Mask-guided Matting in the Wild

Kwanyong Park[1] Sanghyun Woo[1] Seoung Wug Oh[2] In So Kweon[1] Joon-Young Lee[2]

[1]KAIST [2]Adobe Research

Abstract

Mask-guided matting has shown great practicality compared to traditional trimap-based methods. The mask-guided approach takes an easily-obtainable coarse mask as guidance and produces an accurate alpha matte. To extend the success toward practical usage, we tackle mask-guided matting in the wild*, which covers a wide range of categories in their complex context robustly. To this end, we propose a simple yet effective learning framework based on two core insights: 1) learning a generalized matting model that can better understand the given mask guidance and 2) leveraging weak supervision datasets (e.g., instance segmentation dataset) to alleviate the limited diversity and scale of existing matting datasets. Extensive experimental results on multiple benchmarks, consisting of a newly proposed synthetic benchmark (Composition-Wild) and existing natural datasets, demonstrate the superiority of the proposed method. Moreover, we provide appealing results on new practical applications (e.g., panoptic matting and mask-guided video matting), showing the great generality and potential of our model.*

1. Introduction

Image matting aims to predict the opacity of objects, which enables precise separation from surrounding backgrounds. Due to the ill-posed nature of the task, many works [7, 13, 21, 27, 30, 48] have improved matting performance by relying on the manual guidance of a *trimap*. However, pixel-level annotation of foreground/background/unknown is extremely burdensome, restricting its usage in many practical applications such as image/video editing and film production. Recently, many efficient alternatives for user guidance have been proposed, including trimap-free [15, 32], additional background images [22, 34], scribble [43], and the user clicks [45].

Among them, the mask-guided approach [50] shows a great trade-off between performance and intensity of user interaction. It utilizes a coarse mask as guidance, which is much easier to obtain either manually or from off-the-shelf segmentation models [2, 10]. With only the coarse spatial prior, the mask-guided matting model [50] shows comparable or even better performance than the trimap-based competitors [13, 17, 21, 27, 48] on synthetic Composition-1k [48] and a real-world human matting dataset [50]. However, despite the encouraging results, we see the previous state-of-the-art model [50] struggles to obtain desirable alpha matte in complex real-world scenes (see Fig. 1).

Figure 1. **Qualitative Comparisons of MGMatting [50] and Ours in the wild.** The mask guidance is overlaid on images with blue color. Best viewed zoomed in.

With this observation, we tackle *mask-guided matting in the wild*. Specifically, we formulate unique setups and emerging challenges of the new task as follows: (1) We aim to handle objects in their complex context, reflecting the characteristics of natural images. The previous method [50] evaluates their model on iconic-object images [1, 29] where only a *single* object is in the center. As the model can easily find the target object in such images, the model's real instance discrimination ability is, in fact, veiled. On the contrary, in an 'in-the-wild' setting, it is crucial to precisely localize the target object from the given coarse/noisy mask guidance (*i.e.* mask awareness). (2) Our model targets to deal with diverse categories of objects in natural

images. Unlike most previous methods that improve generalization performance at the expense of category-specific regime (*e.g.*, limiting to humans [15] or animals [19]), we aim to understand distinctive matting patterns of vast categories. (3) Limited data problem makes the new setting more complicated. Due to the labeling complexity, annotating alpha matte for objects in common scenes, *e.g.,* the COCO dataset [24], is infeasible. As a sidestep, previous benchmarks [32, 48] extract the alpha matte and foreground colors from images with simple backgrounds. These are composited on various backgrounds [6, 24], and resulting samples are used to train and evaluate matting models. However, due to the inevitable composition artifacts, the models usually show limited generalization performance. In that sense, how to train and evaluate the in-the-wild matting model remains an open question.

Toward this goal, we propose a simple yet effective learning framework for a generalized mask-guided matting model. First, we investigate fundamental reasons for the poor generalization of the previous mask-guided matting model [50] and find that this is mainly from the training data generation process. Specifically, the previous composition process includes instance merging data augmentation, which merges several foreground objects into a single object. While this augmentation is effective in the trimap-based methods [21, 30, 41], it implicitly makes a negative bias for the mask-guided matting model to ignore the guidance. Thus, the model struggles to localize the target objects in complex natural scenes. We alleviate the bias by proposing an instance-wise learning objective, where the model is supervised to segment one of multiple instances according to the guidance. By doing so, the model learns strong semantic representation regarding complex relations and soft transitions between objects. Despite the simplicity of the proposal, this greatly improves performance in the wild.

Second, we explore a practical solution to make the mask-guided model handle various categories of objects robustly. Instead of scaling the matting dataset, we leverage a dataset with weak supervision [14, 46] (*i.e.*, instance segmentation dataset [24]), as the coarse instance masks are easier to obtain over the diverse categories of objects. To effectively hallucinate the fine supervision signal with the weak localization guidance, we come up with a self-training framework [36, 37, 47]. Specifically, a pseudo label is generated based on a weakly-augmented input (both image and instance mask annotation as guidance), which supervises the model prediction on a strongly augmented version of them. During self-training, the model is not only adapting to the in-the-wild scenario in a self-evolving manner but also being robust to noise in both image and guidance.

To verify the in-the-wild performance of mask-guided matting, we formally define an evaluation protocol involving multiple sub-benchmarks: Composition-Wild, AIM-500 [20], COCO [24]. We first design an in-the-wild extension of the popular synthetic Composition-1k benchmark [48], namely Composition-Wild. We simulate complex real-world images by compositing multiple foreground objects. To bring valuable insight on the failure cases of the model, we design sub-metrics for Composition-Wild. In addition, we use the AIM-500 dataset to establish quantitative results on natural images (*i.e.*, with no composition artifacts), although most images are iconic-object images with simple backgrounds. Finally, we provide qualitative outputs of our mask-guided matting model on the COCO dataset [24] which is one of the most representative in-the-wild datasets.

To summarize, we make the following main contributions. 1) To our best knowledge, it is the first work to explore mask-guided matting in the wild. 2) We develop a simple yet effective learning framework leveraging both composited and weak-guidance images. 3) We design an evaluation setup for the new task. 4) We initiate several interesting extensions: video and panoptic matting.

2. Related Works

Natural Image Matting. Most image matting methods require a trimap as additional input, which conveys pixel-level annotations of foreground, background, and unknown region. Traditional image matting methods can be categorized into two groups. Sampling-based methods [5, 8, 11, 35, 44] estimate the alpha matte of the unknown region based on sampled colors of foreground and background. Propagation-based methods [4, 12, 16–18, 38] propagate the neighboring known alpha values to unknown regions according to the affinity between pixels. Both methods primarily rely on color or low-level features, showing limitations in complicated scenarios.

To tackle these issues, many deep-learning based matting methods have been proposed. DIM [48] is a representative work which proposes a convolutional encoder-decoder matting network as well as a large-scale synthetic dataset (Composition-1k) to train the model. Many follow-up works have made tremendous improvements in diverse aspects, such as sophisticated loss design [7, 13, 28] and architectural advances by introducing attention mechanisms [21, 26] or transformer architecture [30].

There is a large volume of trials to relax the heavy user-supplied constraints. Some methods [32, 52] attempt to get rid of trimap. However, they show inferior performance and cannot be generalized to unseen objects in the real world. Sengupta *et al.* [34] utilize additional background images along with other lightweight priors (*e.g.*, segmentation mask) to perform the matting task. Wei *et al.* [45] introduce user click interaction to effectively eliminate the ambiguity of target object. Recently, MGMatting [50] proposes a mask-guided matting framework, where only the

easily obtainable coarse mask is needed as guidance.

In this paper, we further study mask-guided matting in an in-the-wild setting, inspired by the generality and practicality of the framework. Different from MGMatting, we target diverse categories of objects in their complex background, enabling real-world applications.

Class-specific Matting is a special type of matting that limits its target objects to one or few classes such as humans [3] or animals [19]. BSHM [25] leverages the coarse segmentation mask data to build a generalized human matting network. MODNet [15] presents a lightweight network architecture with decomposed multi-scale network designs. GFM [19] first proposes the task of animal matting and a real-world animal matting dataset. In general, known semantics effectively ease the difficulty of matting, and these methods show great generality without the necessity of additional guidance input. However, the specialized methods and trained model cannot be generalized to the in-the-wild setting since limited semantics do not cover diverse patterns of visual worlds.

Apart from previous efforts, we attempt to extend the short category regime of matting networks. For this purpose, we propose an efficient solution by leveraging the instance segmentation dataset [24], which covers diverse categories of objects with weak localization annotations. Given the weak guidance dataset, we study the challenges in developing an in-the-wild mask-guided matting model.

3. Mask-guided Matting in the Wild

In natural scenes, objects are rarely in isolation: diverse categories of entities interact with each other. We aim to robustly separate a target object from the surroundings in such scenes. Specifically, we formulate the problem in the form of mask-guided matting, namely mask-guided matting in the wild. Our matting model takes an easily-obtainable coarse mask as guidance, either manually drawn or predicted from the off-the-shelf segmentation models. With in-the-wild data, the model needs to understand semantics so that it can precisely localize the target object from the given noisy masks. In the meantime, the model also needs to have strong low-level feature representation so that it can capture fine details and accurate opacity of the diverse target matte.

To concrete the new challenges toward such an in-the-wild setting, we first study how the previous model [50] performs. We derive predictions with different mask guidance and summarize the results on Fig. 2. As shown in the upper example, in the synthetic Composition-1k [48] example, the previous model perfectly separates the foreground object regardless of the given guidance. Surprisingly, it produces an alpha matte of the foreground object when the mask of the background region is given as guidance. On the contrary, the model lacks generalization on

Figure 2. **Predictions of MGMatting [50] with different guidance.** (upper) synthetic Composition-1k image [48]. (bottom) Complex real world image [49]. Best viewed zoomed in.

complex real-world scenes (see bottom example). In particular, the model fails to localize the target object and makes wrong predictions on regions of similar color distributions (*e.g.*, shadows) or another nearby object. From the two representative examples, we make several observations: 1) The previous method yields a model overfitted to the (trained) synthetic matting dataset, which memorizes the limited patterns of foreground objects rather than semantically understanding the mask guidance for localizing and segmenting the objects; 2) Such a model fails to precisely localize and segment a target object in the wild; 3) Previously used synthetic benchmark, Composition-1k [48], neither represents the complex nature of the in-the-wild setting nor systematically blocks the shortcut of the overfitting issue. Thus the Composition-1k may not be a suitable benchmark for mask-guided matting in the wild.

To this end, we explore how to train a generalized mask-guided matting model given the limited scale and diversity of the matting dataset (Sec. 4), as well as revisit the evaluation benchmark to test the model (Sec. 5).

4. Proposed Method

The overview of the proposed training framework is illustrated in Fig. 3. Our framework involves three types of datasets: 1) matting dataset that includes accurate alpha matte and corresponding foreground colors; 2) background dataset, where the foreground objects are composited to create training samples; 3) weak supervision dataset that is introduced to extend the matting model to the in-the-wild setting. The main goal of our framework is to learn accurate matting ability from the synthetic composited images and generalize the knowledge to the in-the-wild domain. To this end, we first investigate how the model could learn generalized knowledge from the composited images (Sec. 4.1), then how the model is further improved with weak supervision data (Sec. 4.2).

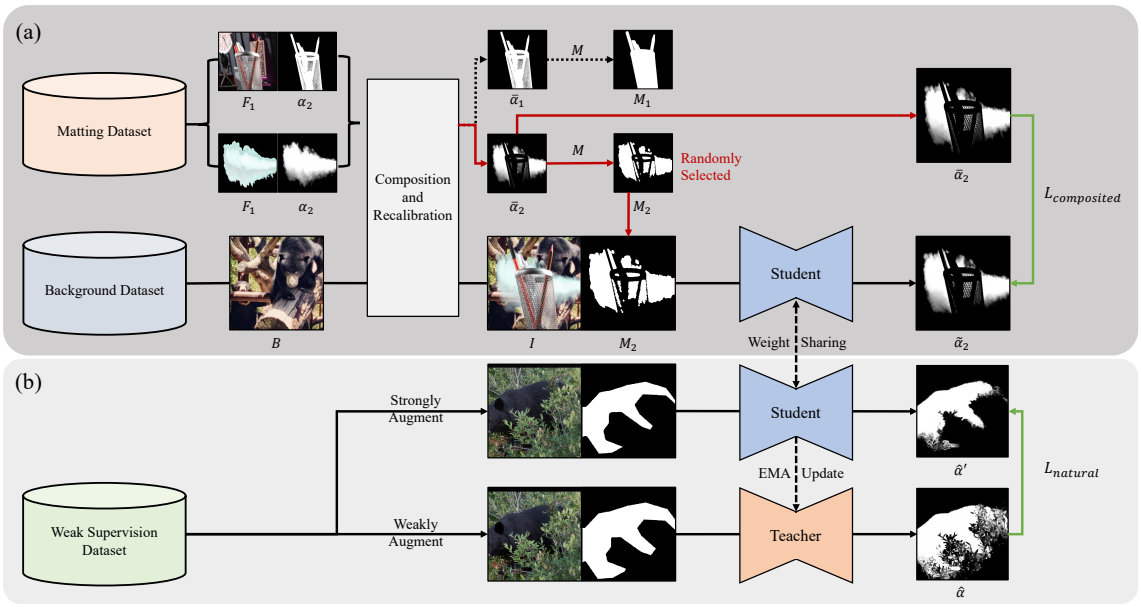


Figure 3. **The Overview of the Proposed Framework.** (a) Learning from composited images with instance-wise learning (Sec. 4.1). (b) Learning from weak-supervision images with self-training (Sec. 4.2).

4.1. Learning from Composited Images

As discussed in the Sec. 3, the previous model [50] shows poor generalization performance in the challenging real-world setting. Specifically, it often fails to identify the target object in a given guidance and produces a matting result for other objects. We found that the problem, the lack of mask awareness, comes from the previous data pipeline to generate training samples.

For a clear understanding, we first briefly summarize the previous training data generation step proposed in MGMatting [50]. Here, for clarity, generic data augmentation such as random-crop, resize, and affine-transform are omitted. A foreground object (consisting of foreground color F and alpha matte α) and a background image B are randomly sampled from the matting and background datasets, respectively. These are composited to form a training image I following the composition formulation as follows:

$$I = Composite(F, \alpha, B) = \alpha F + (1 - \alpha)B. \quad (1)$$

To simulate the mask guidance M, they first binarize the alpha matte with a threshold randomly sampled from 0 to 1, then apply dilation and/or erosion to it. In addition, they propose a stronger mask augmentation, CutMask [50], originally designed for robustness to noise in mask guidance. Like CutMix [51], a random-sized patch is selected, and the content is pasted on a random position.

In addition, MGMatting [50] adopted instance merging augmentation as most state-of-the-art trimap-based matting methods [21, 30, 41] did. Specifically, they select two random foreground objects, merge them into a single object, and composite it on a random background image. The instance merging augmentation effectively alleviates the limited scale and diversity of foreground objects in the matting dataset. However, in the context of mask-guided matting, we found that this augmentation hinders instance-wise understanding and makes a negative bias that the model eventually segments all the existing composited objects regardless of the given mask guidance. The different effects of the same augmentation are originated from the uncertainty of mask guidance. While the trimap brings pixel-level accurate localization of foreground/background/unknown regions, the mask guidance naturally conveys noisy information, and it is hard to localize them accurately. Under such uncertainty, the model tends to find the shortcut, segmenting all the composited objects rather than learning to utilize the uncertain mask guidance.

Instance-wise Learning. We systematically block the above shortcut and enforce the model to learn complex relations and soft transitions between objects. The idea is simple: supervise the model to segment one of them according to the guidance (See Fig. 3-(a)). Specifically, given the two foreground objects (colors F_1,F_2 and opacity α_1,α_2), we sequentially composite these foreground layers (the first object is on top.) on background image B as follows:

$$\begin{aligned} I &= Composite(F_1, \alpha_1, Composite(F_2, \alpha_2, B)) \\ &= \alpha_1 F_1 + \alpha_2(1 - \alpha_1)F_2 + (1 - \alpha_1)(1 - \alpha_2)B. \end{aligned} \quad (2)$$

Based on the above equation, we re-calibrate the fractional contribution of each foreground object to images at pixel level as follows: $\overline{\alpha}_1 = \alpha_1, \overline{\alpha}_2 = \alpha_2(1 - \alpha_1)$. We randomly

select an index $i_{rand} \in \{1, 2\}$ of a target object. With the guidance $M(\overline{\alpha}_i)$, the model predicts the opacity of the target object $\widetilde{\alpha}_i$, which is supervised by corresponding ground-truth alpha matte $\overline{\alpha}_i$. Such instance-wise training makes the model learn fractional differentiation between objects according to the given coarse mask guidance.

4.2. Learning from Images with Weak Supervision

Due to the annotation difficulties, the labels of the alpha matte are practically infeasible to collect for large quantities and diverse categories. By comparison, coarse mask annotations are more abundant and accessible, thus extending the dataset scale and taxonomies is easier to achieve. Motivated by this, we leverage the weak supervision dataset (*i.e.*, instance segmentation [24]), and investigate how these new data benefit the generalization of the mask-guided matting model. The main challenge lies in how to get matting supervision because the instance mask labels are noisy in fine details and do not include opacity annotations.

To effectively tackle this challenge, we design a self-training framework (See Fig. 3-(b)). The framework generates pseudo matting labels $\widehat{\alpha}$ under the guidance of instance masks, and the pseudo labels supervise the matting model on natural images. To obtain high-quality pseudo labels, we adopt several design choices as follows.

Teacher-Student Framework. Motivated by the success in self-/semi-supervised learning [9, 42], a teacher network is introduced, which is a slowly advancing version of a student model via the exponential moving average (EMA). We utilize stable predictions of the teacher network as the pseudo labels $\widehat{\alpha}$ and guide the student model during training.

Weakly-augmented Input. The teacher network takes weakly-augmented samples to generate the pseudo labels. We only apply standard geometric augmentation, such as affine transform and random crop, for both image and mask guidance. Despite the noise in fine details, the ground-truth instance mask conveys a strong localization cue to the model, resulting in high-quality pseudo labels.

For strongly-augmented inputs, the student model produces predictions $\widehat{\alpha}'$, which are supervised to be the same as the pseudo labels $\widehat{\alpha}$ obtained from the teacher model. To form the strongly-augmented version, we design two types of augmentations: image and guidance perturbations.

Image Perturbations. Previous self-training-based methods [36, 37, 47] often employ region-level augmentation (e.g., Gaussian blur) to generate the strongly-augmented images. However, in the matting task, these augmentations are unsuitable since they interpolate the color value across the pixels and disturb each object's fractional contribution. This results in a mismatch of ground-truth alpha matte between the teacher and student inputs, lowering the effectiveness of self-training. We instead adopt linear pixel-level augmentations such as additive Gaussian noise and linear contrast.

Guidance Perturbations. We also modulate the reliability of mask guidance via dilation and/or erosion of the instance mask. Considering the diverse size of objects in natural scenes, we decide the kernel size based on the size of the object. Specifically, we set the kernel size as $\beta\%$ of the shorter side of an object bounding box. The guidance perturbation is crucial to not only the success of self-training but also the robustness against common noise in mask guidance.

4.3. Unified Training

As shown in Fig. 3, under the proposed framework, the (student) mask-guided matting model can jointly learn from composited images with ground-truth labels and natural images with weak supervision. The final loss function is a combination of losses on both images as: $L_{final} = L_{composited} + L_{natural}$.

For the composited images, we use the same loss function as MGMatting [50], which is the summation of l_1 regression loss, composition loss [48], and Laplacian loss [13]. Thus, $L_{composited} = L_{l1}(\widetilde{\alpha}_i, \overline{\alpha}_i) + L_{comp}(\widetilde{\alpha}_i, \overline{\alpha}_i) + L_{lap}(\widetilde{\alpha}_i, \overline{\alpha}_i)$. To calculate the composition loss, we use ground-truth foreground color and alpha matte for non-target objects. For the natural images, we use l_1 regression loss as follows: $L_{natural} = L_{l1}(\widehat{\alpha}', \widehat{\alpha})$.

5. Benchmarks

There are several public matting benchmarks [19, 20, 32, 39, 48, 50] which provide high-quality alpha matte. However, none of them reflects the complexity of in-the-wild images and the consisting samples only contain a single object with simple backgrounds. Thus, we carefully design a new evaluation setup to facilitate an extensive evaluation of the mask-guided matting model in in-the-wild scenarios. The new setup consists of multiple benchmarks, and we describe the distinctive role of each benchmark below.

Composition-Wild. We first propose an in-the-wild extension of standard synthetic benchmark Composition-1k [48], namely Composition-Wild. We simulate the complex context of in-the-wild images by compositing multiple foreground objects onto a background image, similar to Eq.(2). The same foreground object data and background dataset (*i.e.,* PASCAL VOC [6]) are utilized as Compostion-1k. We report the performance over the forefront object from each sample unless otherwise specified. In addition, we design comprehensive metrics to allow in-depth analysis.

- SAD_{FG} reports the SAD (Sum of Absolute Difference) over the target object region where has non-zero ground-truth alpha matte. It indicates how well the prediction captures fine matting details of target objects (*i.e.,* **Target Detail Quality**).

- SAD_{BG} quantifies the wrong predictions on other objects or background regions by measuring the SAD over the non-target object region. It represents the capability of target object localization from the noisy mask guidance (*i.e.,* **Localization**).
- SAD_{OCC} measures the SAD error of occluded objects. It indicates how well the model deals with occluded objects (*i.e.,* **Occlusion handling**).

AIM-500 [20] is a natural image matting benchmark that contains diverse categories of objects. The performance evaluation on the AIM benchmark denotes whether the model generalizes well to natural images rather than overfitting to the synthetic distribution of composited images. However, AIM is limited at capturing matting performance in complex contexts since they only contain images with simple backgrounds. In that regard, the evaluations on Composition-Wild and AIM complement each other and provide a holistic understanding of the mask-guided matting model in the wild.

COCO [24] is a representative example of in-the-wild images. While we cannot access ground-truth alpha mattes on COCO, we conduct qualitative comparisons over diverse objects in their natural context.

6. Experiments

6.1. Implementation Details

Architecture details. We build our learning framework with a state-of-the-art mask-guided matting network, PRN from MGMatting [50], which adopts a standard encoder-decoder structure [21, 33] and progressively refines the output through the decoding process.

Training details. In practice, we adopt the two-stage training: pre-training on composited images and fine-tuning on both composited and natural images. This ensures the quality of pseudo labels at the beginning of the self-training, resulting in better performance. Following the previous works [21, 27, 30, 48], we use DIM [48] and COCO [24] as matting and background datasets, respectively. COCO dataset [24] is also utilized for a weak guidance dataset. We set the kernel ratio β as 0.2.

Evaluation details. We report the widely used four matting metrics, SAD (Sum of Absolute Differences), MSE (Mean Squared Error), Grad (Gradient), and Conn (Connectivity), over the whole image, when the ground-truth alpha matte is available. To simulate the coarse mask guidance, we binarize the alpha matte with threshold 0.5 and then dilate the mask with kernel size 25×25, similar to [53]. We use a single trained model for all the benchmark evaluations and real-world applications.

Figure 4. **Qualitative Results on COCO dataset.** The guidance are either manually drawn [24] (upper) or a prediction of the off-the-self model [10] (bottom). Best viewed zoomed in.

Figure 5. Effect of Instance-wise Learning.

6.2. Qualitative Comparisons

Fig. 1 and Fig. 4 illustrate the qualitative comparisons on images in the COCO validation sets. We test the models under challenging scenarios such as occlusions, an unseen object in a matting dataset, complex surrounding context, and large interior errors on mask guidance. Compared with the previous mask-guided matting model [50], our model shows significantly better localization of a target object, as well as, produces much more accurate alpha mattes for challenging data. More qualitative comparisons can be found in the supplementary material.

6.3. Ablation Study and Analysis

Quantitative evaluations on Composition-Wild and AIM-500 datasets are shown in Table 1.

Effectiveness of Instance-wise Learning. We first analyze how instance-wise learning contributes to learn generalized matting capability from composited images. We set the two baseline models: (naive) MGMatting and MGMatting*. For the (naive) MGMatting, we reproduce the model with the same training recipes proposed in MGMatting [50]. In addition, we report the stronger baseline score of MGMatting* by excluding a heavy guidance augmentation, CutMask [50]. We empirically found that CutMask

Method	Composition-Wild SAD	MSE	Grad	Conn	SAD_{FG}	SAD_{BG}	SAD_{OCC}	AIM-500 [20] SAD	MSE	Grad	Conn
Stage 1: Learning with Composited Images											
MGMatting [50]	583.36	0.1328	139.84	52.54	65.51	517.85	842.88	71.91	0.0268	23.37	21.97
MGMatting* [50]	389.15	0.0727	141.80	44.21	55.74	333.42	524.99	26.18	0.0056	15.81	14.53
+Instance-wise (Ours)	67.27	0.0058	42.35	43.52	49.50	17.77	59.59	21.12	0.0038	16.88	14.72
Stage 2: Learning with Composited and Weak-guidance Images											
Segmentation-training	69.24	0.0069	57.08	46.56	54.36	14.88	65.42	20.94	0.0043	20.46	14.20
Self-training (Ours)	58.16	0.0046	39.04	41.37	47.32	10.84	53.02	16.72	0.0030	14.68	12.02

Table 1. Quantitative evaluation on Composition-Wild and AIM-500 [20]. * denotes the improved version of MGMatting [50]. "Instance-wise" represents instance-wise learning. A lower score is better in all the metrics.

makes substantial *random* localization errors in guidance and lowers the reliability of mask guidance during training. Training with such guidance makes models ignore the mask guidance and overfit to synthetic samples, resulting in inferior in-the-wild performance. On top of MGMatting*, the "Instance-wise Learning" strategy drastically benefits the mask-guided matting model. Unlike the baseline model that segments all the composited objects (Fig. 5-(c)), the proposed strategy allows the model to well discriminate the target object (Fig. 5-(d)). Our model faithfully utilizes the coarse mask guidance and shows a huge gain in localization performance (see SAD_{BG} metric). Although the strategy assumes multi-object scenarios and learns the complex relations between them, it also contributes to the better matting details for an object (see SAD_{FG} in Composition-Wild and SAD in AIM). By design, the instance-wise learning naturally improves occlusion handling by learning the order of layered foreground objects and their fractional contributions.

Effectiveness of Self-training. To extend the mask-guided matting model to the in-the-wild setting, we leverage the weak-supervision dataset under the proposed self-training framework. We compare our self-training framework to a baseline, namely "segmentation-training". This baseline is motivated by previous human-matting methods [23, 25]. It attempts to learn generalized matting capability via simultaneous segmentation training on natural images. To be specific, the model takes strongly augmented instance masks as guidance and is supervised to recover ground-truth instance masks through the additional segmentation head. As deteriorated performance indicates, the segmentation-training strategy fails to scale up to the in-the-wild setting. While it enhances the localization of target objects, supervision from binary instance masks hinders learning diverse matting patterns in the wild (see SAD_{FG} metric), especially for transparent objects. On the contrary, our self-training framework successfully generates the soft-pseudo label and utilizes it as a direct supervision to matting tasks, resulting in huge improvements in all the metrics.

Analysis on Perturbation in Self-training. To demonstrate the effectiveness of perturbations in self-training, we ablate the image and guidance perturbations separately. The results are summarized in Table 2. As degraded performance denoted, encouraging robustness against image and guidance perturbations is crucial to the success of the self-training.

Method	Composition-Wild SAD	MSE	SAD_{FG}	SAD_{BG}	AIM-500 SAD	MSE
Ours	58.16	0.0046	47.32	10.84	16.72	0.0030
No Image.	60.08	0.0048	45.19	14.89	18.57	0.0039
No Guide.	60.14	0.0049	47.95	12.19	18.27	0.0035

Table 2. **Ablation Study on Perturbations in Self-training.**

We further investigate whether the matting-specific design of image perturbation is necessary. To this end, we train three variants with different configurations of image perturbations: (1) Pixel-level linear augmentation (Pixel), (2) Region-level augmentation (Region) and (3) Both of them (Pixel+Region). Below, we report the SAD metric for Composition-Wild and AIM-500, respectively.

Pixel	Region	Pixel+Region
58.16 / 16.72	69.8 / 25.96	61.88 / 19.00

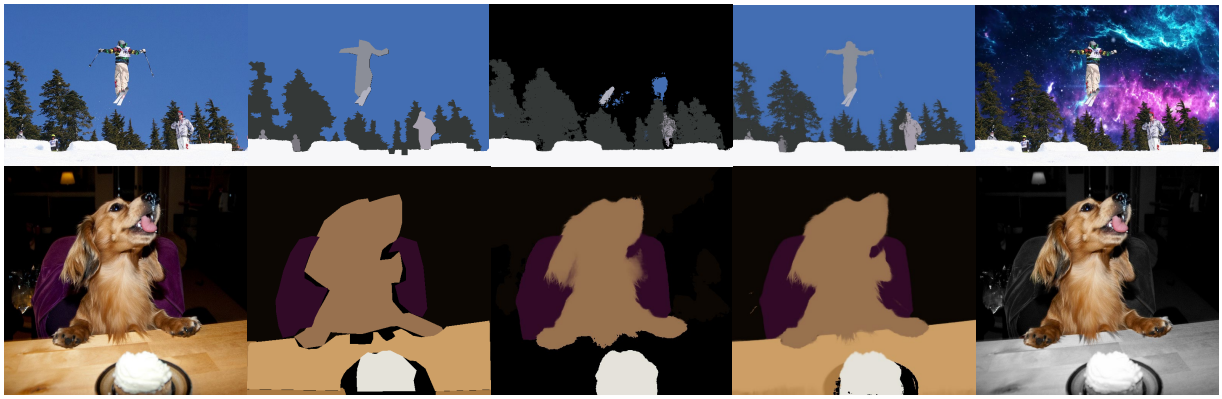
We can observe that the region-level augmentation, which is widely used in other tasks [37, 47], disturbs the effects of self-training objectives. As discussed in Sec. 4.2, this is because the region-level augmentation breaks the underlying opacity of objects and make a mismatch between inputs of the student and the teacher networks. On the contrary, pixel-level linear augmentation brings clear improvements.

7. Real-world Applications

To show the generality and potential of the in-the-wild mask-guided matting model, we illustrate two new applications: mask-guided video matting and panoptic matting. Fig. 6 and Fig. 7 show examples of each application.

(a) Frame (b) Guidance (c) MGMatting (d) Ours (e) Editing Result

Figure 6. **Extension to Mask-guided Video Matting.** We use propagated masks [31] as guidance.



(a) Image (b) Guidance (c) MGMatting (d) Ours (e) Editing Result

Figure 7. **Extension to Panoptic Matting.** Hand drawn panoptic masks [24] are utilized as guidance. We illustrate independent regions with different colors. The black color represents ignored regions of labels or predictions.

7.1. Mask-guided Video Matting

Current video matting solutions [40, 53] require multiple frames of trimap to produce an alpha matte for a video. Therefore, required user interaction is more burdensome in the video domain. With the mask-guided matting model, we explore a more practical scenario where only a single instance mask is given in the first frame. In this challenging setting, we first propagate the given mask to the rest of the video frames using the video object segmentation model [31]. Then, the mask-guided matting model utilizes the propagated mask as guidance and produces alpha matte in a frame-by-frame manner. Our model not only corrects the noise in propagated masks but also predicts reasonable opacity in challenging cases such as motion blur. With the prediction, we also provide visually appealing video editing results (*i.e.,* background replacement).

7.2. Panoptic Matting

Panoptic segmentation aims to parse an image into non-overlapping regions of stuff and things. By the definition of the task format, the soft transition between the regions is overlooked. We explore whether our mask-guided matting model could provide such understanding from panoptic segmentation results. Given a panoptic mask, our model considers the segment of each region separately and produces the corresponding alpha matte. We form panoptic matting results by aggregating the predictions. Unlike the baseline, our model successfully refines the fine details and captures the soft transition between the regions. Interestingly, while the model is trained on objects (*i.e.,* things), our method is well generalized on the stuff regions (See sky region in Fig. 7). Panoptic matting enables us to do region-level image editing and the qualitative results are shown in Fig. 7.

8. Conclusion

In this paper, we propose a simple yet effective learning framework for *mask-guided matting in the wild*. We first introduce instance-wise learning to learn generalized matting ability from the composited images. Then, the model is further extended to handle diverse objects under the self-training on weak-guidance images. Extensive experiments on the newly proposed evaluation benchmark provide an in-depth understanding of the unique challenges. Using our approach, we instantiate new practical applications: mask-guided video matting and panoptic matting. We hope our proposals inspire future work on mask-guided matting.

Acknowledgement This work was partially supported by the NRF (NRF-2020M3H8A1115028, FY2021) and Samsung Electronics Co., Ltd (G01200447).

References

[1] Tamara L Berg and Alexander C Berg. Finding iconic images. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2009. 1

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[3] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 618–626, 2018. 3

[4] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2175–2188, 2013. 2

[5] Yung-Yu Chuang, Brian Curless, David H Salesin, and Richard Szeliski. 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, pages II–II. IEEE, 2001. 2

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 5

[7] Marco Forte and François Pitié. f, b, alpha matting. *arXiv preprint arXiv:2003.07711*, 2020. 1, 2

[8] Eduardo SL Gastal and Manuel M Oliveira. Shared sampling for real-time alpha matting. In *Computer Graphics Forum*, volume 29, pages 575–584. Wiley Online Library, 2010. 2

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 5

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 6

[11] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In *CVPR 2011*, pages 2049–2056. IEEE, 2011. 2

[12] Kaiming He, Jian Sun, and Xiaoou Tang. Fast matting using large kernel matting laplacian matrices. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2165–2172. IEEE, 2010. 2

[13] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4130–4139, 2019. 1, 2, 5

[14] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4233–4241, 2018. 2

[15] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W.H. Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *AAAI*, 2022. 1, 2, 3

[16] Philip Lee and Ying Wu. Nonlocal matting. In *CVPR 2011*, pages 2193–2200. IEEE, 2011. 2

[17] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):228–242, 2007. 1, 2

[18] Anat Levin, Alex Rav-Acha, and Dani Lischinski. Spectral matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(10):1699–1712, 2008. 2

[19] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision*, 130(2):246–266, 2022. 2, 3, 5

[20] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235*, 2021. 2, 5, 6, 7

[21] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11450–11457, 2020. 1, 2, 4, 6

[22] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 1

[23] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 238–247, 2022. 7

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 3, 5, 6, 8

[25] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-sheng Hua. Boosting semantic human matting with coarse annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8572, 2020. 3, 7

[26] 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*, pages 7555–7564, 2021. 2

[27] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3266–3275, 2019. 1, 6

[28] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. Alphagan: Generative adversarial networks for natural image matting. *arXiv preprint arXiv:1807.10088*, 2018. 2

[29] Stephen Palmer. Canonical perspective and the perception of objects. *Attention and performance*, pages 135–151, 1981. 1

[30] GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, and Nojun Kwak. Matteformer: Transformer-based image matting via prior-tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11696–11706, 2022. 1, 2, 4, 6

[31] Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Per-clip video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1352–1361, 2022. 8

[32] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13676–13685, 2020. 1, 2, 5

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 6

[34] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2291–2300, 2020. 1, 2

[35] Ehsan Shahrian, Deepu Rajan, Brian Price, and Scott Cohen. Improving image matting using comprehensive sampling sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 636–643, 2013. 2

[36] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2, 5

[37] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 2, 5, 7

[38] Jian Sun, Jiaya Jia, Chi-Keung Tang, and Heung-Yeung Shum. Poisson matting. In *ACM SIGGRAPH 2004 Papers*, pages 315–321. 2004. 2

[39] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Semantic image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11120–11129, 2021. 5

[40] Yanan Sun, Guanzhi Wang, Qiao Gu, Chi-Keung Tang, and Yu-Wing Tai. Deep video matting via spatio-temporal alignment and aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 8

[41] Jingwei Tang, Yagiz Aksoy, Cengiz Oztireli, Markus Gross, and Tunc Ozan Aydin. Learning-based sampling for natural image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3055–3063, 2019. 2, 4

[42] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 5

[43] Jue Wang and Michael F Cohen. An iterative optimization approach for unified image segmentation and matting. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 936–943. IEEE, 2005. 1

[44] Jue Wang and Michael F Cohen. Optimized color sampling for robust matting. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[45] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Hanqing Zhao, Weiming Zhang, and Nenghai Yu. Improved image matting via real-time user clicks and uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15374–15383, 2021. 1, 2

[46] Sanghyun Woo, Kwanyong Park, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Bridging images and videos: A simple learning framework for large vocabulary video object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, pages 238–258. Springer, 2022. 2

[47] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021. 2, 5, 7

[48] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2970–2979, 2017. 1, 2, 3, 5, 6

[49] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 3

[50] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1154–1163, 2021. 1, 2, 3, 4, 5, 6, 7

[51] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 4

[52] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7469–7478, 2019. 2

[53] Yunke Zhang, Chi Wang, Miaomiao Cui, Peiran Ren, Xuansong Xie, Xian-Sheng Hua, Hujun Bao, Qixing Huang, and Weiwei Xu. Attention-guided temporally coherent video object matting. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5128–5137, 2021. 6, 8