Most matting research resorts to advanced semantics to achieve high-quality alpha mattes, and a direct low-level features combination is usually explored to complement alpha details. However, we argue that appearance-agnostic integration can only provide biased foreground (FG) details and that alpha mattes require different-level feature aggregation for better pixel-wise opacity perception. In this article, we propose an end-to-end hierarchical and progressive attention matting network (HAttMatting++), which can better predict the opacity of the FG from single RGB images without additional input. Specifically, we utilize channel-wise attention (CA) to distill pyramidal features and employ spatial attention (SA) at different levels to filter appearance cues. This progressive attention mechanism can estimate alpha mattes from adaptive semantics and semantics-indicated boundaries. We also introduce a hybrid loss function fusing structural similarity, mean square error, adversarial loss, and setry supervision to guide the network to further improve the overall FG structure. In addition, we construct a large-scale and challenging image matting dataset comprised of 59,000 training images and 1,000 test images (a total of 646 distinct FG alpha mattes), which can further improve the robustness of our hierarchical and progressive aggregation model. Extensive experiments demonstrate that the proposed HAttMatting++ can capture sophisticated FG structures and achieve state-of-the-art performance with single RGB images as input.

CCS Concepts: • Computing methodologies → Image segmentation; Artificial intelligence; Computer vision; Computer vision problems;

Additional Key Words and Phrases: Image matting, alpha matte, hierarchical, progressive, attention

Y. Qiao and Y. Liu contributed equally to this research.
This work was supported in part by the National Natural Science Foundation of China under Grant 61972067/U21A20491, the Innovation Technology Funding of Dalian (2020J26GX036), the Open Research Fund of Beijing Key Laboratory of Big Data Technology for Food Safety (Project No. BTBD-2018KF), and the National Key Research and Development Program of China (No. 2022ZD0210500).
Authors’ addresses: Y. Qiao, Y. Liu, Y. Wang (corresponding author), and X. Yang, Dalian University of Technology, 2 Linggong Road, Dalian, Liaoning, China, 116024; emails: qiaoyu2020@mail.dlut.edu.cn, yuhaoliu7456@gmail.com, {wyx, xinyang}@dlut.edu.cn; Z. Wei (corresponding author), CAS Key Laboratory of Molecular Imaging, Institute of Automation, Beijing, China, 100190; email: ziqi.wei@ia.ac.cn; Q. Cai, Beijing Technology and Business University, Beijing, China, 100048; email: caiq@th.btbu.edu.cn; G. Zhang, Wonxing Technology, 14 Haitian 2 Road, Shenzhen, Guangdong, China, 518110; email: zhangguofeng@wonxing.com.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
1551-6857/2023/02-ART52 $15.00
https://doi.org/10.1145/3540201

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 19, No. 2, Article 52. Publication date: February 2023.
1 INTRODUCTION

Image matting refers to precisely estimating the foreground (FG) opacity from an input image. This problem and its inverse process (known as image composition) have been well studied by both academia and industry. Image matting serves as a prerequisite technology for a broad set of applications, including online image editing, mixed reality, and film production. Formally, it is modeled by solving the following image synthesis equation:

\[ I_z = \alpha_z F_z + (1 - \alpha_z) B_z, \quad \alpha_z \in [0, 1], \]  

(1)

where \( z \) denotes the pixel position in the input image \( I \), \( \alpha_z \), \( F_z \), and \( B_z \) refer to the alpha estimation, FG, and background (BG) at pixel \( z \) separately. The problem is highly ill posed. For each pixel in a given RGB image, the only observed value is the input \( I_z \), which has R, G, B channels. \( F_z \) and \( B_z \) also have three channels, and the alpha value \( \alpha_z \) has one channel. Thus, there are seven values need to be solved, but only three values are known.

In this article, we argue that the structure of the FG resides in two aspects: adaptive semantics and refined boundaries, corresponding to \( \alpha_z = 1 \) and \( \alpha_z \in (0, 1) \) in Equation (1), respectively. To solve such a pixel-wise and ill-posed regression problem, most matting methods usually introduce user-provided trimaps as assistant input. The trimap is composed of black, gray, and white, representing the BG, transition region, and absolute FG separately. The transition region indicates FG boundaries, combined with an absolute FG to jointly guide matting algorithms. Given an RGB image and the corresponding trimap, traditional matting methods [7, 19] explore handcrafted representations or color distribution to predict an alpha matte. However, the color features are inapplicable for structure representation, possibly resulting in artifacts and loss of details when the FG and BG have indistinguishable colors.

Xu et al. formally import deep learning into matting (DIM), and they argue that matting objects share a common structure that can be represented by high-level features. It is noted that DIM involves RGB images in the refinement stage to combine advanced semantics with appearance cues. Advanced semantics indicate the FG category and profiles, whereas appearance cues reveal texture and boundary details. Subsequent matting networks [3, 10, 15, 25] mostly design complicated architectures for advanced semantics extraction and involve appearance cues from the input images or low-level CNN features. There are also some other forms of additional input, like prepared BG images [22, 35] and segmentation masks [57]. However, such research has an obvious dependency on that type of auxiliary and expensive input for better appearance cues and advanced semantics extraction. A well-defined trimap involves fussy manual labeling efforts and time consumption, which is difficult for novice users in practical applications. Correspondingly, prepared BG images or masks share certain constraints to maintain consistency.

Some matting works [6, 8] implement single-image predictions with intermediate segmentation as pseudo trimaps, which partly depress the precision of alpha mattes. The Late Fusion [58] blends FG and BG weight maps from the segmentation network with initial CNN features to predict alpha mattes. However, when semantic segmentation encounters difficulties, the Late Fusion will compromise. The preceding methods directly feed advanced semantics and appearance cues to the optimization or fusion stage. The semantic information is extracted from deep layers to describe
Fig. 1. The alpha mattes produced by our HAttMatting++ and the corresponding composition results with some internet images as a new BG (as shown in the left bottom of the input image).

FG profiles, whereas the appearance cues are mostly from input images or initial layers to focus on texture details and refined boundaries.

In this article, we hold that the advanced semantics and initial appearance cues require essential adaption before combination, and some secondary appearance cues (e.g., middle-level features or parts, hands, leaves, ears) can also be explored to complement potential FG details. On the one hand, natural image matting is a regression problem substantially and not entirely dependent on image semantics. On the other hand, although appearance cues retain sophisticated image texture, they contain the details outside the FG. In addition, some potential FG details can be further enhanced by secondary appearance cues, although the input images or initial CNN features can provide sufficient low-level information. Consequently, directly fusing advanced semantics and initial appearance cues is slightly flawed, and existing matting networks neglect the profound excavation and distillation of such hierarchical features.

This article explores advanced semantics and appearance cues synthetically and contributes an end-to-end hierarchical and progressive attention matting network (HAttMatting++) enabling a progressive manner to aggregate different-level features. Advanced semantics can provide FG category and profiles, whereas appearance cues furnish texture and boundary details. To deeply integrate this hierarchical structure, we perform channel-wise attention on advanced semantics to select matting-adapted features and employ spatial attention on multi-level appearance cues in a top-down manner to filtrate image texture details, then finally aggregate them progressively to predict alpha mattes. Moreover, a hybrid loss composed of mean square error (MSE), structural similarity (SSIM) [46], adversarial loss [13], and sentry supervision is exploited to optimize the whole network training. Extensive experiments show that our attention-guided hierarchical structure aggregation can attain high-quality alpha mattes with only RGB images as input (Figure 1).

The main contributions of this article are as follows:

- We present HAttMatting++, which can achieve high-quality alpha mattes with only RGB images. HAttMatting++ can process variant opacity with different types of objects and has no dependency on auxiliary input. HAttMatting++ can achieve state-of-the-art performance compared to the existing single-input matting methods.

- We present a hierarchical and progressive attention mechanism that can aggregate advanced pyramidal features and different-level appearance cues to produce adaptive semantics and
variant FG details for alpha perception. We resort to a hybrid loss consisting of MSE, SSIM, adversarial loss, and sentry supervision to provide efficient guidance for model training.

- We build a large-scale and challenging matting dataset consisting of 59,600 training images and 1,000 test images (646 distinct FG alpha mattes in total). To the best of our knowledge, this is the largest matting dataset with diverse FG objects, which can further improve the robustness of HAttMatting++. Meanwhile, we have made our dataset publicly available, which can promote further research and evaluation.

A preliminary version of this work was presented in earlier work [29]. In this article, we basically expand our previous work and summarize the main differences as follows.

**Progressive structure aggregation for multi-scale appearance cues exploration.** As shown in Figure 2, instead of directly extracting appearance cues from block1 as in earlier work [29], we employ a top-down spatial attention mechanism to aggregate pyramidal features with progressive appearance cues. Specifically, we consider the low-level feature maps from block2 in ResNeXt [50] as secondary appearance cues, and they are attended by our proposed appearance cues filtration model, then aggregated with pyramidal features. We argue that such secondary appearance cues can be used as a complement to improve some middle-level image texture or details (the leaves, hands, etc.).

**Sentry supervision for advanced semantics enhancement.** We import sentry supervision after the pyramidal features distillation module. As we advocated in our previous conference paper, the aggregation between matting-adapted semantics and appearance cues can be primarily devoted to the alpha matte generation. The advanced matting-adapted semantics play a significant role in the structure aggregation process since the appearance cues filtration module also requires them to suppress redundant BG information. Hence, we add sentry supervision after the pyramidal features as the auxiliary guidance to enhance adaptive semantics. On the one hand, it can further improve pyramidal features distillation and select matting-adapted semantics. On the other hand, it can provide better guidance for subsequent spatial attention and final aggregation. In addition, the whole training procedure and convergence progress can be optimized due to the backward propagation of the sentry supervision.

**More extensive experiments for further analysis and evaluation.** Compared to our previous conference paper, we conduct more experiments to demonstrate our model. Comprehensive ablation study experiments are also performed to compare the potential variation of the network architecture and prove the effectiveness of different components on the baseline network. In addition, we evaluate the proposed HAttMatting++ on more real-world images, and the high-quality alpha mattes can prove the versatility of our model.

2 RELATED WORK

Deep learning brings a major evolution for natural image matting with a high representation of FG structure, and we briefly review image matting from two categories: traditional and deep learning methods.

**Traditional matting.** Existing matting methods mostly achieve FG opacity by virtue of additional input: trimap or scribbles. The trimap is composed of FG, BG, and transition to partition the input image, whereas the scribbles indicate these three category regions by several user-specified scribbles. The transition region signifies FG boundaries, which is the key point for image matting. Although scribbles approaches [19, 20, 54] are convenient for novice users, they significantly deteriorate the alpha matte since there is insufficient information that can be referenced. Therefore, most methods exploit trimaps as essential assistance to perceiving FG structure.
Hierarchical and Progressive Image Matting

Fig. 2. Pipeline of our HAttMatting++. The orange box (Pyramidal Features Distillation) indicates channel-wise attention to distill pyramidal information extracted from ASPP [5]. The blue box (Appearance Cues Filtration) represents spatial attention to filter appearance cues, which are extracted from block1 and block2 in the feature extraction module.

Traditional matting methods primarily resort to color features extracted from the input image to confine transition regions. According to the different ways of using color features, they can be roughly divided into two categories: sampling- and affinity-based methods. Sampling-based methods [17, 32, 36, 44] solve alpha mattes by representing each pixel inside transition regions with a pair of certain FG/BG pixels. According to the local smoothness assumption on the image statistics, most of these methods require that these sampling colors are close to a certain FG or BG. Affinity-based methods [1, 7, 18–20] utilize the affinities of neighboring pixels between certain labels and transition regions to perceive FG boundaries, which often suffer from high computational complexity. One of the most representative methods is closed-form matting [19], which is derived from the matting Laplacian and can produce an alpha matte by solving a sparse linear system of equations.

Both sampling and affinity methods primarily harness color similarity among pixels to predict alpha mattes, incapable of describing the advanced structure of FG. As a result, they usually produce obvious artifacts when the FG and BG share similar colors.

Deep learning matting. Compared with traditional methods that use low-level features, learning-based methods have refreshed all of the previous state-of-the-art records in image and video matting. Similar to other computer vision tasks, matting objects also process a general structure that can be represented by high-level semantic features. Shen et al. [37] proposed the first automatic matting system for portrait photos in an end-to-end fashion. Cho et al. [9] combined the results of Levin et al. [19] and Chen et al. [7] with normalized RGB images as input and learn an end-to-end deep network to predict better alpha mattes. Xu et al. [51] proposed DIM, which integrated RGB images with trimaps as conjunct input, utilizing advanced semantics to estimate alpha mattes. Lutz et al. [26] explored generative adversarial networks (GANs) [13] to generate alpha mattes with pleasing visual effects, which demonstrated the effectiveness of the discriminator model for pixel-wise matting perception. Tang et al. [39] proposed a hybrid sampling and learning-based approach to image matting. Cai et al. [3] and Hou and Liu [15] both established two branches to perceive alpha mattes, and these two branches mutually reinforced each other to refine the final...
results. Lu et al. [25] and Dai et al. [10] unified upsampling operators with the index function to improve the encoder-decoder network. Inspired by image inpainting [56], Li and Lu [21] utilize guided contextual attention to improve the transmission of opacity information. Yu et al. [55] explore patch-based crop and stitch to implement the context dependency in high-resolution matting. Although the trimap expansions in SIM [38] can extend the range of transition capture, they have a restriction on alpha categories. However, all of these matting methods rely on trimap as additional input to enhance their semantic distillation, whereas producing trimap is difficult for common users.

Some matting frameworks [6, 8, 57] resort to semantic segmentation to generate a pseudo trimap, which requires intermediate masks and sometimes can cause FG profiles or boundaries to be incomplete. Wei et al. [47], Yang et al. [53], and Yang et al. [52] sacrificed user interactions and extra feedback time to improve alpha mattes. Although the combination of the multi-scale features in the work of Aksoy et al. [2] can generate alpha mattes automatically, it has a very slow execution. Zhang et al. [58] investigated semantic segmentation variants for FG and BG weight map fusion to obtain alpha mattes. Although they can produce competent results without trimaps, failure cases occur when segmentation is inapplicable. Qiao et al. [30] exploit information integration to regress alpha mattes. Sengupta et al. [35] and Lin et al. [22] employ additional BG images to predict alpha mattes, which must share the environment as default. Trimap-free methods are similar in nature to saliency-based methods [40–42]. Compared to existing matting networks, we perform a profound excavation of the matting information in the input image, and inspired by the recent success of attention mechanism and contextual correlation [27, 28], we propose a hierarchical and progressive aggregation framework to integrate adaptive semantics and refined appearance cues. These improvements enable us to produce high-quality alpha mattes without trimaps.

3 METHODOLOGY

3.1 Overview

We can conclude from Equation (1) that the complete object FG should consist of two parts: (1) the main body indicating the FG category and profiles ($\alpha_z = 1$), and (2) the internal texture and boundary details located in the transition region ($\alpha_z \in (0, 1)$). The former can be obtained by advanced semantic information, whereas the latter usually comes from input images or low-level CNN features, termed appearance cues, and they can be combined to achieve alpha mattes. Although the low-level features and advanced semantic information can complement each other, their vanilla combination may produce some interference in alpha mattes. On the one hand, the low-level features usually contain texture details outside the FG. On the other hand, the highly abstracted advanced semantics typically contain global information and may be too sensitive to object categories, which means that they contribute unequally to alpha mattes that mostly consider the FG context. Consequently, we argue that the advanced semantics and appearance cues need essential processing before combination. First, natural image matting is supposed to handle different types of FG objects, which suggests that we should distill advanced semantics to attend to FG information, and appropriately suppress them to reduce their sensitivity to object class. Second, as shown in Figure 3, appearance cues involve unnecessary BG details, which need to be erased in alpha mattes.

Based on these observations, the core idea of our approach is to select matting-adapted semantic information and eliminate irrelevant BG texture in appearance cues, then aggregate them to predict alpha mattes. For this purpose, we adopt channel-wise attention to distill advanced semantics extracted from atrous spatial pyramid pooling (ASPP) [5], and perform spatial attention on different-level appearance cues to eliminate image texture details outside the FG simultaneously.
Our well-designed hierarchical and progressive attention mechanism can perceive the FG structure from adaptive semantics and refined boundaries, and their combination can achieve better results. Moreover, we design a hybrid loss to further guide the network training by combining MSE, SSIM, and adversarial loss \[13\], which respectively are responsible for pixel-wise precision, structure consistency, and visual quality. We also import sentry supervision after the pyramidal features distillation to improve adaptive semantics.

### 3.2 Network Architecture

#### 3.2.1 Overall Network Design

The pipeline of our proposed HAttMatting++ is unfolded in Figure 2. We harness ResNeXt [50] as the backbone in consideration of their powerful ability to extract high-level semantic information. To obtain a larger receptive field while retaining high-resolution feature maps, we remove the max pooling layers, set the stride of block1 and block4 as 1, and employ dilation convolutions in block4. The advanced feature maps from block4 are then fed to the ASPP [5] module for multi-level semantics capture.

There are five blocks in total in our feature extraction network, and we regard the first three layers (block0 to block2) as shallow layers and the rest (block3 and block4) as deep layers. The low-level features at shallow layers include more detailed information (see Figure 3), such as texture, boundary, and internal structure, but they also contain more noises outside the FG. By contrast, high-level features at deep layers can provide abstract semantic information (see Figure 2), which is beneficial to locating most parts of the FG and suppressing the noises.

Different from previous works [3, 58] that simply fuse the upsampled deep-layer features with the shallow-layer features by concatenation or element-wise summation operation, we adopt a more aggressive yet efficient form called attention-guided hierarchical and progressive structure aggregation. As shown in Figure 2, for the high-level features at deep layers, we apply a pyramidal features distillation module to select matting-adapted semantics and locate the FG regions with strong semantic responses. For the low-level features at shallow layers, we present an appearance cues filtration module to refine the boundaries of the FG and restrain the noises in the BG. Then we combine the distilled pyramidal features with the filtered progressive appearance cues in a top-down manner, and these hierarchical features can jointly contribute to the resolution of alpha mattes. In addition, we utilize the discriminator network referred to as PatchGAN [16, 60] to enhance the visual quality of alpha mattes.
3.2.2 Pyramidal Features Distillation. The extracted pyramidal features are devoted unequally to FG structure regression, hence we perform channel-wise attention on pyramidal features to distill adaptive semantic attributes. As in orange box shown in Figure 2, we present the pyramidal features after upsampling as $F_p^i \in \mathbb{R}^{C \times H/4 \times W/4}$, and here $H$ and $W$ are equal to the height and width of the input image. Similarly, the attended pyramidal features are $F_p^o \in \mathbb{R}^{C \times H/4 \times W/4}$. First, the upsampled $F_p^i \in \mathbb{R}^{C \times H/4 \times W/4}$ are fed to our pyramidal features distillation module, then we utilize global pooling to generalize the feature maps for aggregating spatial information. Subsequently, a shared multi-layer perceptron (MLP) that consists of two consecutive fully connected layers is employed to further distill semantic attributes. Then, through using a sigmoid operation, the range of the final channel-wise attention maps were mapped to $[0, 1]$. The preceding process can be described as follows:

$$F_p^o = \sigma \left( MLP \left( Pool \left( F_p^i, w_c^0 \right) \right), w_c^1 \right) \otimes F_p^i,$$

(2)

where $w_c$ refers to the parameters in our channel-wise attention block, $\sigma$ refers to the sigmoid operation, and $Pool$ represents the global pooling operation. The channel-wise attention can select pyramidal features adapted to image matting, and retain FG profiles and category attributes. The pyramidal features are learned from the deep ResNext block, which are highly abstract semantic information, and thus we need appearance cues to restore FG details in alpha mattes.

3.2.3 Appearance Cues Filtration. Image matting requests precise FG boundaries, whereas high-level pyramidal features are incapable of catching such texture details. Therefore, we bridge a skip connection between the appearance cues and the advanced pyramidal semantics, which can transport low-level structure information for alpha matte generation. The representative feature maps extracted from block1 are illustrated in the second row of Figure 3, and we take these features as our initial appearance cues. These appearance cues can depict a sophisticated image texture, compatible with the boundary accuracy required by alpha matte perception.

The proposed HAttMatting++ can leverage appearance cues to enhance FG boundaries in the results. Correspondingly, we import spatial attention to filter appearance cues located in the BG and emphasize the ones inside the FG simultaneously. Specifically, we use kernel size $1 \times 7$ and $7 \times 1$ to execute horizontal and vertical direction attention, respectively. The blue box in Figure 2 shows our spatial attention. The input of the black arrow indicates the guidance information from the previous stage, which can be depicted as $F_{gda} \in \mathbb{R}^{C \times h \times w}$ ($h$ and $w$ correspond to different resolutions on different-level appearance cues). They are further disposed of via two parallel convolutions with the above $1 \times 7$ and $7 \times 1$ kernels. Their combination is then concatenated and convolved by a standard convolution layer, followed by a sigmoid layer, producing our 2D spatial attention map. This spatial attention map was utilized to handle the appearance cues ($F_a^i \in \mathbb{R}^{C \times h \times w}$) by the operation of an element-wise product, to remove the textures and details that belong to the BG. The preceding process can be described as follows:

$$S_1 = f^{1 \times 7} \left( f^{7 \times 1} \left( f^{7 \times 8} \left( F_{gda}, w_S^0, w_S^1 \right), w_S^2 \right) \right),$$

(3)

$$S_2 = f^{1 \times 7} \left( f^{7 \times 1} \left( f^{7 \times 8} \left( F_{gda}, w_S^0, w_S^1 \right), w_S^2 \right) \right),$$

(4)

$$F_a^o = \sigma \left( f^{1 \times 1} \left( Cat \left( S_1, S_2 \right) \right), w_S^1 \right) \otimes F_a^i,$$

(5)

where $\sigma$ denotes the sigmoid function and $f^{n \times m}$ represents a convolution operation with the filter size of $n \times m$, and $w_S$ represents the parameters in our spatial attention. $F_a^o \in \mathbb{R}^{C \times h \times w}$ is the output of our appearance cues filtration module. Despite that the appearance cues involve sufficient
image texture, only the regions inside or surrounding the FG can contribute to alpha mattes. Therefore, we use distilled pyramidal features as the guidance, combined with low-level information at different levels to jointly improve alpha mattes.

Several research works [4, 48] combine spatial and channel-wise attention, and we have three main differences compared to them. First, we use the high-level features as spatial attention maps to filter low-level features, which can suppress BG appearance cues via attended FG semantics. Second, the channel-wise attention is only imported to distill the pyramidal features, and we use Max-Pool to capture the most representative features. Thus, the dominated semantics can be extracted and enhanced for the FG profiles. Third, the attention mechanism is employed in the multiple decoder phases, guiding the encoder features progressively.

3.2.4 **Top-Down Progressive Features Aggregation.** Since the feature maps of block0 are originally derived from the input images, which are highly crude for alpha perception, we ignore it and just consider the low-level features from block1 and block2 as our appearance cues (denoted as initial appearance cues and secondary appearance cues separately in the following). First, we perform channel-wise attention on the advanced pyramidal features extracted from ASPP [5] to select adaptive semantics. Then, we consider the distilled semantics as guidance and execute the first-stage spatial attention on secondary appearance cues (the \( h \) and \( w \) in Equation (5) are \( H/4 \) and \( W/4 \) separately). The filtered secondary appearance cues and the distilled semantics are aggregated as the subsequent guidance. Next, we carry out another spatial attention on initial appearance cues to further remove redundant BG information (here the \( h \) and \( w \) in Equation (5) are \( H/2 \) and \( W/2 \) separately). After this, we perform the final hierarchical features aggregation, and the alpha mattes can be generated by another upsampling operation. It is worth noting that even if the second spatial attention does not explicitly use the distilled pyramidal features as guidance, it has been implicitly included in the result of the first aggregation. In addition, we add sentry supervision after the pyramidal features distillation module to enhance matting-adapted semantics. It can also provide better guidance for the appearance cues filtration module and promote the convergence of the spatial attention during backward propagation.

The channel-wise and spatial attention can jointly optimize the alpha matte generation: one responsible for pyramidal features selection and the other responsible for appearance cues filtration. This well-designed top-down hierarchical attention mechanism can efficiently attend low-level features and advanced semantics, and their progressive aggregation can produce high-quality alpha mattes with fine-grained details.

### 3.3 Loss Function

Pixel regression related loss function (\( L_1 \) or MSE loss) is usually adopted as the object function for alpha matte prediction [3, 51] and can generate competent alpha mattes via pixel-wise supervision. However, such regression loss only measures the difference in absolute pixel space, without consideration of FG structure. Therefore, we introduce SSIM loss (\( L_{ssim} \)) to calculate structure similarity between the predicted alpha mattes and ground truth. SSIM [46] has demonstrated a striking ability to boost structure consistency in the predicted images [31, 43]. Apart from the aforementioned loss function, we add adversarial loss (\( L_{adv} \)) to promote the visual quality of the predicted alpha mattes. In the proposed HAttMatting++, we employ this hybrid loss function to guide the network training, achieving effective alpha matte optimization. Our loss function is defined as follows:

\[
L_{total} = \lambda_1 L_{adv} + \lambda_2 L_{MSE} + \lambda_3 L_{SSIM} + \lambda_4 L_{sentry}.
\]

\( L_{adv}, L_{MSE}, \) and \( L_{SSIM} \) can improve alpha mattes from visual quality, pixel-wise accuracy, and FG structure similarity separately, whereas \( L_{sentry} \) refers to the sentry supervision after the
pyramidal features distillation module. It was designed to improve the adaptive semantics and promote the convergence of the spatial attention. \(\lambda_1, \lambda_2, \lambda_3,\) and \(\lambda_4\) represent balance coefficients for loss function. \(L_{adv}\) is defined as follows:

\[
L_{adv} = E(I, A)[\log(D(I, A) + \log(1 - D(I, G(I)))]
\]

where \(I\) represents the input image and \(A\) is the predicted alpha matte. \(L_{MSE}\) is expressed as follows:

\[
L_{MSE} = \frac{1}{|\Omega|} \sum_{i} (\alpha^i_p - \alpha^i_g)^2, \quad \alpha^i_p, \alpha^i_g \in [0, 1],
\]

where \(\Omega\) represents pixels set and \(|\Omega|\) is the number of pixels (i.e., the size of the input image). \(\alpha^i_p\) and \(\alpha^i_g\) are the predicted and ground truth alpha values at pixel \(i\), respectively. \(L_{MSE}\) can ensure the pixel-wise accuracy of alpha matte estimation. We establish FG structure optimization via \(L_{SSIM}\) as follows:

\[
L_{SSIM} = 1 - \frac{(2\mu_p\mu_g + c_1)(2\sigma_{p,g} + c_2)}{(\mu^2_p + \mu^2_g + c_1)(\sigma^2_p + \sigma^2_g + c_2)}.
\]

Here \(\mu_p, \mu_g\) and \(\sigma_p, \sigma_g\) are the mean and standard deviations of \(\alpha^i_p\) and \(\alpha^i_g\). With \(L_{SSIM}\) as guidance, our method can further improve FG structure:

\[
L_{sentry} = \frac{1}{|\Omega_{-4}|} \sum_{i} (\alpha^i_p - \alpha^i_g)^2, \quad \alpha^i_p, \alpha^i_g \in [0, 1],
\]

where \(\Omega_{-4}\) represents a quarter of the full pixels set. \(L_{sentry}\) is partly motivated by the side output [14], and we take the same loss function as \(L_{MSE}\) to supervise the pyramidal distillation module. The only difference is that the resolution of \(L_{sentry}\) is decreased by a factor of 4 than the original resolution of \(L_{MSE}\). \(L_{sentry}\) can assist the spatial attention to restrict the noises out of the BG and further improve the pixel-wise accuracy of the final alpha matte.

### 3.4 Implementation Details

We implement HAttMatting++ using PyTorch. During the training process, for the data augmentation, we crop the image and GT pairs centered on random pixels in the transition regions indicated by trimaps, then they are randomly cropped to \(512 \times 512\) and \(640 \times 640\), and \(800 \times 800\). Next, they are resized to a resolution of \(512 \times 512\) and augmented by horizontally random flipping. To accelerate the training process and prevent overfitting, we use the pre-trained ResNeXt-101 network [50] as the feature extraction backbone, whereas the other layers are randomly initialized from a Gaussian distribution. For loss optimization, we use the stochastic gradient descent optimizer with the momentum of 0.9 and a weight of 0.007, adjusted by the “poly” policy [24] with the power of 0.9 for 20 epochs. The balance coefficients \(\lambda_1, \lambda_2, \lambda_3,\) and \(\lambda_4\) in Equation (6) are 0.05, 1, 0.1, and 0.3 during the first epoch and are revised as 0.05, 1, 0.025, and 0.1 for the subsequent 19 epochs. Our HAttMatting++ is trained on three joint Tesla V100 graphics cards with a mini-batch size of 4, and it takes about 24 hours for the network to finish 20 epochs.

### 4 EXPERIMENTS AND ANALYSES

In this section, we evaluate HAttMatting++ on two datasets: the public Adobe Composition-1K [51] and our Distinctions-646. We first compare HAttMatting++ with state-of-the-art methods both quantitatively and qualitatively. Then we perform ablation study for HAttMatting++ and the previous version HAttMatting [29] on both datasets to demonstrate the significance of several crucial components. Finally, we execute HAttMatting++ on real scenarios to generate alpha mattes.
Fig. 4. Some examples of our Distinctions-646 dataset. The first row presents FG images, the second row presents ground truth alpha mattes, and the third row presents the corresponding composited results on new BG. The FG images display irregularly due to truncated boundaries. Please zoom in to see the finer details and texture.

4.1 Datasets and Evaluation Metrics

Datasets. The first dataset is the public Adobe Composition-1K [51]. The training set consists of 431 FG objects with the corresponding ground truth alpha mattes. Each FG image is combined with 100 BG images from the MS COCO dataset [23] to composite the input images. For the test set, Composition-1K contains 50 FG images as well as the corresponding alpha mattes, and 1,000 BG images from the PASCAL VOC2012 dataset [11]. The training and test sets were synthesized through the algorithm provided by Xu et al. [51].

The second is our Distinctions-646 dataset. Adobe Composition-1K contains many consecutive video frames and cropped patches from the same image, and there are only 250 dissimilar FG objects in their training set. To improve the versatility and robustness of the matting during training, we construct our Distinctions-646 dataset composed of 646 distinct FG images. As shown in Figure 4, the dataset we proposed has a high degree of alpha precision and high freedom of FG categories and shapes. For training and test, we divide these FG examples into 596 and 50 and then produce 59,600 training images and 1,000 test examples according to the composition rules in the work of Xu et al. [51]. We have released the download right of our dataset.

Evaluation metrics. We evaluate the alpha mattes produced by our HAttMatting++ following the four common quantitative metrics: the sum of absolute difference (SAD), MSE, the gradient (Grad), and connectivity (Conn) proposed by Rhemann et al. [33]. A better image matting method can mostly produce high-quality alpha mattes, thus reducing the values of the preceding four metrics.

4.2 Comparison to the State of the Art

Evaluation on the Composition-1k test set. Here we compare HAttMatting++ with six traditional handcrafted algorithms: Shared Matting [12], Learning Based [59], Global Matting [32], Closed Form [19], KNN Matting [7], and Information-Flow [1], as well as twelve deep learning based methods: DCNN [9], DIM [51],
Fig. 5. Visual comparisons on the Composition-1k test set. The segments in SSS [2] are hand picked.

AlphaGAN [26], SSS [2], SampleNet [39], Context-Aware [15], IndexNet [25], MSIA [30], GCA [21], A^2U [10], Late Fusion [58], and BGMv2 [22]. The evaluation codes of the preceding methods are from original papers or our implementations using recommended parameters.

SSS [2], MSIA [30], Late Fusion [58], and our HAttMatting series can generate alpha mattes without trimap. For the other methods that rely on trimaps as input, we can generate transition regions along with the areas where alpha values are between 0 and 1, using random dilation in the range of [1, 25]. We harness full-resolution input images for fair contrast, and the visual results are illustrated in Figure 5. The quantitative comparisons are reported in Table 1, and the four metrics are all calculated on the whole image. In Table 1, the methods in gray (Late Fusion and our HAttMatting series) only take RGB images as input, whereas the others require trimap as assistance to guarantee the accuracy of alpha mattes. “Basic” means our baseline network, and the corresponding “Basic +” represents that we assemble different components on the baseline to generate alpha mattes. Here “HAttMatting” means our previous published conference paper. “HAttMatting + multi-scale” and “HAttMatting + Sentry” indicate that we add the secondary appearance cues and sentry supervision, respectively. HAttMatting++ denotes the current full model (with multi-scale appearance cues and sentry supervision).
Table 1. Quantitative Comparisons on the Composition-1k Testing Set

| Methods                  | SAD↓ | MSE↓ | Grad↓ | Conn↓ | Additional Input |
|--------------------------|------|------|-------|-------|-------------------|
| Shared Matting [12]      | 125.37 | 0.029 | 144.28 | 123.53 | Trimap            |
| Learning Based [59]      | 95.04  | 0.018 | 76.63  | 98.92  | Trimap            |
| Global Matting [32]      | 156.88 | 0.042 | 112.28 | 155.08 | Trimap            |
| Closed Form [19]         | 124.68 | 0.025 | 115.31 | 106.06 | Trimap            |
| KNN Matting [7]          | 126.24 | 0.025 | 117.17 | 131.05 | Trimap            |
| DCNN [9]                 | 115.82 | 0.023 | 107.36 | 111.23 | Trimap            |
| Information-Flow [1]     | 70.36  | 0.013 | 42.79  | 70.66  | Trimap            |
| DIM [51]                 | 48.87  | 0.008 | 31.04  | 50.36  | Trimap            |
| AlphaGAN [26]            | 90.94  | 0.018 | 93.92  | 95.29  | Trimap            |
| SampleNet [39]           | 48.03  | 0.008 | 35.19  | 56.55  | Trimap            |
| IndexNet [25]            | 44.52  | 0.005 | 29.88  | 42.37  | Trimap            |
| CA Matting [15]          | 38.73  | 0.004 | 26.13  | 35.89  | Trimap            |
| GCA Matting [15]         | 35.27  | 0.0034 | 19.72  | 31.93  | Trimap            |
| A2U [10]                 | 33.78  | 0.0037 | 18.04  | 31.00  | Trimap            |
| Late Fusion [58]         | 58.34  | 0.011 | 41.63  | 59.74  | No               |
| MSIA [30]                | 47.86  | 0.007 | 28.61  | 43.39  | No               |
| HAttMatting [29]         | 44.01  | 0.007 | 29.26  | 46.41  | No               |
| Basic                    | 126.31 | 0.025 | 111.35 | 118.71 | No               |
| Basic + SSIM             | 102.79 | 0.021 | 88.04  | 110.14 | No               |
| Basic + Low              | 89.39  | 0.016 | 56.67  | 90.03  | No               |
| Basic + CA               | 96.67  | 0.018 | 73.94  | 95.08  | No               |
| Basic + Low + CA         | 72.73  | 0.013 | 49.53  | 65.92  | No               |
| Basic + Low + SA         | 54.91  | 0.011 | 46.21  | 60.40  | No               |
| Basic + Low + CA + SA    | 49.67  | 0.009 | 41.11  | 53.76  | No               |
| HAttMatting + Multi-scale | 45.69  | 0.0065 | 28.61  | 45.83  | No               |
| HAttMatting + Sentry     | 43.46  | 0.0064 | 28.17  | 46.11  | No               |
| HAttMatting++            | 43.27  | 0.0059 | 27.91  | 44.09  | No               |
| BGMv2                    | 15.56  | 0.012 | 11.46  | 13.94  | BG images        |
| HAttMatting++_Human      | 18.40  | 0.011 | 8.59   | 16.13  | No               |

Note: “Human” means that the metrics are only calculated on the human images.

Compared to the conventional handcrafted methods, HAttMatting++ exhibits significant superiority over them, which can be clearly observed in Figure 5 and Table 1. As for the deep learning based approaches, we divide them into two categories: trimap-free methods and trimap-based ones. As shown in Table 1 and Figure 5, compared with the trimap-free methods of SSS [2], MSIA [30], and Late Fusion [58], our method can surpass them by a large margin both quantitatively and qualitatively. Since the results of SSS are separate semantic blocks, and the inference process is quite time consuming, we segment some sample images, then manually select and splice different semantic blocks to form the final alpha mattes. Similar results could be found in the work of Aksoy et al. [2].
Compared with trimap-based methods, our HAttMatting++ has more sophisticated details than DCNN [9] and DIM [51], and is better than SampleNet [39], since we employ a hierarchical and progressive attention mechanism to distill advanced semantics and different-level appearance cues, and their aggregation achieves relatively complete FG profiles and boundaries. Our HAttMatting++ is slightly inferior to Context-Aware [15] and IndexNet [25] on some metrics. The former establishes two branches and resorts to FG image supervision to predict alpha mattes, whereas the latter learns index functions to capture texture and boundary details. Both the cutting-edge GCA [21] and A²U [10] matting can achieve better alpha mattes by profound exploration of trimaps. Although they both generate high-quality alpha mattes, trimaps are strongly required during their training and inference phase, which restricts their effectiveness in practical applications. Our HAttMatting++ only need single RGB images as input, which is quite convenient for novice users. The comparative results with BGMv2 [22] are reported at the bottom of Table 1. The BGMv2 implementation is the same as the original paper (trained only on human images). Here we use the 220 human samples in the Composition-1k testing set rather than five random BG images. Independent of the BG images, we achieve better MSE and Grad metrics than BGMv2. The visual results are shown in Figure 7 (the first row).

Evaluation on Distinctions-646. For our Distinctions-646 dataset, we compare HAttMatting++ with eight recent matting methods, including Shared Matting [12], Learning Based [59], Global Matting [32], Closed Form [19], KNN Matting [7], DCNN [9], Information-Flow [1], and DIM [51], as well as BGMv2 [22]. We also use random dilation to generate high-quality trimaps [51], and relevant metrics are computed on the whole image.

The quantitative comparisons are displayed in Table 2, where the definition of “Basic,” “Basic +,” and “HAttMatting +” are the same as in Table 1. Our HAttMatting++ shows a clear advantage on all four metrics compared to all of the traditional methods, and is slightly better than DIM [51] on SAD and MSE metrics, while obviously better than in the Grad and Conn metrics, which indicates that our model can achieve improved visual quality. It is noted that only our method can generate alpha mattes without trimaps, and all of the other methods demand trimaps to confine the transition region, which effectively improves the performance of these methods. Figure 6 presents a visual comparison with the DIM [51] network. Here we enlarge the transition region to reduce
the accuracy of trimap, and the corresponding alpha mattes with DIM are shown in the fourth column. The deterioration in visual quality is evident with the transition region expanded, which can verify that DIM has a strong dependence on the quality of trimaps. The alpha mattes produced by HAttMatting++ exhibit sophisticated texture details, which mainly benefits from the aggregation of adaptive semantics and valid appearance cues in our model. The comparisons of human samples with BGMv2 are demonstrated in Table 2 and Figure 7 (the second row). We also employ the official BGMv2 weights that are trained on human images, and the proposed HAttMatting achieves better SAD and MSE metrics.

4.3 Ablation Study

The core idea of our HAttMatting++ is to extract adaptive pyramidal features and filter progressive appearance cues, then aggregate them to generate alpha mattes. To accomplish this goal, we employ channel-wise attention (CA) and spatial attention (SA) to re-weight pyramidal features and different-level appearance cues separately. We also introduce SSIM in our loss function to further improve the FG structure.

In this section, we first discuss the potential variation of the proposed network architecture, then analyze the properties of some important components in the conference paper [29] by combining them on the baseline network, and finally expound the presented multi-level information fusion and sentry supervision. All combinations of different components in the following are trained and tested on the Composition-1k dataset and our Distinctions-646 dataset. The correlated evaluation values are summarized in Tables 1 and 2.

4.3.1 Alternative Semantics Extraction Strategies. In the pipeline of our HAttMatting++, we extract high-level semantic features from the block4 of ResNeXt [50], then feed them to ASPP [5] module to capture pyramidal features. The original intention of such architecture is based on the consideration that FG objects always occupy most of the input image in matting, and there is no need to design a multi-scale framework to capture objects of various sizes.

Actually, we have tried other network architectures to obtain pyramidal features. For the fair comparison with different extraction strategies at the architecture level, we remove the sentry supervision and take the previous conference paper as the baseline. As shown in the left of Figure 8, we attempt to concatenate the feature maps from block3 and block4, and assume that the fusion between them can lead to more abundant high-level semantics for locating most of the FG. Theoretically speaking, we can give the block3 branch a weight map 0 to degenerate the architectures.
| Methods                        | SAD↓ | MSE↓ | Grad↓ | Conn↓ | Additional Input |
|-------------------------------|------|------|-------|-------|------------------|
| Shared Matting [12]           | 119.56 | 0.026 | 129.61 | 114.37 | Trimap           |
| Learning Based [59]           | 105.04 | 0.021 | 94.16  | 110.41 | Trimap           |
| Global Matting [32]           | 135.56 | 0.039 | 119.53 | 136.44 | Trimap           |
| Closed Form [19]              | 105.73 | 0.023 | 91.76  | 114.55 | Trimap           |
| KNN Matting [7]               | 116.68 | 0.025 | 103.15 | 121.45 | Trimap           |
| DCNN [9]                      | 103.81 | 0.020 | 82.45  | 99.96  | Trimap           |
| Information-Flow [1]          | 78.89  | 0.016 | 58.72  | 80.47  | Trimap           |
| DIM [51]                      | 47.56  | 0.009 | 43.29  | 55.90  | Trimap           |
| Basic                         | 129.94 | 0.028 | 124.57 | 120.22 | No               |
| Basic + SSIM                  | 121.79 | 0.025 | 110.21 | 117.41 | No               |
| Basic + Low                   | 98.88  | 0.020 | 84.11  | 92.88  | No               |
| Basic + CA                    | 104.23 | 0.022 | 90.87  | 101.9  | No               |
| Basic + Low + CA              | 85.57  | 0.015 | 79.16  | 88.38  | No               |
| Basic + Low + SA              | 78.14  | 0.014 | 60.87  | 71.90  | No               |
| Basic + Low + CA + SA         | 57.31  | 0.011 | 52.14  | 63.02  | No               |
| HAttMatting [29]              | 48.98  | 0.0094 | 41.57  | 49.93  | No               |
| HAttMatting + Multi-scale     | 48.32  | 0.0091 | 41.15  | 48.01  | No               |
| HAttMatting + Sentry          | 47.53  | 0.0090 | 40.91  | 46.34  | No               |
| HAttMatting++                 | **47.38** | **0.0088** | **40.09** | **45.60** | No               |
| BGMv2                         | 11.53  | 0.010 | 9.69   | 9.70   | BG images        |
| HAttMatting++_Human           | 11.46  | 0.010 | 8.91   | 9.88   | No               |

In Figure 8, into our pipeline. However, it is difficult to achieve the same accuracy as our pipeline by combining the semantic features of block3 in the training process, possibly because the coarse semantics from block3 may result in a diversion of the pyramidal features. In addition, another form of semantic fusion we tried is shown in the right of Figure 8. The channel-wise attention and spatial attention were performed on the features extracted from ASPP [5] and block3, then their combination associated with the low-level features extracted from block1 was sent to execute further spatial attention. However, the feature maps from block3 are most highly abstract and only contain limited appearance cues for exploration, which may on the contrary disturb the semantic extraction of the backbone.

For low-level features, we also attempt to utilize the input image as appearance cues, like previous matting works [45, 51, 58], but the complex color distribution will obscure texture details and disturb the refinement of FG boundaries.

The quantitative comparisons are shown in Table 3. The implementation details of all models are the same as HAttMatting++, and all results are evaluated on the Composition-1k test set. "Input Image" means that we only extract spatial cues from the input image, and “Strategy-1” and “Strategy-2” refer to the alternative pyramidal features capture strategies on the left and right of Figure 8, respectively. The alpha mattes produced by “HAttMatting + MS” are better than the other features extraction strategies. In our analysis, we argue that the FG object occupies most of the
Table 3. Comparisons with Different Semantic Features

| Extraction Strategies | Methods               | SAD↓  | MSE↓  | Grad↓  | Conn↓ |
|-----------------------|-----------------------|-------|-------|--------|-------|
| Input Image           |                       | 115.23| 0.0157| 51.26  | 106.06|
| Strategy-1            |                       | 77.29 | 0.0098| 44.93  | 78.63 |
| Strategy-2            |                       | 64.37 | 0.0072| 32.03  | 56.99 |
| HAttMatting + MS      |                       | 45.69 | 0.0065| 28.61  | 45.83 |

Fig. 8. Some potential pyramidal features extraction architectures.

4.3.2 Primary Combination. Here we analyze the effectiveness of different components in the previous version [29]. We take the input image from the Adobe Composition-1k dataset as a visual instance, and the corresponding alpha mattes produced by some following models are shown in the first and second rows of Figure 9. The “Low” here denotes the appearance cues from block1.

Basic. This is our baseline network, which only uses original pyramidal features to generate alpha mattes, and is optimized by \( L_{MSE} \) and \( L_{adv} \).

Basic + SSIM. \( L_{SSIM} \) is involved in our loss function. It can be obviously seen from Tables 1 and 2 that even if only \( L_{SSIM} \) is added to optimize the baseline network, the results can surpass some traditional methods, such as Shared Matting [12], Global Matting [32], Closed Form [19], and KNN Matting [7]. Meanwhile, since DCNN [9] utilizes the results of closed-form matting [19] and KNN matting [7] to predict alpha mattes by relatively mediocre networks, our method is deservedly better than DCNN [9].

Basic + Low. Low-level appearance cues are directly aggregated with pyramidal features, which can furnish sophisticated texture and details for alpha mattes. Compared to the pyramidal features that own most of the advanced semantics for locating the profiles of objects, the low-level features contain most of the texture and fine-grained structure information that can contribute to the boundary of FG. After adding low-dimensional information, all four metrics improve a lot, which is also reasonable.

Basic + CA. On the basis of baseline, we perform channel-wise attention to distill pyramidal features. CA can effectively suppress irrelevantly advanced semantics and reduce the sensitivity of the trained model to object classes, which means that the network can handle diverse kinds of FG objects and the model versatility is enhanced. We observe that the metric of MSE dropped 0.007 (from 0.025 to 0.018) by adding only one CA module. This can also corroborate the importance of CA for locating the profile of FG and restraining the response of unrelated semantics.

Basic + Low + CA. This combination integrates the advantages of the preceding two modules to promote performance (Figure 9). Both components can boost the performance above the baseline...
Fig. 9. The visual comparison of different components. “HAttMatting + MS” and “HAttMatting + Sen” mean that we embed secondary appearance cues and sentry supervision in the previous version, respectively, whereas HAttMatting++ represents our current full model. Each component has a significant improvement for alpha mattes. Please zoom in to see the finer difference.

network, and their combination can effectively aggregate matting-adapted semantics and refined boundaries as expected.

Basic + Low + SA. Our modified SA can eliminate the BG texture in appearance cues, improving the subsequent aggregation process. It is noting that the spatial attention map must exert on the low-level features, and thus the “Basic + SA” comparison is unavailable in our ablation experiment. According to our quantitative and qualitative report, the Low and SA module can bring a huge enhancement to the final alpha mattes. In addition, we can observe from Figure 9 that although spatial attention can filter redundant BG information, they are inapplicable for attending matting-adapted semantics without the CA module (the black dress in the BG is also projected into the alpha mattes).

Basic + Low + CA + SA. We assemble CA, Low, and SA to achieve competent alpha mattes without SSIM.

More detailed quantitative comparisons are shown in Tables 1 and 2. It can be clearly seen that each component can significantly improve our results. The low-level structure information and advanced semantics are essentially complementary to each other in terms of image matting. CA can furnish FG profiles by distilling pyramidal features, whereas SA can exhibit fine-grained internal texture and boundary details by filtrating initial appearance cues, and their aggregation can generate high-quality alpha mattes (see Figure 9).

4.3.3 Expanded Combination. We conduct this additional ablation study to verify the effectiveness of two newly extended components on the basis of the previous conference version. The first is our progressive appearance cues extraction and aggregation, and the second is the sentry supervision. Similarly, these two models are trained and tested on both the Adobe Composition-1K dataset and our Distinctions-646 dataset.
HAttMatting + progressive structure aggregation. Compared to the previous conference paper that only fuses the attended pyramidal features from block4 and appearance cues from block1, we import another information flow from block2 as the secondary appearance cues to execute structure aggregation, under the guidance of our attention mechanism progressively. We argue that this design can further integrate middle-level information (the hands, leaves, etc.) to achieve multi-scale features aggregation. As shown in Figure 9 (this item is denoted as “HAttMatting + MS”), the stripes of the sieve are further distilled (please zoom in to see the yellow box). Furthermore, the three metrics except SAD improved slightly, which can also confirm the validity of secondary appearance cues for complementing necessary FG texture and details.

HAttMatting + sentry supervision. We downsample the ground truth alpha mattes by a factor of 4 and embed it at the back of pyramidal features distillation module as our sentry supervision. As shown in Table 1, there has been a slight decline in all four metrics after we add sentry supervision. We attribute this improvement to the more pure extraction of advanced semantics. First, the sentry supervision can guide the channel-wise attention to further capture the matting-adapted semantics, and prevent the misclassification of pixels inside the FG. This modification can promote the constraint of FG and BG regions, especially for our trimap-free motivation. Second, the resolution of the input images in matting are usually quite high, and such middle-size sentry supervision can shield some unrelated small objects and enhance the main semantics of the FG subject. Third, according to Section 3.2.3, it can produce better guidance to assist the spatial attention to restrict unnecessary BG noises. In addition, the back propagation of the intermediate loss can accelerate the convergence of the whole model while preventing the gradient from disappearing.

Consequently, the proposed model contains all components (i.e., CA, SA, SSIM loss, sentry supervision loss, and GAN loss) and can achieve the best performance, which demonstrates that all components listed previously are necessary for the proposed method to generate better alpha mattes.

4.4 Results on Real-World Images

Figure 10 shows our matting results on real-world images. All evaluation models are trained on the Composition-1k dataset. We dilate and erode our alpha mattes to generate high-quality trimaps (the second column), which are intractable for labeling from scratch. Based on these trimaps, we present the results of GCA [21] and A2U [10] for visual comparison. HAttMatting++ can achieve high-quality alpha mattes without any external input or user interaction. The dense branches and the fine hairs can be predicted with high precision, which is benefited from our progressive appearance cues filtration and aggregation. By contrast, GCA and A2U ignore the stalk (the fourth row) even with high-quality trimaps.

4.5 Failure Cases and Limitations

There are some limitations to our method, and two failure cases are illustrated in Figure 11. The reflection, emptiness, blur, and so forth of natural images can greatly influence our methods (the bottom of the cup is a struggle, and the branches of the dandelion are poorly detected). The reflection and emptiness can provide pseudo information for our hierarchical and progressive features extraction, and result in misleading guidance for top-down aggregation. In addition, our HAttMatting++ also has other limitations. First, the data generation and augmentation rules we refer to on the composite dataset can bring more artifacts and inconsistencies, which will expand the disparity between the natural images and the composite ones (other matting networks, e.g., [21, 39, 51], may be also more or less affected in the same way). Second, our model can only handle simple scenes (similar to most existing matting methods [3, 15, 58], and it will be better with trimaps as assistance when predicting multiple FG objects simultaneously.
52:20

Fig. 10. Results on real-world images.

Fig. 11. Some failure cases on real-world examples.

5 CONCLUSION AND FUTURE WORK
In this article, we propose HAttMatting++, which can predict high-quality alpha mattes from single RGB images. HAttMatting++ employs channel-wise attention to extract matting-adapted semantics and performs spatial attention at different levels to filtrate progressive appearance cues. Extensive experiments demonstrate that our hierarchical and progressive structure aggregation
can effectively distill high- and low-level features from the input images, and achieve high-quality alpha mattes without external trimaps.

In the future, we will explore how to improve the adaptability of the network to further increase the robustness of the model on natural images. The domain adaption [34] and GAN [49] may solve the problem of reducing the gap between synthetic data and natural images. We will also explore how to extract multi-target alpha mattes in a fashion of a trimap-free method.

REFERENCES
[1] Yagiz Aksoy, Tunc Ozan Aydin, and Marc Pollefeys. 2017. Designing effective inter-pixel information flow for natural image matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’17). 228–236.
[2] Yaşğiz Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. 2018. Semantic soft segmentation. ACM Transactions on Graphics 37, 4 (2018), Article 72.
[3] Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. 2019. Disentangled image matting. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV’19). 8818–8827.
[4] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’17). 6298–6306.
[5] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2018. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 4 (2018), 834–848.
[6] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. 2018. Semantic human matting. In Proceedings of the ACM International Conference on Multimedia (MM’18). 618–626.
[7] Qifeng Chen, Dingzeyu Li, and Chi Keung Tang. 2013. KNN matting. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 9 (2013), 2175–2188.
[8] D. Cho, S. Kim, Y. W. Tai, and I. S. Kweon. 2016. Automatic trimap generation and consistent matting for light-field images. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 8 (2016), 1504–1517.
[9] Donghyeon Cho, Yu-Wing Tai, and In So Kweon. 2019. Deep convolutional neural network for natural image matting using initial alpha mattes. IEEE Transactions on Image Processing 28, 3 (2019), 1054–1067.
[10] Yutong Dai, Hao Lu, and Chunhua Shen. 2021. Learning affinity-aware upsampling for deep image matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’21). 6841–6850.
[11] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2010. The PASCAL Visual Object Classes (VOC) challenge. International Journal of Computer Vision 88, 2 (2010), 303–338.
[12] Eduardo S. L. Gastal and Manuel M. Oliveira. 2010. Shared sampling for real-time alpha matting. Computer Graphics Forum 29, 2 (2010), 575–584.
[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Xu Bing, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS’14). 2672–2680.
[14] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H. S. Torr. 2017. Deeply supervised salient object detection with short connections. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’17). 3203–3212.
[15] QiQi Hou and Feng Liu. 2019. Context-aware image matting for simultaneous foreground and alpha estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV’19). 4129–4138.
[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’17). 5967–5976.
[17] L. Karacan, A. Erdem, and E. Erdem. 2015. Image matting with KL-divergence based sparse sampling. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV’15). 424–432.
[18] P. Lee and Ying Wu. 2011. Nonlocal matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’11). 2193–2200.
[19] Anat Levin, Dani Lischinski, and Yair Weiss. 2007. A closed-form solution to natural image matting. IEEE Transactions on Pattern Analysis and Machine Intelligence 30, 2 (2007), 228–242.
[20] Anat Levin, Alex Rav-Acha, and Dani Lischinski. 2008. Spectral matting. IEEE Transactions on Pattern Analysis and Machine Intelligence 30, 10 (2008), 1699–1712.
[21] Yaoyi Li and Hongtao Lu. 2020. Natural image matting via guided contextual attention. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI’20). 11450–11457.
[22] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L. Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2021. Real-time high-resolution background matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’21). 8762–8771.
[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV’14). 740–755.
[24] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. 2015. ParseNet: Looking wider to see better. arXiv preprint arXiv:1506.04579 (2015).
[25] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. 2019. Indices matter: Learning to index for deep image matting. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV’19). 3265–3274.
[26] Sebastian Lutz, Konstantinos Amplianitis, and Aljoscha Smolic. 2018. AlphaGAN: Generative adversarial networks for natural image matting. In Proceedings of the British Machine Vision Conference (BMVC’18). 259.
[27] Haiyang Mei, Yuanyuan Liu, Ziqi Wei, Dongsheng Zhou, Xiaopeng Xiao, and Xin Yang. 2021. Exploring dense context for salient object detection. IEEE Transactions on Circuits and Systems for Video Technology 32, 3 (2021), 1378–1389.
[28] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson W. H. Lau. 2020. Don’t hit me! Glass detection in real-world scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’20).
[29] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. 2020. Attention-guided hierarchical structure aggregation for image matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’20).
[30] Yu Qiao, Yuhao Liu, Qiang Zhu, Xin Yang, Yuxin Wang, Qiang Zhang, and Xiaopeng Wei. 2020. Multi-scale information assembly for image matting. Computer Graphics Forum 39 (2020), 565–574.
[31] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. 2019. BASNet: Boundary-aware salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’19). 7471–7481.
[32] C. Rhemann and C. Rother. 2011. A global sampling method for alpha matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’11). 2049–2056.
[33] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. 2009. A perceptually motivated online benchmark for image matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’09). 1826–1833.
[34] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. 2018. Learning from synthetic data: Addressing domain shift for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’18). 3752–3761.
[35] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2020. Background matting: The world is your green screen. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’20). 2288–2297.
[36] Ehsan Shahrian, Deepu Rajan, Brian Price, and Scott Cohen. 2013. Improving image matting using comprehensive sampling sets. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’13). 636–643.
[37] Xiaoyong Shen, Xin Tao, Hongyan Gao, Chao Zhou, and Jiaya Jia. 2016. Deep automatic portrait matting. In Proceedings of the European Conference on Computer Vision (ECCV’16). 92–107.
[38] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. 2021. Semantic image matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’21). 11120–11129.
[39] Jingwei Tang, Yagiz Aksoy, Cengiz Oztireli, Markus Gross, and Tunc Ozan Aydin. 2019. Learning-based sampling for natural image matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’19). 3050–3058.
[40] Xin Tian, Ke Xu, Xin Yang, Lin Du, Baocai Yin, and Rynson W. H. Lau. 2022. Bi-directional object-context prioritization learning for saliency ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’22).
[41] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson W. H. Lau. 2020. Weakly-supervised salient instance detection. In Proceedings of the British Machine Vision Conference (BMVC’20).
[42] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson W. H. Lau. 2021. Learning to detect instance-level salient objects using complementary image labels. International Journal of Computer Vision 130 (2021), 729–746.
Hierarchical and Progressive Image Matting

52:23

[43] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. 2018. CRNN: Multi-scale guided concurrent reflection removal network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’18). 4777–4785.

[44] Jue Wang and Michael F. Cohen. 2007. Optimized color sampling for robust matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’07). 1–8.

[45] Yu Wang, Yi Niu, Peiyong Duan, Jianwei Lin, and Yuanjie Zheng. 2018. Deep propagation based image matting. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI’18). 999–1006.

[46] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13, 4 (2004), 600–612.

[47] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Hanqing Zhao, Weiming Zhang, and Nenghai Yu. 2021. Improved image matting via real-time user clicks and uncertainty estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’21). 15374–15383.

[48] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV’18). 3–19.

[49] Yuanbo Xiangli, Yubin Deng, Bo Dai, Chen Change Loy, and Dahua Lin. 2020. Real or not real, that is the question. In Proceedings of the International Conference on Learning Representations (ICLR’20).

[50] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’17). 5987–5995.

[51] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. 2017. Deep image matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’17). 311–320.

[52] Xin Yang, Yu Qiao, Shaozhe Chen, Shengfeng He, Baocai Yin, Qiang Zhang, Xiaopeng Wei, and Rynson W. H. Lau. 2020. Smart scribbles for image matting. ACM Transactions on Multimedia Computing Communications and Applications 16, 4 (2020), Article 121, 21 pages.

[53] Xin Yang, Ke Xu, Shaozhe Chen, Shengfeng He, Baocai Yin Yin, and Rynson Lau. 2018. Active matting. In Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS’18). 4590–4600.

[54] Guan Yu, Wei Chen, Xiao Liang, Zhiang Ding, and Qunsheng Peng. 2006. Easy matting—A stroke based approach for continuous image matting. Computer Graphics Forum 25, 3 (2006), 567–576.

[55] Haichao Yu, Ning Xu, Zilong Huang, Yuqian Zhou, and Humphrey Shi. 2021. High-resolution deep image matting. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI’21). 3217–3224.

[56] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2018. Generative image inpainting with contextual attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’18). 5505–5514.

[57] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. 2021. Mask guided matting via progressive refinement network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’21). 1154–1163.

[58] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weizhu Xu. 2019. A late fusion CNN for digital matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’19). 7461–7470.

[59] Yuanjie Zheng and Chandra Kambhamettu. 2009. Learning based digital matting. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV’09). 889–896.

[60] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV’17). 2242–2251.

Received 29 August 2021; revised 11 May 2022; accepted 23 May 2022.

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 19, No. 2, Article 52. Publication date: February 2023.