SDNet: An Extremely Efficient Portrait Matting Model via Self-Distillation

Ziwen Li[1,*], Bo Xu[2,*], Jiake Xie[3], Yong Tang[3], Cheng Lu[2†]
[1]OPPO , [2]Xpeng, [3]Picup.AI
liziwennba@gmail.com

Abstract

Most existing portrait matting models either require expensive auxiliary information or try to decompose the task into sub-tasks that are usually resource-hungry. These challenges limit its application on low-power computing devices. In addition, mobile networks tend to be less powerful than those cumbersome ones in feature representation mining. In this paper, we propose an extremely efficient portrait matting model via self-distillation (SDNet), that aims to provide a solution to performing accurate and effective portrait matting with limited computing resources. Our SDNet contains only 2M parameters, 2.2% of parameters of MGM, and 1.5% of that of Matteformer. We introduce the training pipeline of self-distillation that can improve our lightweight baseline model without any parameter addition, network modification, or over-parameterized teacher models which need well-pretraining. Extensive experiments demonstrate the effectiveness of our self-distillation method and the lightweight SDNet network. Our SDNet outperforms the state-of-the-art (SOTA) lightweight approaches on both synthetic and real-world images.

1. Introduction

Portrait matting is a popular computer vision task that aims to extract accurate alpha mattes of humans in a given image or video. It has significant value in multimedia creation applications regardless of the scenarios, such as background replacement of live conferences, photo/video editions, and movie-making without the need to build a green or blue screen background. Although previous methods [18, 27, 32, 33, 46] utilize pre-defined trimaps (a draft marking foreground, background, and transition areas) as constraint information to reduce solution space and achieve brilliant performance on portrait matting, high prediction accuracy often comes with tedious manual annotation and time costs. Some trimap-free models are proposed to situate these problems without extra auxiliary cues (*e.g.* trimap or sematic mask) by learning the image saliency from object detection or segmentation methodology.

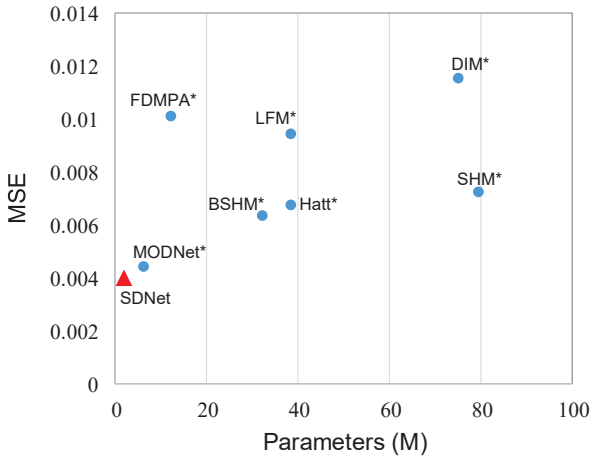

Figure 1. Parameters vs. performance of trimap-free models. * indicates that the models pre-trained on Supervisely Person Segmentation (SPS) dataset [45].

Current trimap-free methods often try to decompose the portrait matting into sequential sub-tasks that first generate pseudo-trimaps, semantic masks, or implicitly learn the transition region distribution before solving matting problems. While multi-stage tasks often come at large model sizes and the expense of high computational costs, which limit matting methods applied in applications on mobile devices with low-power computing capability. One of the direct solutions to situate this is to adopt lightweight backbones such as MobileNet [19] instead of widely-used cumbersome backbones such as ResNet [17] and VGG [44]. In addition, other model compression and acceleration methods such as knowledge distillation (KD) can also be introduced to address this issue. KD is a typically effective approach that is extensively explored in multiple computer vision tasks [31, 47, 49, 51] such as segmentation, it trains a lighter student network with the guidance of an over-parametrized teacher model.

However, there still remain two big challenges in the above solutions. First, lightweight networks tend to be less

*Ziwen Li and Bo Xu contribute equally.
†The corresponding author.

powerful than those cumbersome ones in feature representation mining, which has been widely acknowledged in previous lightweight backbone research [19, 42, 52]. Second, conventional knowledge distillation (KD) often suffers from the following problems: 1) Traditional KD needs a two-step implementation which is to first train a heavy teacher and then transfer its knowledge to a light student. 2) The choice of a teacher model has a great impact on the performance of the student model and the model with the best performance may not be the best teacher [10, 23, 36]. Substantial attempts to select or design the best teacher model and the two-step training mechanism of KD can lead to long time consumption. 3) The student model scarcely exploits all the information transferred from the teacher model, which leads to inefficient knowledge transfer and a significant gap between the student and teacher.

To address the above challenges, in this paper, we propose an Extremely Efficient Portrait Matting Model via Self-Distillation (SDNet), that aims to provide a solution to performing accurate and effective portrait matting on low-power computing devices. We employ ILBlock [8] as the backbone to build a computationally efficient lightweight baseline network of our SDNet. To further improve the matting performance and address the issues of traditional KD, we propose a self-distillation (SD) method that seeks privileged information from downstream features as teacher knowledge to guide the upstream features throughout the decoding process. Compared with traditional KD methods, our SDNet performs only a one-step training process and achieves significant model self-improvement without the modification of network structure or addition of parameters. As illustrated in Figure 1, our model achieves the most significant performance with the least parameters among SOTA trimap-free models.

Overall, the contributions of this paper are as follows:

- We propose a novel extremely efficient and lightweight portrait matting model with only 2M parameters.
- We introduce the training pipeline of self-distillation that can improve our lightweight baseline model without any parameter addition, network modification, or over-parameterized teacher models which need well-pretraining.
- Extensive experiments demonstrate the effectiveness of our self-distillation method and the lightweight SDNet network. Our SDNet outperforms the state-of-the-art (SOTA) approaches on both synthetic and real-world images.

2. Related works

Currently, the matting is generally formulated as an image composite problem, which solves the 7 unknown variables per pixel from only 3 known values:

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i \tag{1}$$

where 3 dimensional RGB color I_i of pixel i, while foreground RGB color F_i, background RGB color B_i, and matte estimation α_i are unknown. In this section, we discuss the SOTA works trying to solve this under-determined equation.

2.1. Classic methods

Classic foreground matting methods can be generally categorized into two approaches: sampling-based and propagation-based. Sampling-based methods [6, 22] sample the known foreground and background color pixels, and then extend these samples to achieve matting in other parts. Various sampling-based algorithms are proposed, $e.g.$, global sampling method [22] and comprehensive sampling [13]. Propagation-based methods [1, 5] reformulate the composite Eq. 1 to propagate the alpha values from the known foreground and background into the unknown region, achieving more reliable matting results. [21] provides a very comprehensive review of various matting algorithms.

2.2. Deep learning-based methods

Classic matting methods are carefully designed to solve the composite equation and its variant versions. However, these methods heavily rely on chromatic cues, which leads to bad quality when the color of the foreground and background show small or no noticeable difference.

Trimap-based methods. Automatic and intelligent matting algorithms are emerging, due to the rapid development of the deep neural network in computer vision. Initially, some attempts were made to combine deep learning networks with classic matting techniques, $e.g.$ KNN matting [5]. Cho *et al.* [9] employ a deep neural network to improve the results of the closed-form matting and KNN matting. These attempts are not end-to-end, so not surprisingly the matting performance is limited by the convolution back-ends. Subsequently, full DL image matting algorithms appear [4, 46]. Xu *et al.* [46] propose a two-stage deep neural network (Deep Image Matting) based on SegNet for alpha matte estimation and contribute a large-scale image matting dataset (Adobe dataset) with ground truth foreground (alpha) matte, which can be composited over a variety of backgrounds to produce training data. We also use this data for the first-step pre-training of our network. Lutz *et al.* [35] introduce a generative adversarial network (GAN) for natural image matting and improve the results of Deep Image Matting [46]. Cai *et al.* [2] investigate the bottleneck of the previous methods that directly estimate the alpha matte from a coarse trimap, and propose to divide the matting problem into trimap adaptation and alpha estimation tasks. Hou *et al.* [18] employs two encoder networks

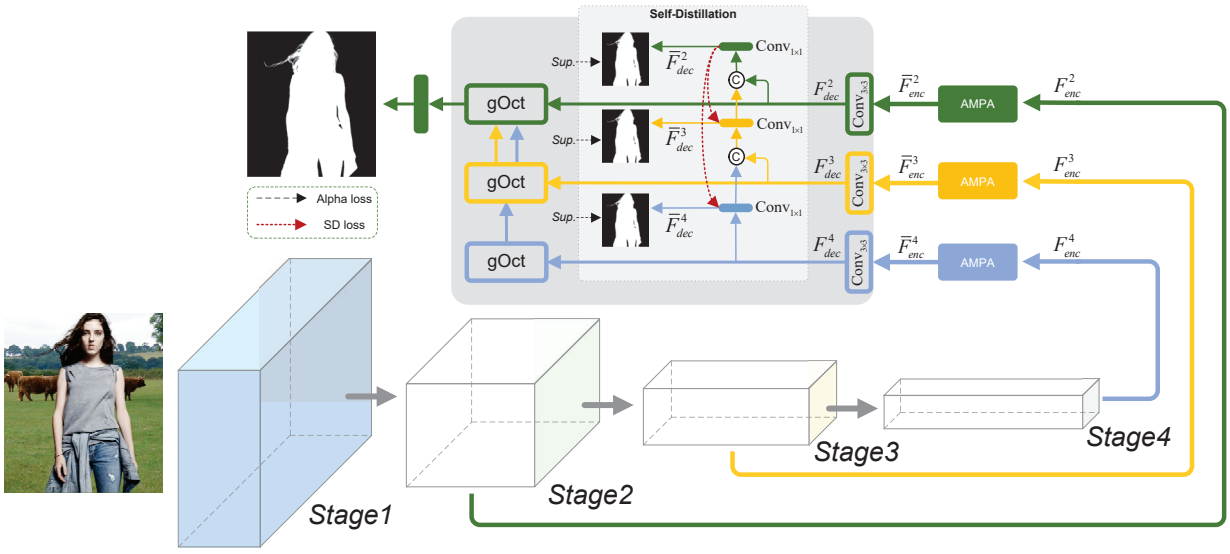

Figure 2. The architecture of the SExtremely Efficient Portrait Matting Model via Self-Distillation (SDNet). The SDNet employs IL-Block [8] as the backbone to build a 4-stage encoder. An Adaptive Multi-Perspective Aggregation (AMPA) module is followed and can aggregate multiple receptive fields in one layer to fully mine both high-level and low-level features. We perform self-distillation in the process of our proposed multi-level aggregation decoder.

to extract essential information for matting, however, it is not robust to faulty trimaps. Forte *et al*. [29] propose a low-cost upgrade to alpha matting networks to also predict the foreground and background colours. They study variations of the training regions and explore a wide range of existing and novel loss functions for optimal prediction.

Additional natural background. Qian *et al*. [39] compute a probability map to classify each pixel into the foreground or background by simple background subtraction. This algorithm is sensitive to the threshold and fails when the colors of the foreground and background are similar. Sengupta *et al*. [43] introduce a self-supervised adversarial approach - Background Matting (BGM), achieving state-of-the-art results. However, as a prerequisite, the photographer needs to take a shoot of natural background first, which is not friendly to the intensive multi-scene shooting application. Liu *et al*. [28] propose the Background Matting V2 that employs two neural networks: a base network computes a low-resolution result which is refined by a second network operating at high-resolution on selective patches.

Trimap-free methods. Currently, a majority of deep image matting algorithms [2, 18, 35, 46] try to estimate a boundary that divides the foreground and background, with the aid of a user-generated trimap. Several trimap-free matting methods [4, 53] predict the trimap first, followed by alpha matting. Qiao *et al*. [40] employ spatial and channel-wise attention to integrating appearance cues and pyramidal features, they 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 in trimap-free matting. Lin *et al*. [29] propose a robust real-time matting method (RVM) training strategy that optimizes the network on both matting and segmentation tasks. Ke *et al*. [24] present a lightweight matting objective decomposition network (MODNet) by optimizing a series of sub-objectives simultaneously via explicit constraints. They also introduce an e-ASPP module to fuse multi-scale features, plus a self-supervised sub-objectives consistency (SOC) strategy to address the domain shift problem, which is common in trimap-free methods.

Besides, most current trimap-free methods focus only on human/portrait matting but ignore the objects that are interacting with or attached to people. In addition, they learn the saliency of the images by data-driven training, which lacks the situational perception between salient objects and the surrounding environment, leading to biased or incomplete foreground prediction, especially in multi-object scenes. This is the main reason why we propose the method of Situational Perception Guided Image Matting. In this paper, we quantitatively evaluate the performance of our model for alpha matting in human-object interactive and multi-object scenes.

2.3. Self-Distillation

Self-distillation is a technique that has gained increasing attention in the deep learning community. It involves training a neural network to mimic its own behavior in or-

der to improve its performance. [51] introduced a method that utilizes a deeper classifier to teach a shallower classifier, resulting in improved performance while maintaining the response time. This approach demonstrates the potential of self-distillation in balancing accuracy and computational efficiency. [47] proposes to maintain the consistency of feature maps and predictions between different distorted data. Their approach highlights the importance of consistency in self-distillation and its role in enhancing model performance. [49] presented a method that pre-trains a student model and subsequently uses it to generate soft labels for self-distillation. Their work shows the effectiveness of using soft labels to guide the learning process in self-distillation. Despite the success of self-distillation in various deep learning applications, its adoption in the context of image matting remains relatively unexplored. In this work, we aim to address this gap by proposing a novel self-distillation method tailored for portrait matting tasks using an extremely lightweight network.

3. Methodology

In this section, we describe our proposed lightweight network SDNet and the self-distillation method which does not need to change the network structure or add any parameters.

3.1. Multi-scale feature exaction

As illustrated in Figure 2, SDNet is based on an encoder-to-decoder architecture. The encoder employs ILBlock [8] as the backbone to extract features within multiple scales. ILBlock [8] is a lightweight feature extractor that is first applied in the salient object detection (SOD) task. ILBlock [8] consists of a vanilla OctConv [7] and two 3×3 simplified gOctConvs [8]. Compared to the manual setting of channels in the vanilla OctConv [7], gOctConv [8] is a flexible self-adaptive convolutional operator which takes inputs of multiple arbitrary scales and does not need lots of effort to re-adjust for models. The output of gOctConv [8] is also flexible and fits well with the matting network.

We split our encoder into 4 stages, which are stacked with 3,4,6,4 ILBlocks [8]. Each stage takes the input of resolution (H_i, W_i) and outputs the features of resolution (H_i, W_i) and (H_i/2, W_i/2). The ILBlocks [8] can enable each stage to integrate multi-scale features while keeping lightweight. Therefore, the output of our multi-stage encoder contains both high-level features with abstract semantics and low-level features with detail distributes that contribute to the matting task. For simplicity, we denote the multi-stage feature representations of the encoder as a set $F_{enc} = \{F^1_{enc}, F^2_{enc}, F^3_{enc}, F^4_{enc}\}$.

3.2. Adaptive Multi-Perspective Aggregation

As described in previous dense prediction work [38], larger receptive fields establish dense connections between feature maps and per-pixel classifiers which improve the accuracy of internal regions, while smaller receptive fields benefit the localization focus on local fine-grained details near the object boundaries. Inspired by this, we propose an Adaptive Multi-Perspective Aggregation (AMPA) module and fuse multiple receptive fields in one layer to enhance the diversity of reception in feature representations that cover both global semantic integrity and detail attributes. Technically, we connect standard depthwise separable convolution [11], atrous convolution [3], and asymmetric convolution [12] parallel for multi-scale fusion. This can be summarized as follows:

$$\begin{aligned} \chi^1_1 &= \mathrm{Conv}_d(F^i_{enc}) \\ \chi^2_1 &= \mathrm{Conv}_{atr}(F^i_{enc}) \\ \chi^3_1 &= \mathrm{Conv}_{asy}(F^i_{enc}) \\ \chi_2 &= \mathrm{RELU}(\mathrm{LN}(\mathrm{Concat}[\chi^1_1, \chi^2_1, \chi^3_1])) \end{aligned} \tag{2}$$

where F^i_{enc} denotes the feature produced by the encoder at each stage, Conv_* denote different types of convolutional operations ($i.e.$, depthwise separable, atrous, and asymmetric convolutions), and BN is the abbreviation of batch normalization, and Concat[*] is the concatenation operation. A 1×1 convolution is followed to squeeze channels of χ_2 to the same number as the input of F^i_{enc}. And then, we utilize the SEblock [20] (see Figure 4) to model the channel-wise attention for effective enhancement of the fused feature representations, as follows:

$$\begin{aligned} \chi_3 &= \mathrm{SE}(\mathrm{Conv}_{1\times1}(\chi_2)) \\ \bar{F}^i_{enc} &= \mathrm{Conv}_{1\times1}(\chi_3) + F^i_{enc} \end{aligned} \tag{3}$$

where we use a residual connection to fuse χ_2 with the input of F^i_{enc} for better optimization. The network structure of AMPA is also shown in Figure 3.

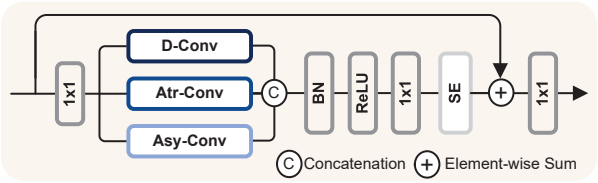

Figure 3. Network structure of Adaptive Multi-Perspective Aggregation (AMPA) module.

3.3. Multi-level aggregation decoder

In order to better aggregate multi-level features and fully mine both semantic abstract information and fine-grained details, we design a multi-level aggregation decoder. As shown in Figure 2, the output features ($\bar{F}^i_{enc}, i = 2, 3, 4$) of the adaptive multi-perspective aggregation (AMPA) module are first sent to 3×3 convolutional layers that output

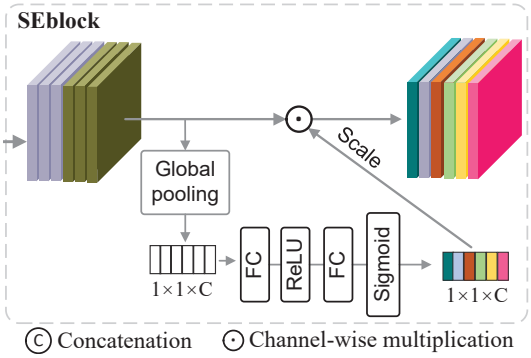

Figure 4. Architecture of the SEBlock.

$F^i_{dec}, i = 2, 3, 4$, respectively. F^4_{dec} is processed by a 1×1 convolutional layer and output $\bar{F}^4_{dec}$ whose channel number is reduced to 1. And then, $\bar{F}^{i+1}_{dec}$ from the previous decoder stage is upsampled and fused with current stage feature $F^i_{dec}, i = 2, 3$ for multi-stage feature aggregation, as follows:

$$\bar{F}^i_{dec} = \text{Conv}_{1\times1}(\text{Concat}[\text{UP}(\bar{F}^{i+1}_{dec}), F^i_{dec}]) \quad (4)$$

where a 1×1 convolution is followed to reduce the channel numbers of $\bar{F}^i_{enc}, i = 2, 3$ to 1, and UP is the upsampled operation.

Inspired by multiple salient object detection methods, we utilize the multi-level supervision strategy to supervise the feature of $\bar{F}^i_{dec}, i = 2, 3, 4$ by the ground truth alpha matte (α_{gt}). We recover the spatial size of $\bar{F}^i_{dec}$ to the same as the original input and the loss is calculated as:

$$L_{aux} = \sum_{i=1}^{3} L_{alpha}$$

where L_{alpha} is the sum of L1 loss L_{l1}, Laplacian loss L_{lap} and composition loss L_c:

$$L^i_{l1} = ||\text{UP}(\bar{F}^i_{dec}) - \alpha_{gt}||_1$$

$$L^i_{lap} = \sum_{n=1}^{5} \frac{2^{n-1}}{5} ||L^n_{pyr}(\text{UP}(\bar{F}^i_{dec})) - L^n_{pyr}(\alpha_{gt})||_1$$

$$L^i_c = ||\text{UP}(\bar{F}^i_{dec}) * F_{gt} + (1 - \text{UP}(\bar{F}^i_{dec})) * B_{gt} - I_{gt}||_1$$

For better multi-scale feature aggregation, we send $\bar{F}^i_{dec}, i = 2, 3, 4$ to three individual gOctConv blocks and the feature aggregation can be summarized as:

$$\begin{aligned} \bar{\chi}_4 &= \text{gOct}(\bar{F}^4_{dec}) \\ \bar{\chi}_3, \bar{\chi}^*_4 &= \text{gOct}(\bar{F}^3_{dec}, \bar{\chi}_4) \\ \bar{\chi}_2 &= \text{gOct}(\bar{F}^2_{dec}, \bar{\chi}_3, \bar{\chi}^*_4) \\ \alpha_f &= \text{Conv}_{1\times1}(\bar{\chi}_2) \end{aligned} \quad (5)$$

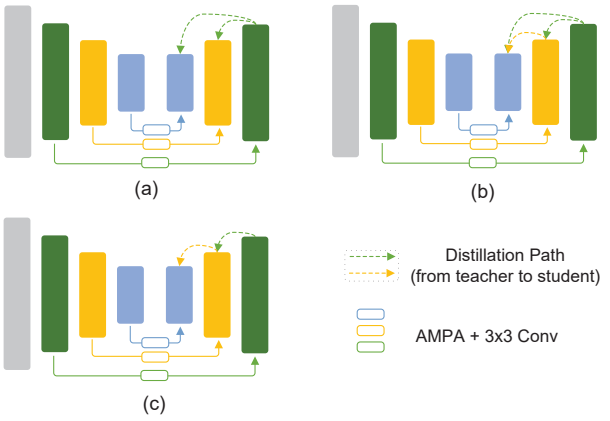

Figure 5. The settings of different reference features: (a) best teacher distillation; (b) dense distillation; (c) transitive distillation.

where the final gOctConv outputs a high-resolution feature $\bar{\chi}_2$ which is used to predict a final alpha matte α_f after a 1×1 convolutional operation. α_f is also supervised by the alpha matte ground truth α_{gt} through L_{alpha}.

3.4. Self-Distillation

We introduce the training pipeline of self-distillation that can improve our lightweight baseline model. Compared to the original model, the self-distillation does not change the network structure or add any parameters addition. Due to the model design of our decoder, we observe that the downstream-stage features often integrate more multi-scale features from upstream-stage features and the downstream features are closer to the predicted values. We believe that privileged information from the downstream features can guide the representations of upstream features. Therefore, we introduce a self-distillation training pipeline, which seeks downstream features as teachers to facilitate upstream features. Specifically, the self-distillation (SD) loss L_d is defined by the cosine similarity between the reference (teacher) feature and the student feature. Inspired by [50], we compare three settings of reference features, as follows:

- **Best teacher distillation.** $\bar{F}^2_{dec}$ is chosen as reference feature to guide $\bar{F}^3_{dec}$ and $\bar{F}^4_{dec}$.
- **Dense distillation.** Each downstream feature (*i.e.*, $\bar{F}^2_{dec}$, ..., $\bar{F}^{i-1}_{dec}$) are utilized as reference features to teach the upstream feature $\bar{F}^i_{dec}$.
- **Transitive distillation.** Each downstream feature is selected as the reference feature only for its adjacent upstream feature.

The self-distillation settings are also illustrated in Figure 5. Among them, the best teacher distillation performs the best on our self-distillation training pipeline. The comparison result is also discussed in the ablation study section.

Methods	Param (M)	Trimap	SAD↓	MSE↓
DIM [46]	75.1	✓	6.9	0.0014
IndexNet [33]	8.2	✓	6.6	0.0013
MODNet [15]	6.5	✓	6.3	0.0011
GCA Matting	24.8	✓	2.6	0.0003
MGM	90.4	✓	2.4	0.0003
FBA Matting	34.7	✓	2.3	0.0003
Matteformer	138.6	✓	**2.0**	**0.0002**
Ours w/o SD	2.0	✓	3.3	0.0005
Ours	2.0	✓	2.7	0.0003
$DIM^\star$ [46]	75.1		11.6	0.0048
$FDMPA^\star$	12.3		11.5	0.0047
$LFM^\star$	38.5		10.1	0.0043
$SHM^\star$	79.6		9.2	0.0031
$HAtt^\star$	38.7		9.4	0.0034
$BSHM^\star$	32.3		8.8	0.0029
$MODNet^\star$	6.5		7.7	0.0023
Ours w/o SD	2.0		7.6	0.0028
Ours	2.0		**6.2**	**0.0016**

Table 1. The quantitative results on the portrait subset of Composition-1k. w/o SD denotes the SDNet baseline network without self-distillation. * indicates that the models pre-trained on Supervisely Person Segmentation (SPS) dataset [45].

Methods	Param (M)	Trimap	SAD↓	MSE↓
DIM [46]	75.1	✓	8.1	0.0025
IndexNet [33]	8.2	✓	7.8	0.0022
MGM	90.4	✓	6.8	0.0006
GCA Matting	24.8	✓	6.0	0.0006
FBA Matting	34.7	✓	4.4	0.0004
Matteformer	138.6	✓	**3.3**	0.0002
Ours w/o SD	2.0	✓	5.9	0.0007
Ours	2.0	✓	4.7	0.0004
LFM	38.5		16.9	0.0087
$HAtt^\star$	38.7		12.6	0.0054
$MODNet^\star$	6.5		9.8	0.0037
Ours w/o SD	2.0		10.3	0.0043
Ours	2.0		**8.7**	**0.0028**

Table 2. The quantitative results on the portrait subset of Distinction-646.

4. Experiments

We first describe the datasets used for training and testing. Subsequently, we compare our results with existing state-of-the-art (SOTA) foreground matting algorithms. Finally, we conduct ablation experiments to show the effectiveness of each branch and module.

Methods	Param (M)	Trimap	SAD↓	MSE↓
DIM [46]	75.1	✓	6.7	0.0016
IndexNet [33]	8.2	✓	6.4	0.0015
MODNet	6.5	✓	5.4	0.0013
GCA Matting	24.8	✓	5.0	0.0013
MGM	90.4	✓	4.6	0.0011
FBA Matting	34.7	✓	3.7	0.0009
Matteformer	138.6	✓	**3.2**	0.0007
Ours w/o SD	2.0	✓	5.4	0.0016
Ours	2.0	✓	4.4	0.0011
DIM [46]	75.1		32.7	0.0221
$DIM^\star$ [46]	75.1		17.8	0.0115
$FDMPA^\star$	12.3		16.0	0.0101
$LFM^\star$	38.5		15.8	0.0094
$SHM^\star$	79.6		15.2	0.0072
$HAtt^\star$	38.7		13.7	0.0067
$BSHM^\star$	32.3		11.4	0.0063
$MODNet^\star$	6.5		8.6	0.0044
Ours w/o SD	2.0		8.8	0.0060
Ours	2.0		**7.2**	**0.0040**

Table 3. The quantitative results on PPM-100.

4.1. Datasets

Adobe Image Matting(AIM) [46] is a widely-used dataset for image matting. We select 78 human foreground images within it for training and 22 for testing. Distinction-646 [40] contains 342 human images for training and 22 for testing. During training, We composite each foreground with a random background from COCO [30] images. We apply random affine, cropping, and color jitters to every training sample. For evaluation, we composite each test foreground with 10 diverse backgrounds from VOC [14]. In addition, we also conduct comparative real-world portrait dataset PPM-100 [25] for testing.

4.2. Evaluation metrics

We report the mean square error (MSE) and the sum of the absolute difference (SAD) between predicted and ground truth alpha mattes. Lower values of these metrics indicate better estimated alpha matte.

4.3. Implementation detail

We use Adam [26] Optimizer for training. The model is trained from scratch without pretraining. The batch size is set to 4 and the initial learning rate is 0.0001. We train our network for 120 epochs and the learning rate is multiplied by 0.1 at 75 and 100 epochs.

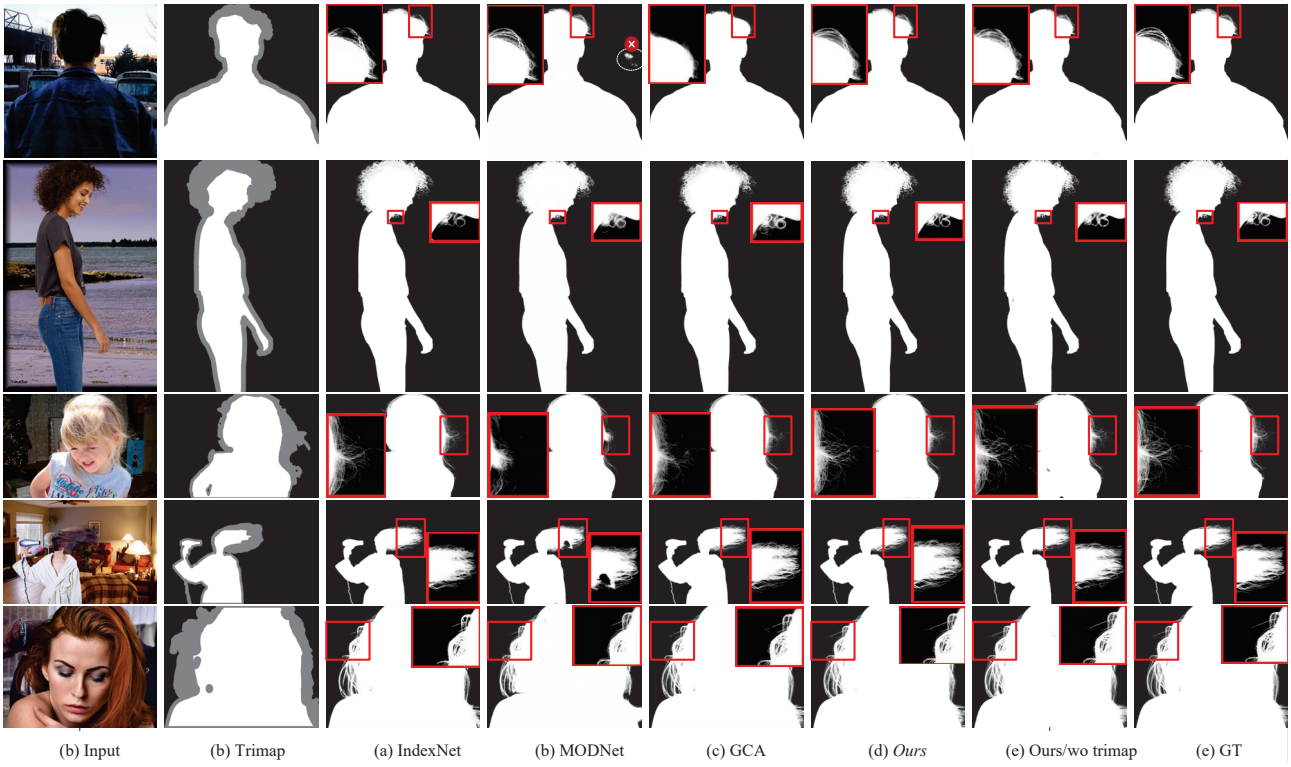

Figure 6. Visual comparisons on public composition datasets. Row 1-2: visualizations on Distinction-646; Row 3-5: visualizations on Composition-1k.

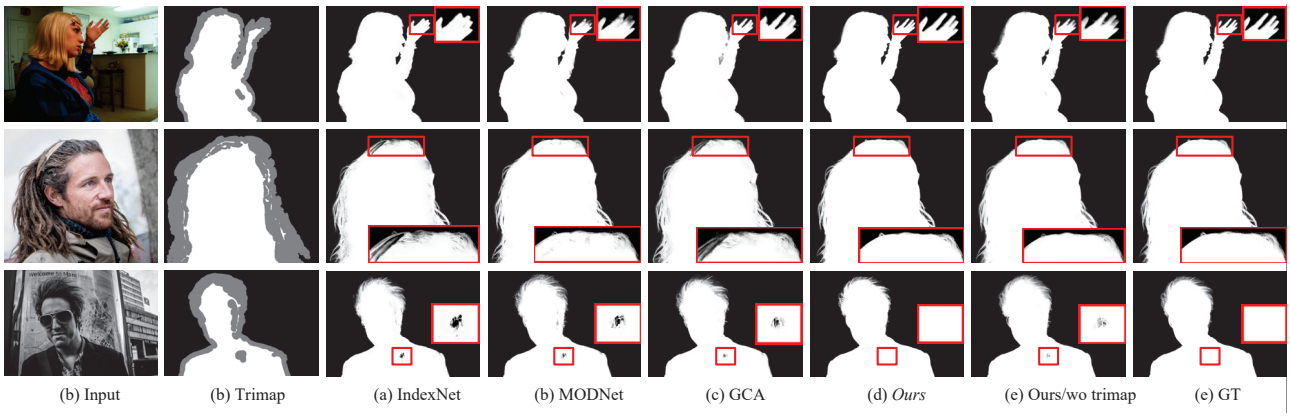

Figure 7. Visual comparisons on PPM-100.

4.4. Comparative study

We conduct comparative study on dual composition benchmarks: Composition-1K [46] and Distinction-646 [40], and a real-world natural portrait benchmark: PPM-100 [25]. We compare our models with the state-of-the-art trimap-based (*i.e.*, DIM [46], IndexNet [34], GCA [27], FBA [16], and Matteformer [37]) and trimap-free (*i.e.*, FDMPA [54], LFM [53], SHM [4], Hatt [40], BSHM [41], and MODNet [25]), and mask-guided (*i.e.*, MGM matting [48]) methods. For MGM, we choose its trimap-based version throughout the experiments.

The quantitative results on the Composition-1K [46] are shown in Table 1. Our trimap-based SDNet is slightly inferior to trimap-based GCA [27], MGM [48], FBA [16], and Matteformer [37] whose backbones are more cumbersome but superior to the trimap-based DIM [46], IndexNet [34], and MODNet [25]. Among trimap-free methods, our SD-

Methods	Parameters(Million)	Size(MB)
DIM [46]	75.06	292.0
IndexNet [33]	8.15	22.9
GCA Matting	24.8	96.5
FBA Matting	34.7	138.8
MGM	90.4	340.0
Matteformer	138.6	513.0
FDMPA	12.3	47.6
LFM	38.5	146.2
SHM	79.6	299.4
BSHM	32.3	124.1
HAtt	38.7	146.6
MODNet [15]	6.48	25.0
Ours	**2.03**	**8.1**

Table 4. The comparison of model size.

SD settings	Trimap	SAD↓	MSE↓
w/o SD	✓	3.3	0.0005
Transitive	✓	2.9	0.0004
Dense	✓	2.8	0.0004
Best teacher	✓	**2.7**	**0.0003**
w/o SD	✓	7.6	0.0028
Transitive		6.4	0.0020
Dense		6.5	0.0021
Best teacher		**6.2**	**0.0016**

Table 5. Ablation of different reference feature settings in self-distillation. w/o SD denotes the SDNet baseline without self-distillation.

Net significantly outperforms other competing SOTA ones without any extra pretraining on segmentation datasets($e.g.$, SPS [45]). Table 2 and Table3 also shows similar experimental effects on the other composited Distinction-646 [40] dataset and real-world PPM-100 [25]. Moreover, as shown in Table 4, our SDNet has the smallest number of parameters (only 2.0 M, 2.2% of parameters of MGM, and 1.5% of that of Matteformer) among both trimap-based and trimap-free models. The above comparative results demonstrate the effectiveness of our lightweight network design and the self-distillation training baseline in both composition and real-world scenes. Further, Our SDNet is lightweight and highly efficient on the portrait matting task and can be considered a novel baseline model that applied to mobile devices. Some representative visualizations on composited benchmarks and real-world PPM-100 are provided in Figure 6 and 7, respectively. The visual comparisons further demonstrate the effectiveness of our lightweight model design and the self-distillation training pipeline.

Models	Trimap	SAD↓	MSE↓
SDNet w/o AMPA	✓	3.1	0.0004
Full SDNet	✓	**2.7**	**0.0003**
SDNet w/o AMPA		9.2	0.0036
Full SDNet		**6.2**	**0.0020**

Table 6. Ablation of the AMPA module.

5. Ablation study

5.1. Effectiveness of different distillation settings

We conduct an ablation study on the portrait subset of the Composition-1K to compare the effectiveness of three distillation methods mentioned in Section3.4. As shown in Table5, although all three distillation methods significantly improved the performance of our SDNet, the best teacher distillation method slightly outperformed the other two in terms of all evaluation metrics. Given the superior performance of the best teacher distillation method, we choose it as the distillation approach in our final model.

5.2. Effectiveness of AMPA module

In order to assess the contribution of the Adaptive Multi-Perspective Aggregation (AMPA) module to the overall performance of our model, we conducted an ablation study by comparing the performance of our SDNet with or w/o the AMPA module. As illustrated in Table 6, the performance of our model w/o AMPA decreased significantly compared to the version with AMPA. The degradation in performance is particularly noticeable for trimap-free methods. We observe that the AMPA module plays a crucial role in capturing multi-level features that cover both global semantic integrity and detail attributes, thanks to its receptive diversity on feature representations by adaptive multi-receptive fields, especially for the trimap-free model versions.

6. Conclusion

In this paper, we propose an Extremely Efficient Portrait Matting Model via Self-Distillation (SDNet), which aims to perform accurate and effective portrait matting with limited resources, such as mobile devices. Our SDNet has only 2M parameters. We introduce an Adaptive Multi-scale Pyramid Attention (AMPA) module that fuses multiple receptive fields in one layer to enhance the diversity of reception in feature representations that cover both global semantic integrity and detail attributes. To improve the performance of our lightweight baseline, we introduce the training pipeline of self-distillation without any parameter addition, network modification, or over-parameterized teacher models from traditional distillation methods. Extensive experiments demonstrate the effectiveness of our self-distillation training pipeline and the lightweight model design.

References

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

[2] Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. Disentangled image matting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8819–8828, 2019. 2, 3

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

[4] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 618–626, 2018. 2, 3, 7

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

[6] Xiaowu Chen, Dongqing Zou, Steven Zhiying Zhou, Qinping Zhao, and Ping Tan. Image matting with local and nonlocal smooth priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1902–1907, 2013. 2

[7] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3435–3444, 2019. 4

[8] Ming-Ming Cheng, Shang-Hua Gao, Ali Borji, Yong-Qiang Tan, Zheng Lin, and Meng Wang. A highly efficient model to study the semantics of salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8006–8021, 2021. 2, 3, 4

[9] Donghyeon Cho, Yu-Wing Tai, and Inso Kweon. Natural image matting using deep convolutional neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 626–643. Springer, 2016. 2

[10] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019. 2

[11] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 4

[12] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1911–1920, 2019. 4

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

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

[15] Marco Forte and François Pitié. f, b, alpha matting. *arXiv preprint arXiv:2003.07711*, 2020. 6, 8

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

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[18] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4130–4139, 2019. 1, 2, 3

[19] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 1, 2

[20] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[21] Wang Jue and Cohen Michael F. *Image and video matting: a survey*. Now Publishers Inc, 2008. 2

[22] He Kaiming, Rhemann Christoph, Rother Carsten, Tang Xiaoou, and Sun Jian. A global sampling method for alpha matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2049–2056. IEEE, 2011. 2

[23] Minsoo Kang, Jonghwan Mun, and Bohyung Han. Towards oracle knowledge distillation with neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4404–4411, 2020. 2

[24] Zhanghan Ke, Kaican Li, Yurou Zhou, Qiuhua Wu, Xiangyu Mao, Qiong Yan, and Rynson WH Lau. Is a green screen really necessary for real-time portrait matting? *arXiv preprint arXiv:2011.11961*, 2020. 3

[25] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W.H. Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *AAAI*, 2022. 6, 7, 8

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[27] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11450–11457, 2020. 1, 7

[28] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 3

[29] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 238–247, 2022. 3

[30] 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 *Proceedings of the European conference on computer vision (ECCV)*, pages 740–755. Springer, 2014. 6

[31] Ruiping Liu, Kailun Yang, Alina Roitberg, Jiaming Zhang, Kunyu Peng, Huayao Liu, and Rainer Stiefelhagen. Transkd: Transformer knowledge distillation for efficient semantic segmentation. *arXiv preprint arXiv:2202.13393*, 2022. 1

[32] Yuhao Liu, Jiake Xie, Xiao Shi, Yu Qiao, Yujie Huang, Yong Tang, and Xin Yang. Tripartite information mining and integration for image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7555–7564, 2021. 1

[33] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3266–3275, 2019. 1, 6, 8

[34] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 7

[35] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. Alphagan: Generative adversarial networks for natural image matting. In *British Machine Vision Conference (BMVC)*, page 259. BMVA Press, 2018. 2, 3

[36] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020. 2

[37] GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, and Nojun Kwak. Matteformer: Transformer-based image matting via prior-tokens. *arXiv preprint arXiv:2203.15662*, 2022. 7

[38] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4353–4361, 2017. 4

[39] Richard J Qian and M Ibrahim Sezan. Video background replacement without a blue screen. In *Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348)*, volume 4, pages 143–146. IEEE, 1999. 3

[40] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 6, 7, 8

[41] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13676–13685, 2020. 7

[42] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 2

[43] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2291–2300, 2020. 3

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[45] supervise.ly. Supervisely person dataset. 2018. 1, 6, 8

[46] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2970–2979, 2017. 1, 2, 3, 6, 7, 8

[47] Ting-Bing Xu and Cheng-Lin Liu. Data-distortion guided self-distillation for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5565–5572, 2019. 1, 4

[48] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. *arXiv preprint arXiv:2012.06722*, 2020. 7

[49] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3911, 2020. 1, 4

[50] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2021. 5

[51] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019. 1, 4

[52] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 2

[53] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7469–7478, 2019. 3, 7

[54] Bingke Zhu, Yingying Chen, Jinqiao Wang, Si Liu, Bo Zhang, and Ming Tang. Fast deep matting for portrait animation on mobile phone. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 297–305, 2017. 7