Attention-Guided Hierarchical Structure Aggregation for Image Matting

Yu Qiao[1,*], Yuhao Liu[1,*],Xin Yang[1,4,†],Dongsheng Zhou[2],Mingliang Xu[3],Qiang Zhang[2],Xiaopeng Wei[1,†]
[1] Dalian University of Technology, [2] Dalian University, [3] Zhengzhou University
[4] Beijing Technology and Business University
{coachqiao2018,yuhaoLiu7456}@gmail.com,{xinyang,zhangq,xpwei}@dlut.edu.cn,donyson@126.com
iexumingliang@zzu.edu.cn

Abstract

Existing deep learning based matting algorithms primarily resort to high-level semantic features to improve the overall structure of alpha mattes. However, we argue that advanced semantics extracted from CNNs contribute unequally for alpha perception and we are supposed to reconcile advanced semantic information with low-level appearance cues to refine the foreground details. In this paper, we propose an end-to-end Hierarchical Attention Matting Network (HAttMatting), which can predict the better structure of alpha mattes from single RGB images without additional input. Specifically, we employ spatial and channel-wise attention to integrate appearance cues and pyramidal features in a novel fashion. This blended attention mechanism can perceive alpha mattes from refined boundaries and adaptive semantics. We also introduce a hybrid loss function fusing Structural SIMilarity (SSIM), Mean Square Error (MSE) and Adversarial loss to guide the network to further improve the overall foreground structure. Besides, we construct a large-scale image matting dataset comprised of $59,600$ *training images and* 1000 *test images (total* 646 *distinct foreground alpha mattes), which can further improve the robustness of our hierarchical structure aggregation model. Extensive experiments demonstrate that the proposed HAttMatting can capture sophisticated foreground structure and achieve state-of-the-art performance with single RGB images as input.*

1. Introduction

Image matting refers to precisely estimate the foreground opacity from an input image. This problem as well as 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:

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

where z denotes the pixel position in the input image I. α_z, F_z and B_z refer to the alpha estimation, foreground (FG) and background (BG) at pixel z separately. The problem is highly ill-posed, for each pixel in a given RGB image, 7 values need to be solved but only 3 values are known.

The digital matting is a pixel-wise FG regression essentially, and we hold that the structure of FG resides two aspects: adaptive semantics and refined boundaries, corresponding to $\alpha_z = 1$ and $\alpha_z \in (0, 1)$ in Eq. 1. Existing matting methods usually solve Eq. 1 by introducing user-provided trimaps as assistant input. The trimap is composed of black, gray and white, representing BG, transition region and absolute FG respectively. The transition region indicates FG boundaries, combined with FG to jointly guide matting algorithms. Given an RGB image and the corresponding trimap, traditional matting methods explore 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 FG and BG have indistinguishable colors.

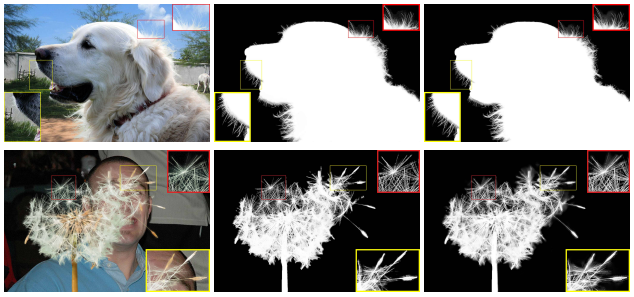


Figure 1: The alpha mattes produced by our *HAttMatting* on the Composition-1k test set [37].

*Joint first authors. †Joint corresponding authors, and they led this project. Project page: https://wukaoliu.github.io/HAttMatting/.

Deep Image Matting (*DIM*) [37] formally imports deep learning into matting, and they argue that matting objects share a common structure which can be represented by high-level features. It is noting that *DIM* involves *RGB* images in the refinement stage to combine advanced semantics with appearance cues. Advanced semantics indicate *FG* category and profiles, while appearance cues reveal texture and boundary details. Subsequent matting networks [3, 15, 23, 34] mostly design complicated architectures for advanced semantics extraction, and fuse appearance cues from input images or low-level *CNN* features.

However, their appearance cues and advanced semantics are all dependent on trimaps as auxiliary and expensive input. A well-defined trimap involves fussy manual labeling efforts and time consumption, which is difficult for novice users in practical applications. Some matting works [5, 7] rely on segmentation to generate trimaps, which partly depress the precision of alpha mattes. The Late Fusion [40] blends *FG* and *BG* weight map from segmentation network with initial *CNN* features to predict alpha mattes with single *RGB* images as input. However, when semantic segmentation encounters difficulties, the late fusion will compromise. The above methods directly feed advanced semantics and appearance cues to optimization or fusion stage, while we hold that they require proper filtration before combination. On one hand, natural image matting is a regression problem substantially and not entirely dependent on image semantics, which means semantic properties extracted by deep network contribute unequally for *FG* structure. On the other hand, as illustrated in Fig. 3, while appearance cues retain sophisticated image texture, they also contain the details outside *FG*. However, existing matting networks neglect the profound excavation and distillation of such hierarchical features.

This paper explores advanced semantics and appearance cues synthetically, and contributes an end-to-end Hierarchical Attention Matting Network (*HAttMatting*) enabling such hierarchical structure aggregation. Advanced semantics can provide *FG* category and profiles, while 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 appearance cues to filtrate image texture details, and finally aggregate them to predict alpha mattes. Moreover, a hybrid loss composed of Mean Square Error (*MSE*), Structural SIMilarity (*SSIM*) [35] and Adversarial Loss [13] is exploited to optimize the whole network training. Extensive experiments show that our attention-guided hierarchical structure aggregation can perceive high-quality alpha mattes with only *RGB* images as input.

The main contributions of this paper are:

- We present an end-to-end Hierarchical Attention Matting Network (*HAttMatting*), which can achieve high-quality alpha mattes without any additional input. The *HAttMatting* is very convenient for novice users and can be effectively applied to different kinds of objects.
- We design a hierarchical attention mechanism which can aggregate appearance cues and advanced pyramidal features to produce fine-grained boundaries and adaptive semantics.
- We resort to a hybrid loss consist of Mean Square Error (*MSE*), Structural SIMilarity (*SSIM*) and Adversarial Loss [13] to improve alpha perception, providing efficient guidance for our *HAttMatting* training.
- We create a large-scale matting dataset with $59,600$ training images and 1000 test images, total 646 distinct foreground alpha mattes. To the best of our knowledge, this is the biggest matting dataset with diverse foreground objects, which can further improve the robustness of our *HAttMatting*.

2. Related Work

Deep learning brings a huge evolution for natural image matting with the highly abstract 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: trimaps or scribbles. The trimap is composed of *FG*, *BG* and transition region to partition the input *RGB* image, while scribbles indicate these three labels by several user-specified scribbles. The transition region suggests *FG* boundaries, which is the key point for image matting. Although scribbles approaches [19, 20, 32, 39] are convenient for novice users, they significantly deteriorate alpha mattes because there is insufficient information can be referenced. Therefore, a majority of methods harness trimaps as essential assistance to perceive *FG* structure.

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 divided into two categories: sampling-based and affinity-based methods. Sampling-based methods [9, 11, 17, 26, 28, 33] solve alpha mattes by representing each pixel inside transition regions with a pair of certain *FG*/*BG* pixels. Affinity-based methods [1, 6, 14, 18, 19, 20, 29] perceive *FG* boundaries via the affinities of neighbouring pixels between certain labels and transition region. Both sampling and affinity methods primarily leverage color features to predict alpha mattes, incapable of describing the advanced structure of *FG*. When *FG* and *BG* share similar colors, traditional approaches usually produce obvious artifacts.

Deep-learning matting. Similar to other computer vision tasks, matting objects also possess a general struc-

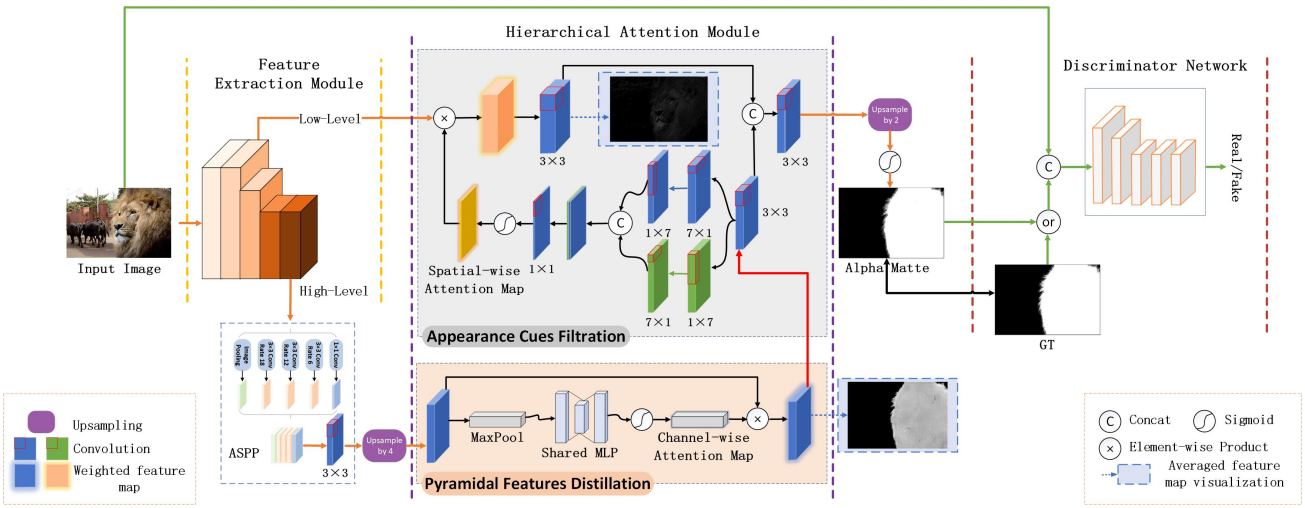


Figure 2: Pipeline of our *HAttMatting*. The orange box (Pyramidal Features Distillation) indicates channel-wise attention to distill pyramidal information extracted from *ASPP* [4]. The gray box (Appearance Cues Filtration) represents spatial attention to filter appearance cues, which are extracted from block1 in the feature extraction module.

ture that can be represented by high-level semantic features. Cho *et al.* [8] concatenated results from [19] and [6] with input image, and employed this 5-channels input to predict alpha mattes. Xu *et al.* [37] proposed deep image matting (*DIM*), which integrated *RGB* images with trimaps as conjunct input, utilizing advanced semantics to estimate alpha mattes. Tang *et al.* [30] proposed a hybrid sampling- and learning-based approach to matting. Cai *et al.* [3] and Hou *et al.* [15] both established two branches to perceive alpha mattes, and these two branches mutually reinforced each other to refine the final results. Hao *et al.* [23] unified upsampling operators with the index function to improve encoder-decoder network. However, all these matting networks rely on trimap to enhance their semantic distillation, while producing trimap is difficult for common users. Some matting frameworks [5, 7] leverage segmentation to generate trimaps, which usually causes *FG* profiles or boundaries incomplete. Yang *et al.* [38] used LSTM and reinforcement learning to produce competent trimap, requiring simple user interaction and extra feedback time. While the multi-scale features combination in [2] can generate alpha mattes automatically, it has a very slow execution. Zhang *et al.* [40] investigated semantic segmentation variant for *FG* and *BG* weight map fusion to obtain alpha mattes. Although they implement matting without trimaps, failure cases occur when segmentation is inapplicable.

3. Methodology

3.1. Overview

We can conclude from Eq. 1 that the complete object *FG* should consist of two parts: 1) the main body indicating *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 suggested by advanced semantics, while the latter usually comes from input images or low-level CNN features, termed as appearance cues, and their combination can achieve alpha mattes. In this paper, we argue that advanced semantics and appearance cues need proper 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 *FG* information, and appropriately suppress them to reduce their sensitivity to object classes. Second, as shown in Fig. 3, appearance cues involve unnecessary *BG* details, which need to be erased in alpha mattes.

Based on the above analysis, the core idea of our approach is to select matting-adapted semantic information and eliminate redundant *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*) [4], and perform spatial attention on appearance cues to eliminate image texture details outside *FG* simultaneously. Our well-designed hierarchical attention mechanism can perceive *FG* structure from adaptive semantics and refined boundaries, and their aggregation can achieve better alpha mattes. Moreover, we design a hybrid loss to guide network training by combining Mean Square Error (*MSE*), Structural SIMilarity (*SSIM*) and Adversarial loss [13], which are respectively responsible for pixel-wise precision, structure consistency and visual quality.

3.2. Network Architecture

Overall network design. The pipeline of our proposed *HAttMatting* is unfolded in Fig. 2. We harness

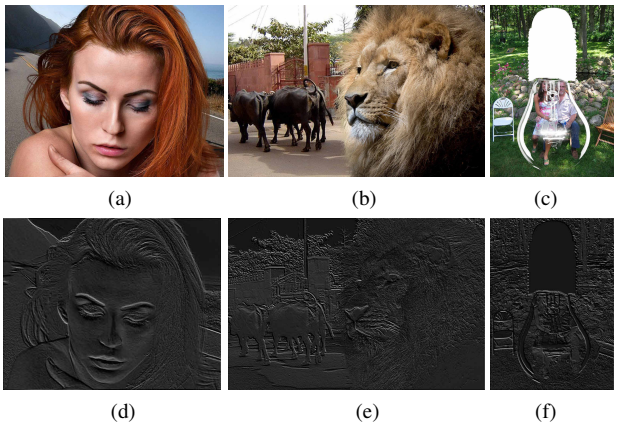


Figure 3: The input images and corresponding appearance cues extracted from the ResNeXt block1. Here we select one of the 256 channels for better visual presentation.

ResNeXt [36] as the backbone network in consideration of their powerful ability to extract high-level semantic information. A series of parameter adjustments are performed on the backbone to obtain a larger receptive field. The advanced feature maps from block4 are then fed to *ASPP* [4] module for multi-scale semantics capture. Correspondingly, we average the feature maps of block1 as appearance cues in our method (Fig. 3). The *HAttMatting* employs channel-wise attention to distill pyramidal features, and performs spatial attention on appearance cues to suppress redundant *BG* details. Besides, we utilize the discriminator network refer to PatchGAN [16, 42] to enhance the visual quality of alpha mattes.

Pyramidal features distillation. The extracted pyramidal features devote unequally to *FG* structure regression, hence we perform channel-wise attention on pyramidal features to distill adaptive semantic attributes. As the orange box is shown in Fig. 2, we upsample pyramidal features with factor 4, and then utilize global pooling to generalize the feature maps. Then a shared MLP is employed to distill semantic attributes. We use a sigmoid layer to compute channel-wise attention map, and multiply it times upsampled pyramidal features to achieve semantics distillation. 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 deep ResNext block, which are highly abstract semantic information, thus we need appearance cues to generate details in alpha mattes.

Appearance cues filtration. Image matting requests precise *FG* boundaries, while high-level pyramidal features are incapable of providing such texture details. Therefore, we bridge a skip connection between ResNeXt block1 and upsampling (Fig. 2) operation, which can transport appearance cues for alpha matte generation. The block1 can capture image texture and details from the input image, sharing the same spatial resolution as the first upsampling. The feature maps extracted from block1 are illustrated in the second row of Fig. 3, we take these low-level features as our appearance cues. These appearance cues can depict 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. Despite the appearance cues exhibit sufficient image texture, only the regions inside or surrounding *FG* can contribute to alpha mattes. Therefore, we import spatial attention to filter appearance cues located in *BG* and emphasize the ones inside *FG* simultaneously. Specifically, we use kernel size $1 * 7$ and $7 * 1$ to execute horizontal and vertical direction attention respectively. The gray box in Fig. 2 shows our spatial attention. The attended pyramidal semantics are further disposed of via two parallel convolutions with the above two fiter kernels. Then their concatenation serves as attention mechanism to handle initial appearance cues, removing the texture and details that belong to *BG*. After this, we concatenate the filtered appearance cues and distilled pyramidal features to achieve alpha mattes. The aggregation of channel-wise and spatial attention jointly optimize the alpha matte generation: one responsible for pyramidal features selection and the other responsible for appearance cues filtration. This well-designed hierarchical attention mechanism can efficiently attend low-level and semantic features, and their aggregation produce high-quality alpha mattes with fine-grained details.

3.3. Loss Function

Pixel regression related loss functions ($\mathcal{L}1$ or *MSE* loss) are usually adopted as the loss function for alpha matte prediction [3, 37]. They can generate competent alpha mattes via pixel-wise supervision. However, such regression loss only measures the difference in absolute pixels space, without consideration of *FG* structure. Therefore, we introduce *SSIM* loss ($\mathcal{L}_{SSIM}$) to calculate structure similarity between the predicted alpha mattes and ground truth. Structural SIMilarity (*SSIM*) [35] has demonstrated a striking ability to boost structure consistency in the predicted images [25, 31]. Apart from the aforementioned loss functions, we add adversarial loss ($\mathcal{L}_{adv}$) [13] 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:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{MSE} + \lambda_3 \mathcal{L}_{SSIM}, \qquad (2)$$

$\mathcal{L}_{adv}$, $\mathcal{L}_{MSE}$ and $\mathcal{L}_{SSIM}$ can improve alpha mattes from visual quality, pixel-wise accuracy and *FG* structure similarity separately. λ_1, λ_2 and λ_3 represent balance coeffi-

cients for loss functions. $\mathcal{L}_{adv}$ is defined as:

$$\mathcal{L}_{adv} = E_{(I,A)}[log(D(I,A)) + log(1 - D(I, G(I)))], \quad (3)$$

where I represents the input image and A is the predicted alpha matte. $\mathcal{L}_{MSE}$ is expressed as:

$$\mathcal{L}_{MSE} = \frac{1}{|\Omega|} \sum_{i}^{\Omega} (\alpha_p^i - \alpha_g^i)^2, \quad \alpha_p^i, \alpha_g^i \in [0, 1], \quad (4)$$

where Ω represents pixels set and $|\Omega|$ is the number of pixels (*i.e.* the size of the input image). α_p^i and α_g^i are the predicted and ground truth alpha values at pixel i respectively. $\mathcal{L}_{MSE}$ can ensure the pixel-wise accuracy of alpha matte estimation. We establish *FG* structure optimization via $\mathcal{L}_{SSIM}$ as:

$$\mathcal{L}_{SSIM} = 1 - \frac{(2\mu_p\mu_g + c_1)(2\sigma_{pg} + c_2)}{(\mu_p^2 + \mu_g^2 + c_1)(\sigma_p^2 + \sigma_g^2 + c_2)}. \quad (5)$$

here μ_p, μ_g and σ_p, σ_g are the mean and standard deviations of α_p^i and α_g^i separately. With $\mathcal{L}_{SSIM}$ as guidance, our method can further improve *FG* structure.

3.4. Implementation Details

We implement *HAttMatting* using PyTorch. For training, all input images are randomly cropped to 512 $\times$ 512 and 640 $\times$ 640 and 800 $\times$ 800. Then, they were resized to a resolution of 512 $\times$ 512 and augmented by horizontally random flipping. In order to accelerate the training process and prevent over-fitting, we use the pre-trained ResNeXt-101 network [36] as the feature extraction network, while the other layers are randomly initialized from a Gaussian distribution. For loss optimization, we use the stochastic gradient descent (*SGD*) optimizer with the momentum of 0.9 and a weight decay of 0.0005. The learning rate is initialized to 0.007, adjusted by the "poly" policy [22] with the power of 0.9 for 20 epochs. The balance coefficients λ_1, λ_2 and λ_3 in Eq. 2 are 0.05, 1 and 0.1 during the first epoch and revised as 0.05, 1 and 0.025 for subsequent 19 epochs. Our *HAttMatting* is trained on a single GPU with a mini-batch size of 4, and it takes about 58 hours for the network to converge on Tesla P100 graphics cards.

4. Experiments

In this section, we evaluate *HAttMatting* on two datasets: the public Adobe Composition-1k [37] and our Distinctions-646. We first compare *HAttMatting* with state-of-the-art methods both quantitatively and qualitatively. Then we perform an ablation study for *HAttMatting* on both datasets to demonstrate the significance of several crucial components. Finally, we execute *HAttMatting* on real scenarios to generate alpha mattes.

4.1. Datasets and Evaluation Metrics

Datasets. The first dataset is the public Adobe Composition-1k [37]. 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 MS COCO dataset [21] to composite the input images. For test set, the Composition-1k contains 50 *FG* images as well as the corresponding alpha mattes, and 1000 *BG* images from PASCAL VOC2012 dataset [10]. The training and test sets were synthesized through the algorithm provided by [37].

The second is our Distinctions-646 dataset. The Adobe Composition-1K contains many consecutive video frames, and cropped patches from the same image, and there are actually only about 250 dissimilar *FG* objects in their training set. To improve the versatility and robustness of the matting network during training, we construct our Distinctions-646 dataset comprised of 646 distinct *FG* images. We divide these *FG* examples into 596 and 50, and then produce $59,600$ training images and 1000 test images according to the composition rules in [37].

Evaluation metrics. We evaluate the alpha mattes following four common quantitative metrics: the sum of absolute differences (SAD), mean square error (MSE), the gradient (Grad) and connectivity (Conn) proposed by [27]. A better image matting method shall produce high-quality alpha mattes, thus reducing the values of the above all four metrics.

4.2. Comparison to the State-of-the-art

Evaluation on the Composition-1k test set. Here we compare *HAttMatting* with 6 traditional matting methods: Shared Matting [12], Learning Based [41], Global Matting [26], ClosedForm [19], KNN Matting [6], Information-Flow [1], and 8 deep learning based methods: DCNN [8], DIM [37], AlphaGAN [24], SSS [2], SampleNet [30], Context-aware [15], IndexNet [23], Late Fusion [40]. SSS, Late Fusion and our *HAttMatting* can generate alpha mattes without trimap. For the other methods, we feed *RGB* images and trimaps produced by 25 pixels random dilation refer to [37]. We use full-resolution input images for fair contrast and the visual results are illustrated in Fig. 4. The quantitative comparisons are reported in Tab. 1, and the four metrics are all calculated on the whole image.

The *HAttMatting* exhibits significant superiority over traditional methods, which can be clearly observed in Fig. 4 and Tab. 1. Compared to deep learning based approaches, the *HAttMatting* has more sophisticated details than DCNN, DIM, SSS and Late Fusion, and is better than SampleNet, since we employ hierarchical attention mechanism to distill advanced semantics and appearance cues, and their aggregation achieves complete *FG* profiles and boundaries. Our *HAttMatting* is slightly inferior to Context-Aware and IndexNet. The former establishes two branches and resorts

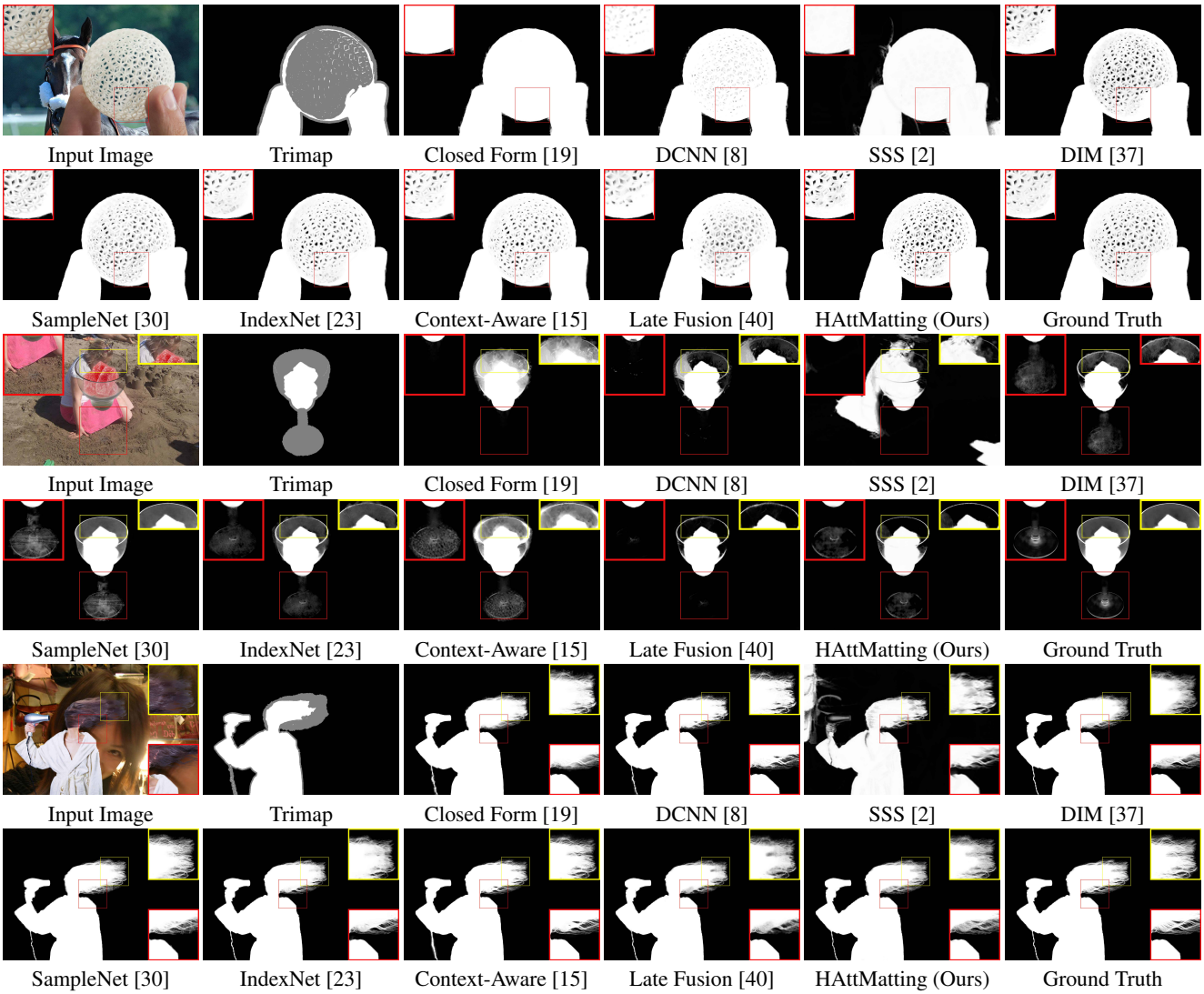


Figure 4: The visual comparisons on the Composition-1k test set. The segments in SSS [2] are hand-picked.

to *FG* image supervision to predict alpha mattes, while the latter learns index functions to capture texture and boundary details. 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 very convenient for novice users.

Evaluation on our Distinctions-646. For our Distinctions-646 dataset, we compare *HAttMatting* with 8 recent state-of-the-art matting methods, including Shared Matting [12], Learning Based [41], Global Matting [26], ClosedForm [19], KNN Matting [6], DCNN [8], Information-Flow [1] and DIM [37]. For other deep learning based methods, since their training codes are unavailable for us, we can not evaluate them on our dataset. We also use random dilation to generate high-quality trimaps [37] and relevant metrics are computed on the whole image.

The quantitative comparisons are displayed in Tab. 2. Our *HAttMatting* shows a clear advantage on all four metrics compared to all the traditional methods, and is better than DIM [37] on Grad and Conn metrics, while slightly worse than it in SAD metric. It is noting that only our method can generate alpha mattes without trimaps, and all the other methods demand trimaps to confine the transition region, which effectively improves the performance of these methods. Fig. 5 illustrates the visual comparison with DIM [37] 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

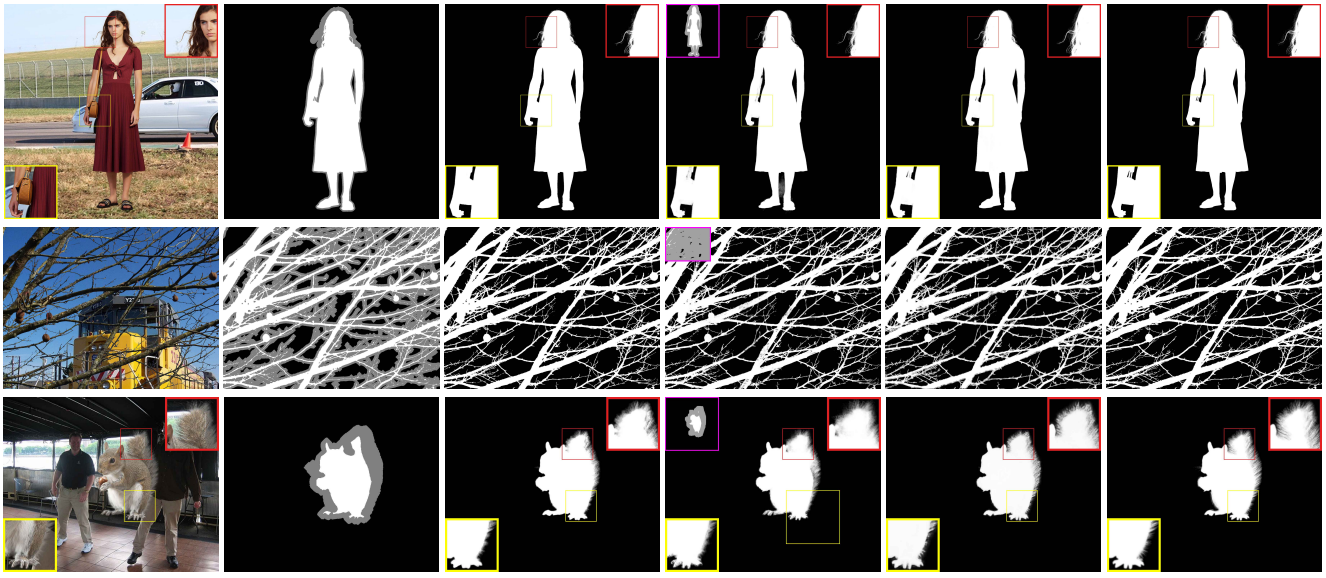


Figure 5: The visual comparisons on our Distinctions-646 test set. The "DIM+Large" means that we feed DIM with trimaps that have larger transition region, while our method can generate high-quality alpha mattes without trimaps.

Methods	SAD↓	MSE↓	Grad↓	Conn↓
Shared Matting [12]	125.37	0.029	144.28	123.53
Learning Based [41]	95.04	0.018	76.63	98.92
Global Matting [26]	156.88	0.042	112.28	155.08
ClosedForm [19]	124.68	0.025	115.31	106.06
KNN Matting [6]	126.24	0.025	117.17	131.05
DCNN [8]	115.82	0.023	107.36	111.23
Information-Flow [1]	70.36	0.013	42.79	70.66
DIM [37]	48.87	0.008	31.04	50.36
AlphaGAN [24]	90.94	0.018	93.92	95.29
SampleNet [30]	48.03	0.008	35.19	56.55
Context-Aware [15]	38.73	0.004	26.13	35.89
IndexNet [23]	44.52	0.005	29.88	42.37
Late Fusion [40]	58.34	0.011	41.63	59.74
HAttMatting (Ours)	**44.01**	**0.007**	**29.26**	**46.41**
Basic	126.31	0.025	111.35	118.71
Basic + SSIM	102.79	0.021	88.04	110.14
Basic + Low	89.39	0.016	56.67	90.03
Basic + CA	96.67	0.018	73.94	95.08
Basic + Low + CA	72.73	0.013	49.53	65.92
Basic + Low + SA	54.91	0.011	46.21	60.40
Basic + Low + CA + SA	49.67	0.009	41.11	53.76

Table 1: The quantitative comparisons on Composition-1k test set. The methods in gray (the Late Fusion and our *HAttMatting*) only take *RGB* images as input, while the others require trimap as assistance to guarantee the accuracy of alpha mattes. "Basic" means our baseline network, and the corresponding "Basic+" represent that we assemble different components on the baseline to generate alpha mattes.

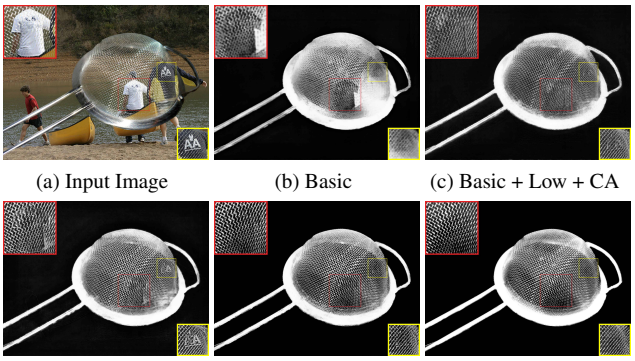
region expanded, which can verify that DIM has a strong dependence on the quality of trimaps. The alpha mattes produced by *HAttMatting* exhibit sophisticated texture details, which mainly benefits from the aggregation of adaptive semantics and valid appearance cues in our model.

Methods	SAD↓	MSE↓	Grad↓	Conn↓
Shared Matting [12]	119.56	0.026	129.61	114.37
Learning Based [41]	105.04	0.021	94.16	110.41
Global Matting [26]	135.56	0.039	119.53	136.44
ClosedForm [19]	105.73	0.023	91.76	114.55
KNN Matting [6]	116.68	0.025	103.15	121.45
DCNN [8]	103.81	0.020	82.45	99.96
Information-Flow [1]	78.89	0.016	58.72	80.47
DIM [37]	47.56	0.009	43.29	55.90
Basic	129.94	0.028	124.57	120.22
Basic + SSIM	121.79	0.025	110.21	117.41
Basic + Low	98.88	0.020	84.11	92.88
Basic + CA	104.23	0.022	90.87	101.9
Basic + Low + CA	85.57	0.015	79.16	88.38
Basic + Low + SA	78.14	0.014	60.87	71.90
Basic + Low + CA + SA	57.31	0.011	52.14	63.02
HAttMatting (Ours)	**48.98**	**0.009**	**41.57**	**49.93**

Table 2: The quantitative comparisons on our Distinctions-646 test set. The definition of "Basic" is the same with Tab. 1.

4.3. Ablation Study

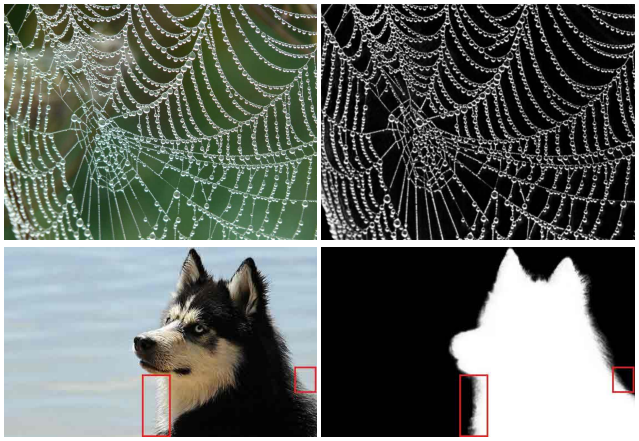
The core idea of our *HAttMatting* is to extract adaptive pyramidal features and filter low-level appearance cues, and then aggregate them to generate alpha mattes. To accomplish this goal, we employ channel-wise attention (CA) and spatial attention (SA) to re-weight pyramidal features and



(a) Input Image (b) Basic (c) Basic + Low + CA

(d) Basic + Low + SA (e) *HAttMatting* (f) Ground Truth

Figure 6: The visual comparison of different components. Each component has a significant improvement for alpha mattes.



Input Image Alpha Matte

Figure 7: The results on real world images.

appearance cues separately. We also introduce *SSIM* in our loss function to further improve the *FG* structure. And here we make different combinations of these components, and verify the significance of them on the Composition-1k and our Distinctions-646 datasets. **Basic:** This is our baseline network, which only uses original pyramidal features to generate alpha mattes, and optimized by $\mathcal{L}_{MSE}$ and $\mathcal{L}_{adv}$. **Basic + SSIM:** $\mathcal{L}_{SSIM}$ is involved in our loss function. **Basic + Low:** Low-level appearance cues are directly aggregated with pyramidal features, which can furnish sophisticated texture and details for alpha mattes. **Basic + CA:** On the basis of baseline, we perform channel-wise attention to distill pyramidal features. CA can effectively suppress unnecessary advanced semantics and reduce the sensitivity of the trained model to *FG* classes, which means the network can handle diverse *FG* objects and the model versatility is enhanced. **Basic + Low + CA:** This combination integrates the advantages of the above two modules to promote performance. **Basic + Low + SA:** Our modified SA can eliminate the *BG* texture in appearance cues, improving subsequent aggregation process. **Basic + Low + CA + SA:** We assemble CA, Low and SA to achieve competent alpha mattes without SSIM.

The quantitative results are shown in Tab. 1 and Tab. 2. It can be clearly seen that each component can significantly improve our results. The visual comparison is illustrated in Fig. 6. CA can furnish *FG* profiles (Fig. 6c) while SA can exhibit fine-grained internal texture and boundary details (Fig. 6d), and their aggregation can generate high-quality alpha mattes (Fig. 6e).

4.4. Results on Real-world Images

Fig. 7 shows our matting results on real-world images*. The evaluation model is trained on the Composition-1k dataset. We can see that *HAttMatting* is able to achieve high-quality alpha mattes without any external input or user interaction. However, if the input image has some blur (the hairs below the mouth of the dog), the *HAttMatting* can only predict ambiguous *FG* boundaries. The blur in the input images can obstruct our appearance cues filtration, and discount subsequent aggregation process.

*Please see the supplementary material for more matting results.

5. Conclusions and Future work

In this paper, we propose an Hierarchical Attention Matting Network (*HAttMatting*), which can predict high-quality alpha mattes from single *RGB* images. The *HAttMatting* employs channel-wise attention to extract matting-adapted semantics and performs spatial attention to filtrate appearance cues. Extensive experiments demonstrate that our hierarchical structure aggregation can effectively distill high-level and low-level features from the input images, and achieve high-quality alpha mattes without external trimaps.

In the future, we will explore more effective strategies to improve our attention mechanism, which we believe can more effectively aggregate advanced semantics and appearance cues, thus further improve the versatility and robustness of our network.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 91748104, Grant 61972067, Grant 61632006, Grant U1811463, Grant U1908214, Grant 61751203, in part by the National Key Research and Development Program of China under Grant 2018AAA0102003, Grant 2018YFC0910506, in part by the Open Research Fund of Beijing Key Laboratory of Big Data Technology for Food Safety (Project No. BTBD-2018KF).

References

[1] Yagiz Aksoy, Tunc Ozan Aydin, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In *CVPR*, 2017.

[2] Yağız Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. Semantic soft segmentation. *ACM TOG*, 2018.

[3] Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. Disentangled image matting. In *ICCV*, 2019.

[4] L. C. Chen, G Papandreou, I Kokkinos, K Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2018.

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

[6] Qifeng Chen, Dingzeyu Li, and Chi Keung Tang. Knn matting. *IEEE TPAMI*, 2013.

[7] D Cho, S Kim, Y. W. Tai, and I. S. Kweon. Automatic trimap generation and consistent matting for light-field images. *IEEE TPAMI*, 2016.

[8] Donghyeon Cho, Yu Wing Tai, and Inso Kweon. Natural image matting using deep convolutional neural networks. In *ECCV*, 2016.

[9] Yung Yu Chuang, B. Curless, D. H. Salesin, and R. Szeliski. A bayesian approach to digital matting. In *CVPR*, 2003.

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[11] Xiaoxue Feng, Xiaohui Liang, and Zili Zhang. A cluster sampling method for image matting via sparse coding. In *ECCV*, 2016.

[12] Eduardo S. L. Gastal and Manuel M. Oliveira. Shared sampling for real-time alpha matting. *CGF*, 2010.

[13] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Xu Bing, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

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

[15] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *ICCV*, 2019.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[17] L. Karacan, A. Erdem, and E. Erdem. Image matting with kl-divergence based sparse sampling. In *ICCV*, 2015.

[18] P Lee and Ying Wu. Nonlocal matting. In *CVPR*, 2011.

[19] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE TPAMI*, 2007.

[20] Anat Levin, Alex Rav-Acha, and Dani Lischinski. Spectral matting. *IEEE TPAMI*, 2008.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[22] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[23] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *ICCV*, 2019.

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

[25] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019.

[26] C. Rhemann and C. Rother. A global sampling method for alpha matting. In *CVPR*, 2011.

[27] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. In *CVPR*, 2009.

[28] Ehsan Shahrian, Deepu Rajan, Brian Price, and Scott Cohen. Improving image matting using comprehensive sampling sets. In *CVPR*, 2013.

[29] Jian Sun, Jiaya Jia, Chi Keung Tang, and Heung Yeung Shum. Poisson matting. *ACM TOG*, 2004.

[30] Jingwei Tang, Yagiz Aksoy, Cengiz Oztireli, Markus Gross, and Tunc Ozan Aydin. Learning-based sampling for natural image matting. In *CVPR*, 2019.

[31] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. Crrn: Multi-scale guided concurrent reflection removal network. In *CVPR*, 2018.

[32] Jue Wang and Michael F. Cohen. An iterative optimization approach for unified image segmentation and matting. In *ICCV*, 2005.

[33] Jue Wang and Michael F. Cohen. Optimized color sampling for robust matting. In *CVPR*, 2007.

[34] Yu Wang, Yi Niu, Peiyong Duan, Jianwei Lin, and Yuanjie Zheng. Deep propagation based image matting. In *IJCAI*, 2018.

[35] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004.

[36] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

[37] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *CVPR*, 2017.

[38] Xin Yang, Ke Xu, Shaozhe Chen, Shengfeng He, Baocai Yin Yin, and Rynson Lau. Active matting. In *NeurIPS*, 2018.

[39] Guan Yu, Wei Chen, Xiao Liang, Zi'ang Ding, and Qunsheng Peng. Easy matting - a stroke based approach for continuous image matting. *CGF*, 2006.

[40] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *CVPR*, 2019.

[41] Yuanjie Zheng and Chandra Kambhamettu. Learning based digital matting. In *ICCV*, 2009.

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *ICCV*, 2017.