Mask-guided Matting in the Wild

Kwanyong Park\textsuperscript{1}  Sanghyun Woo\textsuperscript{1}  Seoung Wug Oh\textsuperscript{2}  In So Kweon\textsuperscript{1}  Joon-Young Lee\textsuperscript{2}

\textsuperscript{1}KAIST  \textsuperscript{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).

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 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 concretize 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 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).
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 $\alpha$) 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:

$$I = \text{Composite}(F, \alpha, B) = \alpha F + (1 - \alpha)B.$$  \hspace{1cm} (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 $\alpha_1, \alpha_2$), we sequentially composite these foreground layers (the first object is on top.) on background image $B$ as follows:

$$I = \text{Composite}(F_1, \alpha_1, \text{Composite}(F_2, \alpha_2, B)) = \alpha_1 F_1 + \alpha_2 (1 - \alpha_1) F_2 + (1 - \alpha_1)(1 - \alpha_2) B.$$ \hspace{1cm} (2)

Based on the above equation, we re-calibrate the fractional contribution of each foreground object to images at pixel level as follows: $\bar{\alpha}_1 = \alpha_1, \bar{\alpha}_2 = \alpha_2(1 - \alpha_1)$. We randomly...
select an index $i_{rand} \in \{1, 2\}$ of a target object. With the guidance $M(\pi_{t})$, the model predicts the opacity of the target object $\tilde{\alpha}_{t}$, which is supervised by corresponding ground-truth alpha matte $\pi_{t}$. 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 $\hat{\alpha}$ 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 $\hat{\alpha}$ 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 $\hat{\alpha'}$, which are supervised to be the same as the pseudo labels $\hat{\alpha}$ 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 $3\%$ 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: $L_{final} = L_{composed} + L_{natural}$.

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, $L_{composed} = L_{l_{1}}(\hat{\alpha}_{t}, \tilde{\alpha}_{t}) + L_{comp}(\tilde{\alpha}_{t}, \pi_{t}) + L_{lap}(\tilde{\alpha}_{t}, \pi_{t})$. 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: $L_{natural} = L_{l_{1}}(\hat{\alpha'}, \tilde{\alpha})$.

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 Composition-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\textsubscript{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 over-fitting 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 $\beta$ 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$\times$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.

(a) Image (b) Guidance (c) MGMatting* (d) +IWL

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
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Composition-Wild</th>
<th>AIM-500 [20]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SAD</td>
<td>MSE</td>
</tr>
<tr>
<td>Stage 1: Learning with Composited Images</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGMatting [50]</td>
<td>58.36</td>
<td>0.1328</td>
</tr>
<tr>
<td>MGMatting* [50]</td>
<td>389.15</td>
<td>0.0727</td>
</tr>
<tr>
<td>+Instance-wise (Ours)</td>
<td>67.27</td>
<td>0.0058</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage 2: Learning with Composited and Weak-guidance Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segmentation-training</td>
</tr>
<tr>
<td>Self-training (Ours)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Composition-Wild</th>
<th>AIM-500</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SAD</td>
<td>MSE</td>
</tr>
<tr>
<td>Ours</td>
<td>58.16</td>
<td>0.0046</td>
</tr>
<tr>
<td>No Image.</td>
<td>60.08</td>
<td>0.0048</td>
</tr>
<tr>
<td>No Guide.</td>
<td>60.14</td>
<td>0.0049</td>
</tr>
</tbody>
</table>

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.
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


[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


[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


[53] Yunke Zhang, Chi Wang, Xiaowu Cao, 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