou Technology, which was supported by Kuaishou Technology and the Research Grant Council of the Hong Kong SAR under grant no. 16201420.

equation where an image I is the combination of foreground F layer, background B layer modulated by alpha α:

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

To adapt to multiple instance matting, we modify the 2-layer Equation 1 to one of multi-instance layered composition, where each instance layer is attenuated by its corresponding α:

$$I = \sum_{i=0}^{n} \alpha_i L_i, \qquad \text{s.t.} \sum_{i=0}^{n} \alpha_i = 1 \tag{2}$$

where L_i and α_i respectively denote the foreground and alpha matte for instance $i > 0$; L_0 and α_0 respectively represent the background and its corresponding alpha matte; n is the number of instances. This equation had also appeared in [5, 31], but all such relevant matting and segmentation tasks were not instance aware. The goal of instance matting is to solve for target mattes α_i for all $i > 0$.

By exploring the complex relation among multiple instances, we propose a new instance matting framework, called **InstMatt**, where a novel **mutual guidance strategy** enables a deep model to decompose mingled compositing colors into their respective instances. Our mutual guidance strategy takes both the relation between instances and the background, and the relation among instances into consideration. Besides, a **multi-instance refinement** module is carefully designed and engineered for interchanging information among instances to synchronize predictions for further refinement. Equipped with the novel mutual guidance and multi-instance refinement, our InstMatt is able to not only produce high-quality human alpha matte but also distinguish multiple human instances shown in Figure 1.

With this new HIM task, existing evaluation metrics for instance segmentation or matting are insufficient, which were designed for either one of the tasks. We propose a new metric, called instance matting quality (**IMQ**), that simultaneously measures instance recognition quality and alpha matte quality. To provide a general and comprehensive validation on instance matting techniques, we construct an instance matting benchmark, **HIM2K**, which consists of a synthetic image benchmark and a natural image benchmark totaling 2,000 images with high-quality matte ground truths.

To demonstrate the promise of our technical contributions beyond human instance matting, we present preliminary results on matting multi-object instances not limited to humans, a fruitful future direction to explore.

2. Related Work

2.1. Matting

Natural Image Matting. Image matting is a pixel-level task, aiming to extract alpha matte for a foreground object. Traditional matting methods can be summarized into two approaches. Sampling-based methods [16, 20, 23, 23] collect a set of known foreground and background samples to estimate unknown alpha values. Propagation-based methods [6,7,12,21,30,31] assume neighboring pixels are correlated, and use their affinities to propagate alpha from known regions to unknown regions. Traditional methods rely on low-level or statistical features, which can easily fail on complex cases due to their limited feature representation.

The wide application of deep convolutional neural network (CNN) addresses this feature representation issue to a great extent. DCNN [15] and DIM [55] are the first representative methods to apply CNN in matting, which are followed by a series of valuable works advancing the state-of-the-art matting performance. Deep learning-based methods can be further grouped into three approaches. Trimap (or mask) based methods [9, 17, 18, 24, 25, 33, 41–43, 48, 49, 52, 56] take an additional trimap to focus the model on the target foreground object. With careful network design, these methods have achieved excellent performance. User-supplied constraints are relaxed in [35, 45] by using an extra photo taken without the relevant foreground object for providing useful prior information. Trimap-free methods [44,58] erase the dependence on additional input. These methods resort attention or salience to localize foreground object and extract the corresponding alpha matte.

Human Matting. Human matting is a class-specific image matting task, where the semantic information of the foreground object, namely, human is known. Known human semantics effectively guides relevant human matting methods and thus they usually do not require additional input. Deep learning-based human matting was first proposed in [46] and then improved in SHM [11]. A method was proposed in BSHM [38] which makes use of coarse annotated data for boosting performance. MODNet [28] addresses automatic and fast human matting using a light-weight network considering both low-resolution semantics and high-resolution details. In RVM [36] a video human matting framework was proposed using a recurrent decoder to improve robustness. Further, a cascade framework is proposed in [57] to extract alpha matte from low-to-high resolution.

2.2. Segmentation

Instance Segmentation. Instance segmentation simultaneously requires instance-level and pixel-level predictions. The existing methods can be classified into three categories. Top-down methods [8,10,13,22,26,27,34,40] first detect instances and then segment the object within detected bounding boxes. On the contrary, bottom-up methods [19,39] first learns the embeddings for each pixel and then group them into instances. Direct methods [53, 54] are box-free and grouping-free. They predict instance masks with classification in one shot without a detection or clustering step.

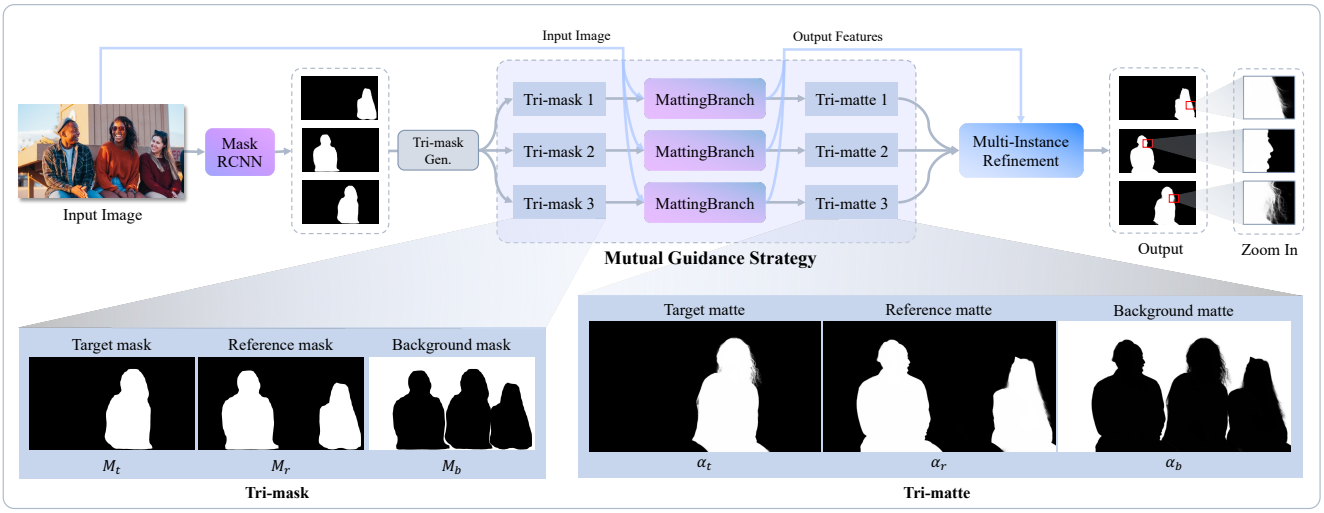

Figure 2. Overall **InstMatt** framework consisting of mutual guidance and multi-instance refinement. We first apply MaskRCNN to obtain instance masks, and then generate **tri-mask** for each instance to provide **mutual guidance** for the matting branch. Through mutual guidance strategy, we upgrade coarse tri-masks into fine **tri-mattes** for all instances. Finally, a **multi-instance refinement** module (illustrated in Figure 3) is designed to make use of the information difference of underlining tri-mattes to further promote the instance matte quality.

Soft Segmentation. Soft segmentation is a pixel-level task, decomposing an image into several segments where each pixel may belong partially to multiple segments. Different decomposition methods lead to different segments. For instance, soft color segmentation methods [3, 4, 47, 50, 51] decompose an image into soft layers of homogeneous colors; spectral matting [31] clusters an image into a set of spectral segments; SSS [5] decomposes an image into soft semantic segments via aggregating high-level embeddings with local-level textures.

2.3. Instance Matting

Instance matting maps each pixel into a set of soft or fractional alphas each tagged with an unique instance ID. Besides inheriting the difficulties from instance segmentation and soft segmentation, instance matting introduces new algorithmic challenges. Specifically, compared to instance segmentation, each pixel in instance matting can partially belong to more than one instance; compared to soft segmentation, each pixel can belong to multiple instances of the same class. To the best of our knowledge, there is no unified framework that can simultaneously address these technical challenges brought by the new instance matting problem. In this paper, we take human instance matting as an example, and propose a framework to address the aforementioned issues via our novel mutual guidance and multi-instance refinement.

3. Method

Our HIM framework, called **InstMatt**, consists of two steps, first recognizing instances and then extracting their respective alpha mattes. This allows the model to globally discover instances and then refine them according to local context. Figure 2 illustrates the whole framework.

3.1. Observations

Sparsity. Equation 2 indicates that a given pixel can belong to multiple instances and hence $\alpha_i, i = 1 \cdots n$. However, in real-life images even containing many instances, each pixel usually consists of no more than two non-zero αs, belonging to an instance and the background, or two overlapping instances, thus satisfying the sparsity observation of multi-instance matting.

Mutual Information and Tri-mattes. To estimate target instance alpha matte α_i, the other instances $j \neq i$ can be regarded as reference information. Note that we do not regard the other instances as part of background since they have different semantic representation. Therefore, we can re-formulate Equation 2 into the following equation using *three* components, i.e., target instance $\mathcal{T}$, the other instances if any $\mathcal{R}$ (also named reference instances), and the background $\mathcal{B}$:

$$I = \underbrace{\alpha_i L_i}_{\text{target } (\mathcal{T})} + \underbrace{\alpha_0 L_0}_{\text{background } (\mathcal{B})} + \underbrace{\sum_{j=1 \text{ and } j\neq i}^{n} \alpha_j L_j}_{\text{reference instances } (\mathcal{R})} \quad (3)$$

If we treat the component $\mathcal{R}$ as a new combined layer, Equation 3 is then simplified into a sparse representation as Equation 4, which considers the sparsity constraint:

$$I = \alpha_t L_t + \alpha_b L_b + \alpha_r L_r, \quad \text{s.t. } \alpha_t + \alpha_b + \alpha_r = 1 \quad (4)$$

where subscripts t,r,b represent the three components $\mathcal{T}$, $\mathcal{R}$,$\mathcal{B}$ respectively. For a target instance, Equation 4 implies

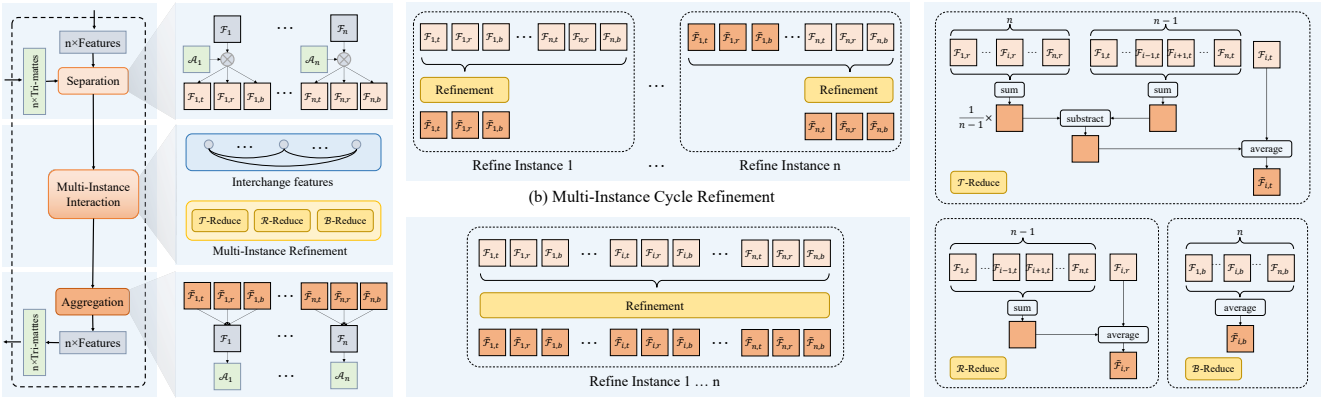

Figure 3. (a) Structure of our multi-instance refinement module, where instances exchange information among each other to refine their features through a multi-instance interaction layer. Two representative multi-instance refinement strategies, i.e. (b) cycle refinement and (c) parallel refinement are proposed and discussed. Figure (d) illustrates three feature reduction operations used in the two refinement ways.

that the alpha matte of each pixel can be correspondingly decomposed into three components, α_t, α_r and α_b (where one or two of them can be zero). These three components provide mutual information for one another, and they are collectively termed **tri-mattes**.

3.2. Mutual Guidance Strategy

Given an image, we first apply MaskRCNN [22] to extract coarse masks M for human instances. The challenge lies in turning the coarse mask into precise alpha matte for each instance. When only one instance exists, the task reduces into conventional human matting, which can be addressed by [56] or other matting techniques. To handle multiple instances, according to the above observations, we propose a novel mutual guidance strategy implemented using a *tri-mask*. Tri-mask $\mathcal{M}$ is defined as the concatenation of M_t, M_r and M_b, which respectively mask the region of $\mathcal{T}$, $\mathcal{R}$ and $\mathcal{B}$. For instance i, $M_{i,t}$, $M_{i,r}$ and $M_{i,b}$ are computed using the following tri-mask generation formulas,

$$M_{i,t} = M_i, \qquad M_{i,r} = \bigcup_{j=1 \text{ and } j \neq i}^{n} M_j \tag{5}$$

$$M_{i,b} = 1 - M_{i,t} \cup M_{i,r} \tag{6}$$

Afterward, for each instance, we feed as input the concatenation of the image and its tri-mask into a matting branch for extracting its *tri-matte* $\mathcal{A}$, which is the concatenation of the alpha matte α_t, α_r and α_b. The matting branch is an encoder-decoder matting network adopting the same structure with the network used in [56]. After the matting branch, we extract the tri-mattes for all instances. To supervise $\mathcal{A}$, multi-instance constraints are employed which will be introduced in Section 3.4.

Prior information in tri-mask provides comprehensive guidance for the model in pixel decomposition. On the one hand, the mutual exclusion among M_t, M_r and M_b guides the model to distinguish human instances from the background. On the other hand, the separation between M_t and M_r guides the model to differentiate instances. Subject to the constraint $\alpha_t + \alpha_r + \alpha_b = 1$, we force the model to learn a mutual exclusive decomposition in a contrastive manner.

3.3. Multi-Instance Refinement

Given n instances, n tri-mattes, i.e., n triplets of $(\alpha_t, \alpha_r, \alpha_b)$ are derived via the aforementioned mutual guidance, which encourages intra-instance but not inter-intance consistencies, which may lead to misalignment among overlapping tri-mattes from different instances. We utilize such inter-instance inconsistencies to correct potential error of the estimated alpha mattes. Based on tri-mattes, we design a multi-instance refinement module (MIR), illustrated in Figure 3 to further promote the quality of alpha mattes for all target instances.

Overall Structure. Our multi-instance refinement module comprises of three steps: separation, interaction and aggregation as shown in Figure 3-(a). For each instance, we use $\mathcal{F}_i$ to represent the feature from the final layer before the prediction head in the matting branch. Though $\mathcal{F}_i$ embodies the information for $\mathcal{T}$, $\mathcal{R}$ and $\mathcal{B}$, it is infeasible to perform individual operation on these three components. Thus, we use tri-matte to provide spatial attention so as to obtain the separate features for $\mathcal{T}$, $\mathcal{R}$ and $\mathcal{B}$. Specifically, multiplied by $\alpha_{i,t}$, $\alpha_{i,r}$ and $\alpha_{i,b}$, we obtain three features $\mathcal{F}_{i,t}$, $\mathcal{F}_{i,r}$ and $\mathcal{F}_{i,t}$, $i \in \{1, 2, ..., n\}$.

Separate representations for $\mathcal{T}$, $\mathcal{R}$ and $\mathcal{B}$ enable free communications and interactions to a large extent among instances. In the second step, a novel **multi-instance interaction** layer is proposed, in which each instance sends its features to other instances and receives the features from other instances. As the number of features varies with the number of instances, feature reduction operation is required to integrate these received features for refinement. Specifi-

cally, the refinement consists of three reduction operations, i.e., $\mathcal{T}$-reduce, $\mathcal{R}$-reduce, and $\mathcal{B}$-reduce, which are defined in Equation 7–9 (Figure 3-(d)).

$$\tilde{\mathcal{F}}_{i,t} = \frac{1}{2}(\mathcal{F}_{i,t} + \frac{1}{n-1}\sum_{j=1}^{n}\mathcal{F}_{j,r} - \sum_{j=1 \text{ and } j\neq i}^{n}\mathcal{F}_{j,t}) \quad (7)$$

$$\tilde{\mathcal{F}}_{i,r} = \frac{1}{2}(\mathcal{F}_{i,r} + \sum_{j=1 \text{ and } j\neq i}^{n}\mathcal{F}_{j,t}) \quad (8)$$

$$\tilde{\mathcal{F}}_{i,b} = \frac{1}{n}\sum_{j=1}^{n}\mathcal{F}_{j,b} \quad (9)$$

Equation (7)–(9) can be regarded as an averaging process. Such 'averaging' can provide communication among instances obtained from each individual branch to alleviate uncertainty and stabilize the convergence. After the multi-instance interaction layer, we reunify $\tilde{\mathcal{F}}_{i,t}$, $\tilde{\mathcal{F}}_{i,r}$ and $\tilde{\mathcal{F}}_{i,t}$ to produce an enhanced feature for tri-matte estimation.

Cycle versus Parallel Refinement. In the multi-instance interaction layer, after instances interchanging features information, there are numerous refinement possibilities since instances can refine their features concurrently or successively. Here, we discuss two representative refinement strategies in the multi-instance interaction layer, i.e. cycle refinement and parallel refinement shown in Figure 3-(b) and (c) respectively:

- *Cycle refinement*. Instances refine their features with the help of other features sequentially. For example, instance 1 first refines its feature and then sends its refined feature to all other instances. Next, instance 2 refines its features with the refined features from instance 1 and the unrefined features from the rest instances, and so on. Finally, instance n refines its features based on the refined features from all the other instances.
- *Parallel refinement*. Instances refine their features with the help of other features simultaneously. All instances refine their features based on the unrefined features from the other instances.

Both refinement strategies are effective in utilizing multi-instance mutual information to alleviate the effect of outliers. Since cycle refinement is order-sensitive, parallel refinement is preferable in non-interactive applications. We adopt parallel refinement in this paper. More comparisons and implementation details can be found in the supplementary materials.

3.4. Multi-Instance Constraint

Conventional matting losses, i.e., alpha loss and pyramid Laplacian loss, are still applicable in instance matting. Specially, we apply alpha loss and pyramid Laplacian loss for α_t, α_r and α_b separately. Their summations are denoted by L_α and L_{lap}.

Alpha loss and pyramid Laplacian loss directly regularize the distance between the estimated alpha matte and the ground truth, not considering composition constraint and alpha constraint among multiple instances as well as the background. We adapt the composition loss to accommodate multi-instance composition constraint as Equation 10,

$$L_{mc} = ||\alpha_t F_t + \alpha_r F_r + \alpha_b F_b - I||_1 \quad (10)$$

In addition, we employ multi-instance alpha constraint on tri-matte as Equation 11 to reduce the solution space:

$$L_{m\alpha} = ||\alpha_t + \alpha_r + \alpha_b - 1||_1 \quad (11)$$

Finally, the total loss is the summation of the aforementioned losses as Equation 12,

$$L = L_\alpha + L_{lap} + L_{mc} + L_{m\alpha} \quad (12)$$

We apply the loss defined in Equation 12 for the tri-mattes from both the matting branch and the multi-instance refinement module.

4. Benchmark

Existing benchmarks are designed for instance segmentation such as COCO dataset [37], or matting such as Composition-1K [55], but not for instance matting. They cannot provide a comprehensive evaluation for instance matting. In this paper, we propose a human instance matting benchmark called HIM2K, which is composed of two subsets, synthetic image subset and natural image subset respectively containing 1,680 and 320 images.

Synthetic Subset. We collect a variety of human images and carefully extract the human foregrounds. Then, we randomly select 2–5 such foregrounds F_i, and iteratively composite them onto a non-human background image sampled from BG20K [32] following Equation 13 below, where I_0 is the background image:

$$I_i = \alpha_i F_i + (1-\alpha_i) I_{i-1}, i \in \{1, ..., n\} \quad (13)$$

Expanding Equation 13 for each foreground object layer, a uniform formula can be derived as Equation 14:

$$I_i = I_0 \prod_{j=1}^{i}(1-\alpha_j) + \sum_{j=1}^{i}\alpha_j F_j \prod_{k=j}^{i}(1-\alpha_k) \quad (14)$$

If we use layer L to represent a foreground image F or background image I_0, Equation 14 for the last iteration can be simplified as Equation 15 which is the same as Equation 2:

$$I = \sum_{i=0}^{n}\alpha_i' L_i \quad (15)$$

Figure 4. HIM2K examples: top is synthetic and bottom is natural.

where $\alpha_i^{'}$ denotes the alpha matte of i-th layer L_i, the target to be estimated for instance i when $i > 0$.

Natural Subset. In light of the domain gap between synthetic and real images, we construct a natural subset for fair evaluation. The natural subset consists of 320 images containing multiple human instances of a variety of poses and scenarios, with ground truth alpha matte obtained by manual labeling using Photoshop. Despite the possibly imperfect (still reasonably accurate) annotation, we found that more than 98% of regions contain no more than 3 overlapping areas, which makes annotated ground truth trustworthy. Evaluation on the natural subset can validate the effectiveness and stability of different methods on real-world photos. Figure 4 shows examples from the two subsets.

5. IMQ Metric

In this section, we introduce a new metric for instance matting. Existing metrics are designed for either matting or instance segmentation including semantic segmentation. Instance segmentation metrics, such as mask average precision (mask AP), are used for measuring the binary instance mask quality, and thus unsuitable for evaluating alpha matte with fractional values in transitional region. On the other hand, the most widely used matting metrics, namely, the four errors MAD (or SAD), MSE, Gradient and Connectivity, measure alpha matte quality without instance awareness. The above limitations of existing metrics necessitate a new metric, which we call instance matting quality (IMQ).

Instance Matting Quality. IMQ measures instance matte quality giving attention to both instance recognition quality and matting quality. Inspired by the panoptic quality [29], IMQ is defined by Equation 16:

$$\text{IMQ} = \frac{\sum_{\alpha,\hat{\alpha}\in TP} S(\alpha,\hat{\alpha})}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{16}$$

where S is the similarity measurement function; TP, FP, and FN are respectively the true positive, false positive and false negative sets; α and $\hat{\alpha}$ are the predicted and ground truth instance alpha matte. The computation of IMQ has two steps: *instance matching* and *similarity measurement* as revealed in Equation 16.

Instance Matching. To match the predicted instance mattes with ground-truth instance mattes, the matching criterion is intersection-over-union (IoU) between α and $\hat{\alpha}$. We first quantify each instance matte into a binary mask by applying $\alpha > 0$ before computing IoU matrix. Based on the IoU matrix, we apply Hungarian matching [1], a greedy assignment strategy to achieve one-to-one assignment. All assigned predicted instance mattes are treated as TP candidates, where a candidate is assigned to TP if its IoU is above a threshold (0.5 is adopted in this paper). After settling the TP set, the FP set and FN set can be derived easily.

Similarity Measurement. The similarity measurement criterion is defined as Equation 17 below, where w is a balance factor, and $\mathcal{E}$ is an error function, e.g., MSE,

$$S(\alpha,\hat{\alpha}) = 1 - \min(w\mathcal{E}(\alpha,\hat{\alpha}), 1) \tag{17}$$

We denote IMQ applying MSE error function to measure similarity as IMQ_{mse}. If we replace the error function $\mathcal{E}$ by MAD, Gradient and Connectivity, we respectively obtain IMQ_{mad}, IMQ_{grad} and IMQ_{conn}.

Analysis. Similar to panoptic quality, IMQ can be decomposed into two components as Equation 18,

$$\text{IMQ} = \underbrace{\frac{\sum_{\alpha,\hat{\alpha}\in TP} \mathcal{S}(\alpha,\hat{\alpha})}{|TP|}}_{\text{Matting Quality (MQ)}} \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{Recognition Quality (RQ)}} \tag{18}$$

RQ has a similar expression to F_1-score, a metric widely used in recognition tasks, while MQ measures the matting quality for TP set. Different from existing instance segmentation and matting metrics, collaboration of RQ and MQ provides a fair and comprehensive evaluation for instance matte quality.

6. Experiments

In this section, we introduce our synthetic training dataset, evaluation and ablation studies. More details about the implementation including the network structure, data augmentations and training schedule can be found in the supplementary materials.

6.1. Synthetic Training Dataset

Since there is no off-the-shelf human instance matting training dataset, we construct our synthetic training dataset following [55], by compositing human instances onto background images. Specifically, for the foreground, we collect 38,618 human instances with matting annotations from Adobe Image Matting dataset [55], Distinctions-646 [44] and self-collected dataset. For the background, we use non-human high-resolution images from [32, 49].

To produce a synthetic image, we randomly pick 2 to 5 instances from the foreground set, and composite them

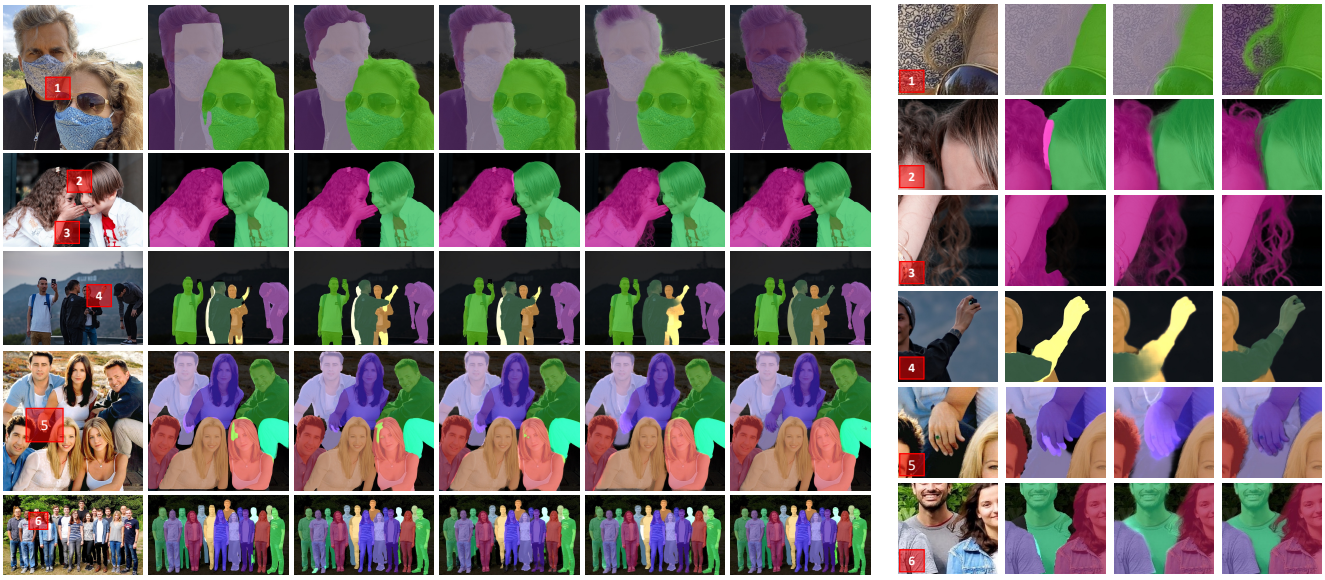

Figure 5. Qualitative comparisons on real-world images. Right shows the of zoom-ins of patch 1–6.

Figure 6. Qualitative results on COCO dataset.

Figure 7. Adaptation to other classes, i.e., cat and dog.

onto a background image. Random crop and zoom are applied on each foreground image. To avoid degenerate cases, such as a totally occluded instance, we composite instances with a random gap or overlap within some reasonable range. The composition is an iterative procedure following Equation 13. Finally, a total of 35,000 synthetic images with multiple instances are included in our training dataset.

6.2. Evaluation

Human Instance Matting. We perform joint qualitative and quantitative evaluations on multiple datasets, including HIM2K, RWP636 [56], SPD [2], COCO [37] dataset as well as more complex real-world images.

HIM2K is the proposed benchmark for human instance matting. Since our method is the first work to address instance matting, we compare our method with instance segmentation methods [14, 22] and a straightforward extension on existing state-of-the-art matting methods [18, 33, 48, 56] based on the masks from MaskRCNN [22]. To validate the effectiveness of our method, we also conduct comparisons on a human matting benchmark, Real World Portrait 636 (RWP636), and a human segmentation dataset, Supervisely Person dataset (SPD). SPD consists of 5418 images with fine mask annotations. We split a subset comprising of 500 images from SPD as the testing dataset. Table 1 and 2 tabulate the quantitative results on the three testing sets, showing our method achieves the state-of-the art performance.

Figure 5 shows qualitative comparisons on complex images, demonstrating that instance matting is capable of solving challenging cases with multiple and overlapping instances, which cannot be addressed by other existing instance segmentation or matting techniques. Note on the other hand while COCO is a widely used testing dataset in detection and segmentation tasks, the mask annotations are labeled by rough polygons thus making COCO inappropriate in quantitative results comparison for the instance matting task. Thus, we instead conduct qualitative comparisons on COCO dataset in Figure 6. Compared with instance segmentation algorithms, our InstMatt framework is significantly better in handling complex matting scenarios

Method	HIM2K (Synthetic Subset)				HIM2K (Natural Subset)				RWP636	
	IMQ_{mad}	IMQ_{mse}	IMQ_{grad}	IMQ_{conn}	IMQ_{mad}	IMQ_{mse}	IMQ_{grad}	IMQ_{conn}	IMQ_{mad}	IMQ_{mse}
MaskRCNN [22]	18.37	25.65	0.45	19.07	24.22	33.74	2.27	26.65	20.26	25.36
MaskRCNN + CascadePSP [14]	40.85	51.64	29.59	43.37	64.58	74.66	60.02	67.20	42.20	52.91
MaskRCNN + GCA [33]	37.76	51.56	38.33	39.90	45.72	61.40	44.77	48.81	33.87	46.47
MaskRCNN + SIM [48]	43.02	52.90	40.63	44.29	54.43	66.67	49.56	58.12	34.66	46.60
MaskRCNN + FBA [18]	36.01	51.44	37.86	38.81	34.81	48.32	36.29	37.23	35.00	47.54
MaskRCNN + MaskGuided [56]	51.67	67.08	53.03	55.38	57.98	71.12	66.53	60.86	30.64	53.16
InstMatt (Ours)	**63.59**	**78.14**	**64.50**	**67.71**	**70.26**	**81.34**	**74.90**	**72.60**	**51.10**	**73.09**

Table 1. Quantitative comparisons on HIM2K and RWP636 [56]. The balance factor w in Equation 17 is set to 10. For IMQ_{mad}, IMQ_{mse}, IMQ_{grad} and IMQ_{conn}, the higher, the better. Bold numbers indicate the best performance.

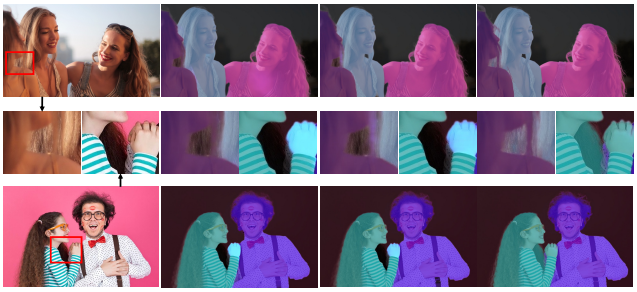

Figure 8. Comparisons among different mask guidance settings. See in particular the zoom-ins showing the best results when all three components are enabled, where the blonde's hairs and the man's shoulder are clearly delineated.

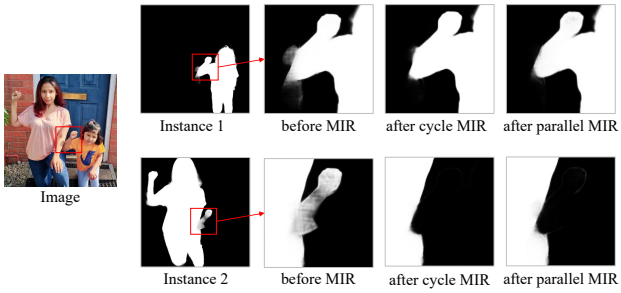

Figure 9. Alpha matte before and after multi-instance refinement.

at the instance level, such as defocus, motion, blurry or thin hairy structures.

Instance Matting Beyond Humans. This paper takes human instance matting as our focused contribution in instance matting. Notably, our method, including mutual guidance, multi-instance refinement, multi-instance constraints, and the proposed instance matting metric IMQ as well can be also applied to instance matting on other semantic classes. We adapt our method on another two popular classes, i.e., cat and dog. Preliminary results shown in Figure 7 indicate that our method may generalize well to other semantic classes in instance matting.

6.3. Ablation Study

Tri-mask. The tri-mask provides mutual guidance for both instances versus background as well as instances versus instances. Table 3 tabulates the results on models with different mask guidance settings. The tri-mask guides the model to assign each pixel partially to the target instance, other instances or the background. Notably, with tri-mask, some missing part due to occlusion are recovered as shown in the examples in Figure 8. The representation of the missing part is similar to that of the target instance, which cannot be ascribed to background or other instances due to the mutual exclusive supervision.

Method	IMQ_{mad}	IMQ_{mse}
MaskRCNN [22]	18.44	18.48
MaskRCNN + CascadePSP [14]	30.54	33.37
InstMatt (Ours)	**30.67**	**39.56**

Table 2. Quantitative results on SPD [2].

M_t	M_r	M_b	MIR	IMQ_{mad}	IMQ_{mse}
✓	✗	✗	✗	57.98	71.12
✓	✓	✗	✗	62.25	74.35
✓	✓	✓	✗	69.40	79.74
✓	✓	✓	✓	**70.26**	**81.34**

Table 3. Results on tri-mask and multi-instance refinement.

Multi-Instance Refinement. Multi-instance refinement aligns alpha matte predictions among multiple tri-mattes. Table 3 shows that the IMQ_{mse} of our model with and without multi-instance refinement module is 81.34 and 79.74, indicating an improvement from our multi-instance refinement. Figure 9 further shows that multi-instance refinement is helpful in erasing outliers due to the information synchronization among different instances.

7. Conclusion

In this paper, we propose a new task, instance matting with human instance matting as the first significant example by proposing a novel instance matting framework. Our InstMatt utilizes mutual exclusive guidance to guide the matting branch to extract alpha matte for each instance, which is followed by a multi-instance refinement module to synchronize information among co-occurring instances. InstMatt is capable of handling challenging cases with multiple and overlapping instances, which can be adapted to other semantic class instance matting beyond human instances. We hope the proposed method, alongside with the new instance matting metric and the human instance matting benchmark, will encourage more future works.

References

[1] Hungarian algorithm. https://en.wikipedia.org/wiki/Hungarian_algorithm. 6

[2] supervisely. https://supervise.ly. 7, 8

[3] Yagiz Aksoy, Tunç Ozan Aydin, Marc Pollefeys, and Aljosa Smolic. Interactive high-quality green-screen keying via color unmixing. *ACM Trans. Graph.*, 35(5):152:1–152:12, 2016. 3

[4] Yagiz Aksoy, Tunç Ozan Aydin, Aljosa Smolic, and Marc Pollefeys. Unmixing-based soft color segmentation for image manipulation. *ACM Trans. Graph.*, 36(2):19:1–19:19, 2017. 3

[5] Yağiz Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. Semantic soft segmentation. *ACM Transactions on Graphics*, 37(4):1–13, 2018. 1, 2, 3

[6] Yagiz Aksoy, Tunc Ozan Aydin, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 2

[7] Xue Bai and Guillermo Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *International Conference on Computer Vision*, 2007. 2

[8] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT: real-time instance segmentation. In *IEEE/CVF International Conference on Computer Vision*, 2019. 2

[9] Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. Disentangled image matting. In *International Conference on Computer Vision*, 2019. 2

[10] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[11] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *ACM Multimedia Conference*, 2018. 2

[12] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2012. 2

[13] Xinlei Chen, Ross B. Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *IEEE/CVF International Conference on Computer Vision*, 2019. 2

[14] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 7, 8

[15] Donghyeon Cho, Yu-Wing Tai, and In-So Kweon. Natural image matting using deep convolutional neural networks. In *European Conference on Computer Vision*, 2016. 2

[16] Yung-Yu Chuang, Brian Curless, David Salesin, and Richard Szeliski. A bayesian approach to digital matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2001. 2

[17] Yutong Dai, Hao Lu, and Chunhua Shen. Learning affinity-aware upsampling for deep image matting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[18] Marco Forte and François Pitié. F, b, alpha matting. *CoRR*, abs/2003.07711, 2020. 2, 7, 8

[19] Naiyu Gao, Yanhu Shan, Yupei Wang, Xin Zhao, Yinan Yu, Ming Yang, and Kaiqi Huang. SSAP: single-shot instance segmentation with affinity pyramid. In *IEEE/CVF International Conference on Computer Vision*, 2019. 2

[20] Eduardo SL Gastal and Manuel M Oliveira. Shared sampling for real-time alpha matting. *Computer Graphics Forum*, 29(2):575–584, 2010. 2

[21] Leo Grady, Thomas Schiwietz, Shmuel Aharon, and Rüdiger Westermann. Random walks for interactive alpha-matting. In *Proceedings of VIIP*, 2005. 2

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, 2017. 1, 2, 4, 7, 8

[23] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2011. 2

[24] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *International Conference on Computer Vision*, 2019. 2

[25] Guanqing Hu and James J. Clark. Instance segmentation based semantic matting for compositing applications. In *Conference on Computer and Robot Vision*, 2019. 2

[26] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring R-CNN. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[27] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[28] Zhanghan Ke, Kaican Li, Yurou Zhou, Qiuhua Wu, Xiangyu Mao, Qiong Yan, and Rynson W. H. Lau. Is a green screen really necessary for real-time portrait matting? *CoRR*, abs/2011.11961, 2020. 2

[29] Alexander Kirillov, Kaiming He, Ross B. Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 6

[30] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2006. 2

[31] Anat Levin, Alex Rav-Acha, and Dani Lischinski. Spectral matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2007. 2, 3

[32] Jizhizi Li, Jing Zhang, Stephen J. Maybank, and Dacheng Tao. End-to-end animal image matting. *CoRR*, abs/2010.16188, 2020. 5, 6

[33] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *AAAI Conference on Artificial Intelligence*, 2020. 2, 7, 8

[34] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[35] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[36] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. *CoRR*, abs/2108.11515, 2021. 1, 2

[37] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *European Conference on Computer Vision*, 2014. 5, 7

[38] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-Sheng Hua. Boosting semantic human matting with coarse annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[39] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. SGN: sequential grouping networks for instance segmentation. In *IEEE International Conference on Computer Vision*, 2017. 2

[40] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[41] Yuhao Liu, Jiake Xie, Xiao Shi, Yu Qiao, Yujie Huang, Yong Tang, and Xin Yang. Tripartite information mining and integration for image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7555–7564, October 2021. 2

[42] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *International Conference on Computer Vision*, 2019. 2

[43] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. Alphagan: Generative adversarial networks for natural image matting. In *British Machine Vision Conference*, 2018. 2

[44] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2, 6

[45] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[46] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision*, 2016. 2

[47] Dheeraj Singaraju and René Vidal. Estimation of alpha mattes for multiple image layers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(7):1295–1309, 2011. 3

[48] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Semantic image matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 7, 8

[49] Yanan Sun, Guanzhi Wang, Qiao Gu, Chi-Keung Tang, and Yu-Wing Tai. Deep video matting via spatio-temporal alignment and aggregation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 6

[50] Yu-Wing Tai, Jiaya Jia, and Chi-Keung Tang. Soft color segmentation and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(9):1520–1537, 2007. 3

[51] Jianchao Tan, Jyh-Ming Lien, and Yotam I. Gingold. Decomposing images into layers via rgb-space geometry. *ACM Trans. Graph.*, 36(1):7:1–7:14, 2017. 3

[52] Tiantian Wang, Sifei Liu, Yapeng Tian, Kai Li, and Ming-Hsuan Yang. Video matting via consistency-regularized graph neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[53] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: segmenting objects by locations. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision*, 2020. 2

[54] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems*, 2020. 2

[55] Ning Xu, Brian L. Price, Scott Cohen, and Thomas S. Huang. Deep image matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 2, 5, 6

[56] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan L. Yuille. Mask guided matting via progressive refinement network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 4, 7, 8

[57] Zijian Yu, Xuhui Li, Huijuan Huang, Wen Zheng, and Li Chen. Cascade image matting with deformable graph refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[58] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion CNN for digital matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2