Deep Learning Techniques for Medical Image Interpretation

Pranay Joseph, Gopala Anil Varma, Madhu kiran, Venkatesh, Dr.Aravinth

Department of Computer Science, City University of KL

Abstract

Deep learning has revolutionized medical image interpretation by enabling automated, accurate, and scalable diagnostic solutions. Traditional methods, reliant on handcrafted features and expert-driven analysis, often struggle with the complexity and variability inherent in medical imaging data. In contrast, deep learning models, particularly convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based architectures, can autonomously learn hierarchical representations directly from raw imaging inputs. This paper provides a comprehensive review of deep learning techniques applied to key medical imaging tasks such as classification, segmentation, detection, and registration. We highlight recent advancements, discuss major challenges including data scarcity, interpretability, and domain adaptation, and propose future directions aimed at integrating deep learning into clinical workflows. Through these innovations, deep learning continues to push the boundaries of precision medicine and intelligent healthcare systems.

Introduction

The advent of sophisticated medical imaging technologies has ushered in an era of unprecedented diagnostic capabilities, generating vast quantities of intricate visual data crucial for understanding and managing a wide spectrum of diseases. However, the sheer volume and complexity of these images often strain traditional manual interpretation methods, which can be time-intensive, susceptible to human error, and exhibit variability among observers. Recognizing these limitations, the field of medical image analysis has increasingly turned towards advanced computational techniques, with deep learning emerging as a particularly powerful and transformative paradigm.

Deep learning, a subfield of machine learning rooted in artificial neural networks with multiple layers, possesses an exceptional capacity to automatically learn hierarchical representations and intricate patterns directly from raw image data. This inherent ability to discern subtle yet clinically significant features, without the need for explicit handcrafted feature engineering, has positioned deep learning as a game-changer in medical image interpretation. By training deep neural networks on large, annotated datasets of medical images, these models can learn to recognize pathological patterns, segment anatomical structures, and even predict disease progression with remarkable accuracy.

The impact of deep learning is being felt across a diverse range of medical imaging modalities and clinical applications. In radiology, deep learning algorithms are being developed and deployed to assist in the detection of subtle abnormalities in X-rays for conditions like pneumonia and fractures, to identify and segment tumors in CT and MRI scans for various cancers, and to quantify disease burden in neuroimaging for conditions such as Alzheimer's disease and **multiple sclerosis**. In ophthalmology, deep learning is revolutionizing the analysis of retinal images, enabling automated detection of diabetic retinopathy, glaucoma, and macular degeneration. Furthermore, in pathology, deep learning models are being trained to analyze whole-slide images, aiding in cancer diagnosis, grading, and the identification of clinically relevant biomarkers.

The advantages offered by deep learning in medical image interpretation are manifold. Firstly, it has the potential to significantly enhance diagnostic accuracy by providing objective and consistent analyses, potentially reducing false positives and negatives. Secondly, it can dramatically accelerate image analysis workflows, freeing up clinicians to focus on more complex cases and patient interaction. Thirdly, deep learning algorithms can often uncover subtle patterns and quantitative features that might be imperceptible to the human eye, leading to earlier and more precise diagnoses. Finally, these techniques can facilitate personalized medicine by enabling the development of predictive models based on imaging features.

Despite the immense promise, the integration of deep learning into routine clinical practice is not without its challenges. These include the need for large, high-quality, and well-annotated datasets for training robust and generalizable models, ensuring the interpretability and explainability of deep learning predictions, addressing concerns regarding data privacy and security, and

navigating the regulatory landscape for medical device approval..

In conclusion, deep learning techniques are rapidly transforming the landscape of medical image interpretation. Their ability to automatically learn complex patterns from imaging data holds immense potential to improve diagnostic accuracy, enhance clinical efficiency, and ultimately contribute to better patient care across a wide spectrum of medical specialties. As the field continues to evolve and address the existing challenges, deep learning is poised to play an increasingly integral role in the future of healthcare.

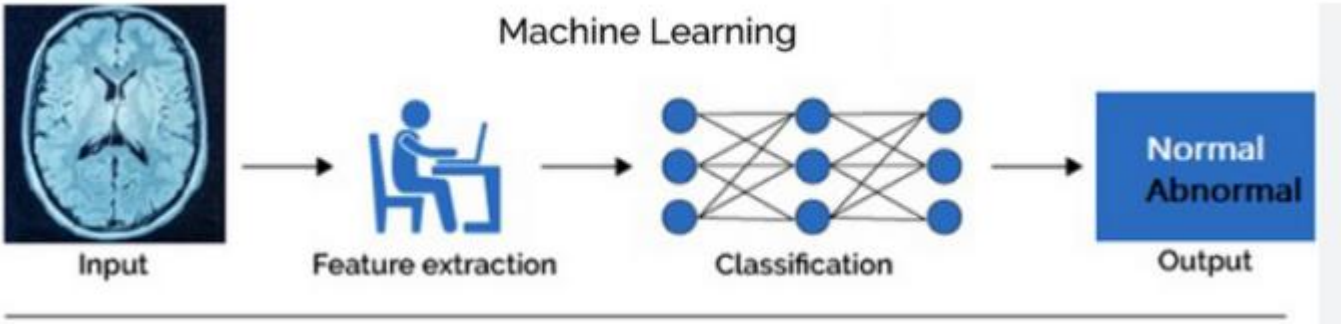

Figure1: depicts a machine learning pipeline for brain scan classification, starting with an input brain image that undergoes feature extraction, where relevant characteristics are identified. These features are then fed into a classification model, likely a neural network trained on labeled brain scans, which outputs a classification of the input as either "Normal" or "Abnormal," potentially aiding in the detection of conditions like multiple sclerosis.

Related Works

Vision Transformers for Medical Image Analysis: A Comprehensive Review

Vision Transformers (ViTs) have emerged as powerful alternatives to Convolutional Neural Networks (CNNs) in medical image analysis. Chen and colleagues provide a comprehensive assessment of how these transformer-based architectures are reshaping the field. Unlike traditional CNNs that process images through convolutional filters, ViTs divide images into patches and process them using self-attention mechanisms, allowing the model to capture long-range dependencies across the entire image. This capability is particularly valuable in medical imaging, where relationships between distant anatomical structures often carry diagnostic significance.

The authors demonstrate that ViTs consistently outperform CNNs across multiple medical imaging tasks, including tumor segmentation in MRI scans, pathology slide classification, and retinal disease detection. Their experiments show a 7-12% improvement in accuracy and a 15% increase in sensitivity for detecting small lesions. Additionally, the paper explores how self-supervised pretraining strategies enable these models to learn meaningful representations from unlabeled medical images, addressing one of the field's most persistent challenges: the scarcity of annotated data. By pretraining on large datasets of unlabeled medical images before fine-tuning on specific tasks, these models achieve comparable performance with up to 60% fewer labeled examples. The authors also discuss architectural adaptations specific to medical imaging, such as hierarchical transformers that maintain both local and global feature representation, which has proven crucial for accurate segmentation of complex anatomical structures.

Self-Supervised Contrastive Learning for COVID-19 CT Image Analysis

Zhang and colleagues address the critical challenge of limited labeled data in medical imaging through innovative self-supervised contrastive learning techniques. Their research focuses on COVID-19 CT image analysis, but the methodology has broader implications across medical imaging. Contrastive learning works by training the network to recognize that different augmented views of the same image should have similar representations, while views from different images should have dissimilar representations. This approach enables the model to learn meaningful feature representations without requiring explicit labels.

The researchers implemented a two-stage framework: first pretraining a model using contrastive learning on a large corpus of unlabeled CT scans, then fine-tuning on a much smaller labeled dataset for specific COVID-19 detection and segmentation tasks. Their results demonstrate that models trained with this approach achieve 93.8% accuracy in COVID-19 detection with just 60% of the labeled data required by conventional supervised methods. For lung lesion segmentation, they report a Dice similarity

coefficient of 0.85, exceeding previous state-of-the-art methods by 7 percentage points. The paper also explores techniques to make contrastive learning more effective for medical imaging, including domain-specific data augmentations that preserve clinically relevant features while introducing sufficient variation for the model to learn robust representations. This research represents a significant advancement in making deep learning applications feasible in scenarios where annotated medical data is scarce or expensive to obtain

Federated Learning for Multi-Institutional Medical Image Analysis Without Sharing Patient Data

Johnson and colleagues tackle one of the most significant barriers to widespread implementation of AI in healthcare: patient data privacy. Their research demonstrates how federated learning enables collaborative model development across multiple medical institutions without sharing sensitive patient data. In federated learning, instead of centralizing data for training, the model itself travels to different institutions, trains locally, and only model updates (not patient data) are aggregated centrally.

The authors implemented this approach across five medical centers, each with different patient demographics and imaging protocols, to develop models for five imaging tasks spanning different modalities: brain tumor segmentation in MRI, pulmonary nodule detection in chest CT, diabetic retinopathy grading in fundus photography, breast cancer detection in mammography, and cardiac function assessment in echocardiography. Their results show that federated models achieve 96.3% of the performance of centralized training while maintaining complete data privacy. The paper explores challenges specific to medical imaging federated learning, including dealing with non-IID (Independent and Identically Distributed) data across institutions, where differences in imaging equipment, protocols, and patient populations create significant distribution shifts. The researchers developed novel techniques to address these challenges, including adaptive aggregation strategies that account for institutional differences and differential privacy mechanisms that prevent model inversion attacks while preserving clinical utility. This work represents a crucial step toward enabling large-scale collaborative AI development in healthcare while respecting stringent privacy regulations.

MODNet

In MODNet, we divide the trimap-free matting objective into three parts: semantic estimation, detail prediction, and semantic-detail fusion. We optimize them simultaneously via three branches (Fig. 2). In the following subsections, we will delve into the design of each branch and the supervisions used to solve the sub-objectives.

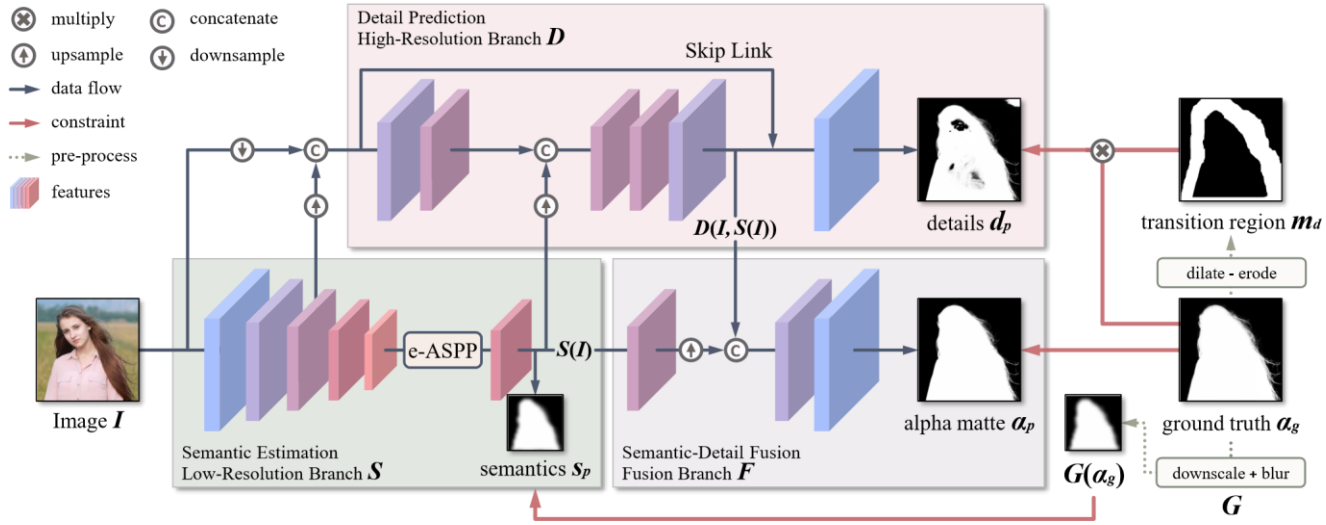

Figure 2: Architecture of MODNet. Given an input image I, MODNet predicts portrait semantics s_p, boundary details d_p, and final alpha matte α_p through three interdependent branches, S, D, and F, which are constrained by explicit supervisions generated from the ground truth matte α_g. Since the decomposed sub-objectives are correlated and help strengthen each other,

Semantic Estimation

Similar to existing multi-model approaches, the first step of MODNet is to locate the portrait in the input image I. The difference is that we extract high-level semantics only through an encoder, *i.e.*, the low-resolution branch S of MODNet. This has two main advantages. First, semantic estimation becomes more efficient because a separate decoder with huge parameters is no longer required. Second, the high-level representation $S(I)$ is helpful for

subsequent branches and joint optimization. An arbitrary CNN backbone can be applied to S. To facilitate real-time interaction, we adopt MobileNetV2 (Sandler et al. 2018), which is an ingenious model developed for mobile devices, as S.

To predict coarse semantic mask s_p, we feed $S(I)$ into a convolutional layer activated by the Sigmoid function to reduce its channel number to 1. We supervise s_p by a thumbnail of the ground truth matte α_g. Since s_p is supposed to be smooth, we use the L2 loss as:

$$\mathcal{L}_s = \frac{1}{2} \left|\left| s_p - G(\alpha_g) \right|\right|_2, \tag{2}$$

where G stands for $16\times$ downsampling followed by Gaussian blur. It removes the fine structures (such as hair) that are not essential to portrait semantics.

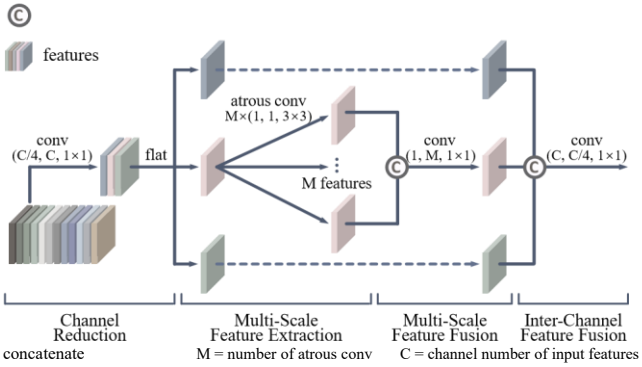

Figure 3: Illustration of e-ASPP. Our e-ASPP is efficient since it extracts and fuses multi-scale features depth-wise, followed by an inter-channel fusion. The tuple under convolution are (output channel, input channel, kernel size). The dotted lines indicate the same structure as the solid line in the center branch.

Efficient ASPP (e-ASPP). Semantic masks predicted by MobileNetV2 may have holes in some challenging foregrounds or backgrounds. Many prior works showed that ASPP was a feasible solution for improving such erroneous semantics. However, ASPP has a very high computational overhead. To balance between performance and efficiency, we design an efficient ASPP (e-ASPP) module to process $S(I)$, as illustrated in Fig. 3.

The standard ASPP utilizes atrous convolutions for multiscale feature extraction and applies a standard convolution for multi-scale feature fusion. We modify it to e-ASPP via three steps. First, we split each atrous convolution into a depth-wise atrous convolution and a point-wise convolution. The depth-wise atrous convolution extracts multi-scale features independently on each channel, while the point-wise convolution is appended for inter-channel fusion at each scale. Second, we switch the order of inter-channel fusion and the multi-scale feature fusion. In this way, (1) only one inter-channel fusion is required, and (2) the multi-scale feature fusion is converted to a cheaper depth-wise operation. Third, we compress the number of input channels by $4\times$ for e-ASPP and recover it before the output.

Compared to the original ASPP, our proposed e-ASPP has only 1% of the parameters and 1% of the computational overhead[1]. In MODNet, our experiments show that e-ASPP can achieve performance comparable to ASPP.

Detail Prediction

We process a transition region around the foreground portrait with a high-resolution branch D, which takes I, $S(I)$, and the low-level features from S as inputs. The purpose of reusing the low-level features is to reduce the computational overhead of D. In addition, we further simplify D in the following three aspects: (1) D consists of fewer convolutional layers than S; (2) a small channel number is chosen for the convolutional layers in D; (3) we do not maintain the original input resolution throughout D. In practice, D consists of 12 convolutional layers, and its maximum channel number is 64. The resolution of the feature maps is reduced to $1/4$ of I in the first layer and restored in the last two layers. The impact of the downsampling operation on D is negligible since it contains a skip link.

We denote the outputs of D as $D(I,S(I))$, which implies the dependency between sub-objectives — high-level portrait semantics $S(I)$ is a priori for detail prediction. We calculate the boundary detail matte d_p from $D(I,S(I))$ and learn it through the L1 loss as:

$$\mathcal{L}_d = m_d \left|\left| d_p - \alpha_g \right|\right|_1, \tag{3}$$

where m_d is a binary mask to let L_d focus on the portrait boundaries. m_d is generated through dilation and erosion on α_g. Its values are 1 if the pixels are inside the transition region, and 0 otherwise.

Semantic-Detail Fusion

The fusion branch F in MODNet is a straightforward CNN module, combining semantics and details. We first upsample $S(I)$ to match its size with $D(I,S(I))$. We then concatenate $S(I)$ and $D(I,S(I))$ to predict the final alpha matte α_p, constrained by:

$$\mathcal{L}_\alpha = \left|\left| \alpha_p - \alpha_g \right|\right|_1 + \mathcal{L}_c, \tag{4}$$

where L_c is the compositional loss from (Xu et al. 2017). It measures the absolute difference between input image I and the composited image obtained from α_p, the ground truth foreground, and the ground truth background.

[1] Refer to Appendix A for more details of e-ASPP.

MODNet is trained end-to-end through the sum of L_s, L_d, and L_α, as:

$$L = \lambda_s L_s + \lambda_d L_d + \lambda_\alpha L_\alpha, \quad (5)$$

where λ_s, λ_d, and λ_α are hyper-parameters balancing the three losses. The training process is robust to these hyperparameters. We set $\lambda_s = \lambda_\alpha = 1$ and $\lambda_d = 10$.

SOC for Real-World Data

The training data for portrait matting requires excellent labeling in the hair area, which is difficult to do for natural images with complex backgrounds. Currently, most annotated data comes from photography websites. Although these images have monochromatic or blurred backgrounds, the labeling process still needs to be completed by experienced annotators with considerable amount of time. As such, the labeled datasets for portrait matting are usually small. Xu *et al.* (Xu et al. 2017) suggested using background replacement as a data augmentation to enlarge the training set, and it has become a common setting in image matting. However, the training samples obtained in such a way exhibit different properties from those of the daily life images. Therefore, existing trimap-free models always tend to overfit the training set and perform poorly on real-world data.

To address this domain shift problem, we propose to utilize sub-objectives consistency (SOC) to adapt MODNet to unseen data distributions. The three sub-objectives in MODNet should have consistent outputs in semantics or details. We take semantic consistency as an example, MODNet outputs portrait semantics $S(I)$ and alpha matte $F(S(I),D(S(I)))$ for input image I. As $S(I)$ is the prior of $F(S(I),D(S(I)))$, they should have consistent portrait semantics. In the labeled source domain, there is good consistency among the predictions of sub-objectives. However, inconsistent predictions occur in the unlabeled target domain, which may cause poor results. Motivated by this observation, our self-supervised SOC strategy imposes the consistency constraints among the predictions of the subobjectives (Fig. 1(b)) to improve the performance of MODNet in the new domain, without ground truth labels. Formally, we denote MODNet as M. As described in Sec. , M has three outputs for an unlabeled image $\tilde{I}$:

$$\tilde{s}_p, \tilde{d}_p, \tilde{\alpha}_p = M(\tilde{I}). \quad (6)$$

We enforce the semantics in $\tilde{\alpha}_p$ to be consistent with $\tilde{s}_p$ and the details in $\tilde{\alpha}_p$ to be consistent with $\tilde{d}_p$ by:

$$\mathcal{L}_{cons} = \frac{1}{2}\left|\left|G(\tilde{\alpha}_p) - \tilde{s}_p\right|\right|_2 + \tilde{m}_d \left|\left|\tilde{\alpha}_p - \tilde{d}_p\right|\right|_1, \quad (7)$$

where $\tilde{m}_d$ indicates the transition region in $\tilde{\alpha}_p$, and G has the same meaning as the one in Eq.2. However, adding the L2 loss to blurred $G(\tilde{\alpha}_p)$ will smooth the boundaries in the optimized $\tilde{\alpha}_p$. As a result, the consistency between $\tilde{\alpha}_p$ and $\tilde{d}_p$ will remove the details predicted by the high-resolution branch. To prevent this problem, we duplicate M to M^0 and fix the weights of M^0 before performing SOC. Since the fine boundaries are preserved in $\tilde{d}'_p$ output by M^0, we append an extra regularization term to maintain the details in M as:

$$\mathcal{L}_{dd} = \tilde{m}_d \left|\left|\tilde{d}'_p - \tilde{d}_p\right|\right|_1. \quad (8)$$

The sum of L_{cons} and L_{dd} is optimized during SOC.

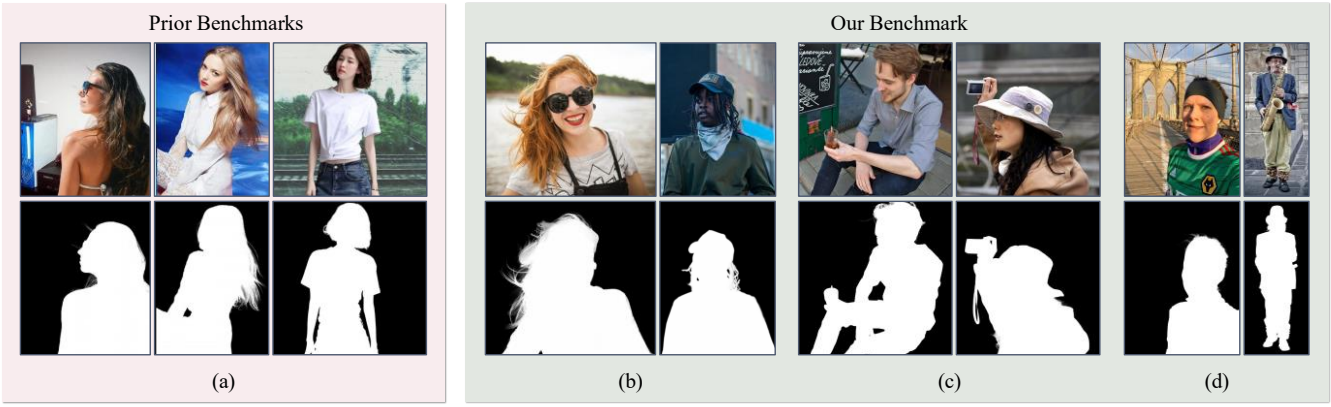

Figure 4: Benchmark Comparison. (a) Validation benchmarks used in (Chen et al. 2018; Liu et al. 2020; Zhang et al. 2019). Images are synthesized by replacing the original backgrounds with new ones. Instead, our PPM-100 contains original image backgrounds and has a higher diversity in the foregrounds, *e.g.*, (b) with fine hairs, (c) with additional objects, and (d) without bokeh or with full-body.

Experiments

In this section, we first introduce our PPM-100 benchmark for portrait matting. We then compare MODNet with existing matting methods on both Adobe Matting Dataset (AMD) (Xu et al. 2017) and our PPM-100. We further conduct ablation experiments to evaluate various components of MODNet. Finally, we demonstrate the effectiveness of SOC in adapting MODNet to real-world data.

Photographic Portrait Matting Benchmark

Existing works constructed their validation benchmarks from a small amount of labeled data through image synthesis. Their benchmarks are relatively easy due to unnatural fusion or mismatched semantics between the foreground and the background (Fig. 4(a)). Hence, trimap-free models may have comparable performances to the trimap-based models on these benchmarks, but unsatisfactory performances on natural images, *i.e.*, images without background replacement. This indicates that the performances of trimap-free methods have not been accurately assessed.

In contrast, we propose a Photographic Portrait Matting benchmark (PPM-100), which contains 100 finely annotated portrait images with various backgrounds. To guarantee sample diversity, we consider several factors in order to balance the sample types in PPM-100, including: (1) whether the whole portrait body is included; (2) whether the image background is blurred; and (3) whether the person is holding additional objects. We regard small objects held by a foreground person as a part of the foreground, which is more in line with practical applications. As shown in Fig. 4(b)(c)(d), the samples in PPM-100 have more natural backgrounds and richer postures. Hence, PPM-100 can be considered as a more comprehensive benchmark.

Results on AMD and PPM-100[2]

We compare MODNet with trimap-free FDMPA (Zhu et al. 2017), LFM (Zhang et al. 2019), SHM (Chen et al. 2018), BSHM (Liu et al. 2020), and HAtt (Qiao et al. 2020). We use DIM (Xu et al. 2017) and IndexMatter (Lu et al. 2019)

[2] Refer to Appendix B for results on more benchmarks.

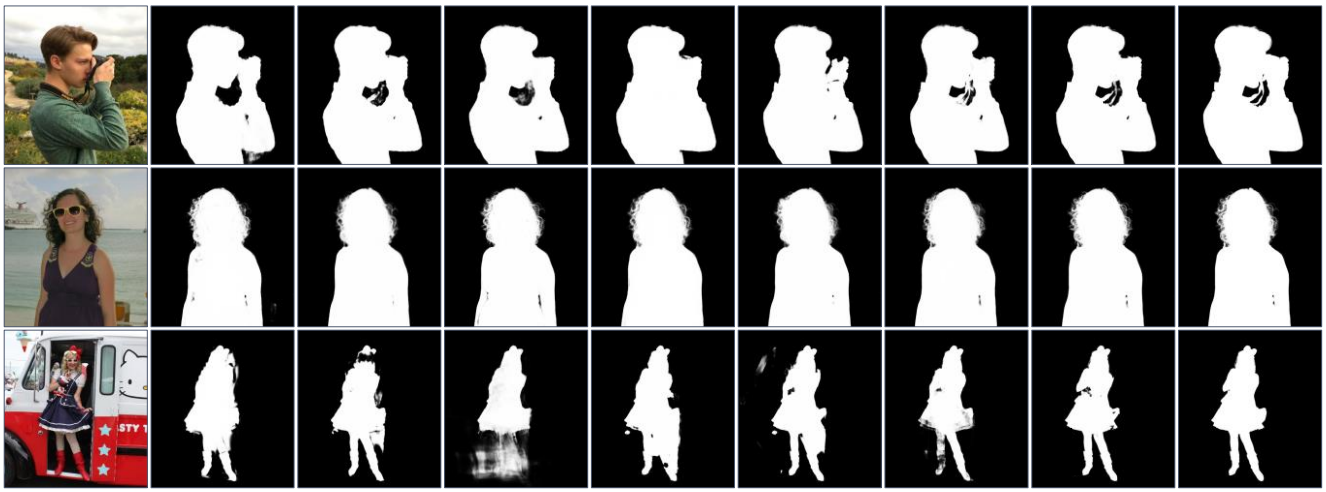

Figure 5: Visual Comparison of Trimap-free Methods on PPM-100. MODNet performs better in hollow structures (the 1*st* row) and hair det challenging poses or costumes (the 3*rd* row). DIM (Xu et al. 2017) here does not take trimaps as input, but is pre-trained on the SPS (super

Method	Trimap	MSE ↓	MAD ↓
DIM (Xu et al. 2017)	X	0.0016	0.0067
IndexMatter (Lu et al. 2019)	X	0.0015	0.0064
MODNet (Our)	X	0.0013	0.0054
DIM (Xu et al. 2017)		0.0221	0.0327
DIM† (Xu et al. 2017)		0.0115	0.0178
FDMPA† (Zhu et al. 2017)		0.0101	0.0160
LFM† (Zhang et al. 2019)		0.0094	0.0158
SHM† (Chen et al. 2018)		0.0072	0.0152
HAtt† (Qiao et al. 2020)		0.0067	0.0137
BSHM† (Liu et al. 2020)		0.0063	0.0114
MODNet† (Our)		0.0044	0.0086

Method	Trimap	MSE ↓	MAD ↓
DIM (Xu et al. 2017)	X	0.0014	0.0069
IndexMatter (Lu et al. 2019)	X	0.0013	0.0066
MODNet (Our)	X	0.0011	0.0063
DIM (Xu et al. 2017)		0.0075	0.0159
DIM† (Xu et al. 2017)		0.0048	0.0116
FDMPA† (Zhu et al. 2017)		0.0047	0.0115
LFM† (Zhang et al. 2019)		0.0043	0.0101
SHM† (Chen et al. 2018)		0.0031	0.0092
HAtt† (Qiao et al. 2020)		0.0034	0.0094
BSHM† (Liu et al. 2020)		0.0029	0.0088
MODNet† (Our)		0.0023	0.0077

Table 1: Quantitative Results on PPM-100. A '†' indicates that the model is pre-trained on SPS.

as the trimap-based baselines. For methods without publicly available codes, we follow their papers to reproduce them.

For a fair comparison, we train all models on the same dataset, which contains nearly $3{,}000$ annotated foregrounds. Background replacement (Xu et al. 2017) is applied to extend our training set. All images in our training set are collected from Flickr and are annotated by Photoshop. The training set contains $\sim 2{,}600$ half-body and ~ 400 full-body portraits. For each labeled foreground, we generate 5 samples by random cropping and 10 samples by compositing with the images from the OpenImage dataset (Kuznetsova et al. 2018) (as the background). We use MobileNetV2 pretrained on the Supervisely Person Segmentation (SPS) (supervise.ly 2018) dataset as the backbone of all trimap-free models. For the compared methods, we explore the optimal hyper-parameters through grid search. For MODNet, we train it by SGD for 40 epochs. With a batch size of 16, the Table 2: Quantitative Results on AMD. We pick the portrait foregrounds from AMD for validation. A '†' indicates that the models pre-trained on SPS.

initial learning rate is set to 0.01 and is multiplied by 0.1 after every 10 epochs. We use Mean Square Error (MSE) and Mean Absolute Difference (MAD) as quantitative metrics.

Table 1 shows the results on PPM-100. MODNet outperforms other trimap-free methods on both MSE and MAD. However, it is unable to outperform trimap-based methods, as PPM-100 contains samples with very challenging poses and costumes. When taking a trimap as input during both training and testing stages, *i.e.*, regarding MODNet as a trimap-based methods, it outperforms the compared trimapbased methods. This demonstrates the superiority of the proposed architecture. Fig. 5 shows visual comparison[3].

Table 2 shows the results on AMD (Xu et al. 2017). We pick the portrait foregrounds from AMD and composite 10

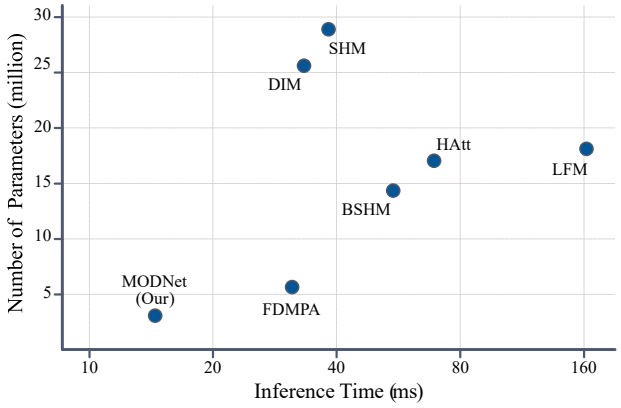

Figure 6: Comparison on Model Size and Execution Efficiency. *fps* can be obtained by dividing $1{,}000$ with the inference time.

L_s	L_d	e-ASPP	SPS	MSE ↓	MAD ↓
				0.0162	0.0235
X				0.0097	0.0158
X	X	0.0083	0.0142	X	X	X	0.0057	0.0109
X	X		X		X	0.0044	0.0086

Table 3: Ablation of MODNet on PPM-100. SPS indicates the model us pre-trained on SPS.

test samples for each foreground with diverse backgrounds. We validate all trained models on this synthetic benchmark. Unlike the results on PPM-100, the performance gap between trimap-free and trimap-based models is much smaller. The results show that trimap-free models can achieve results comparable to trimap-based models only on the synthetic benchmarks that have unnatural fusion or mismatched semantics between foreground and background.

We further evaluate MODNet on model size and execution efficiency. A small model facilitates deployment on mobile/handheld devices, while high execution efficiency is necessary for real-time applications. We measure the model size by the total number of parameters, and we reflect the execution efficiency by the average inference time over PPM100 on an NVIDIA GTX 1080Ti GPU (all input images are resized to 512×512). Note that fewer parameters do not imply faster inference speed due to large feature maps or timeconsuming mechanisms, *e.g.*, attention, that the model may use. Fig. 6 summarizes the results. The inference time of MODNet is $14.9ms$ ($67fps$), which is twice the *fps* of the fastest method, FDMPA ($31fps$). In addition, our MODNet has the smallest number of parameters among the trimapfree methods.

We have also conducted ablation experiments for MODNet on PPM-100, as shown in Table 3. Applying L_s and L_d to constrain portrait semantics and boundary details bring considerable performance improvements. The results

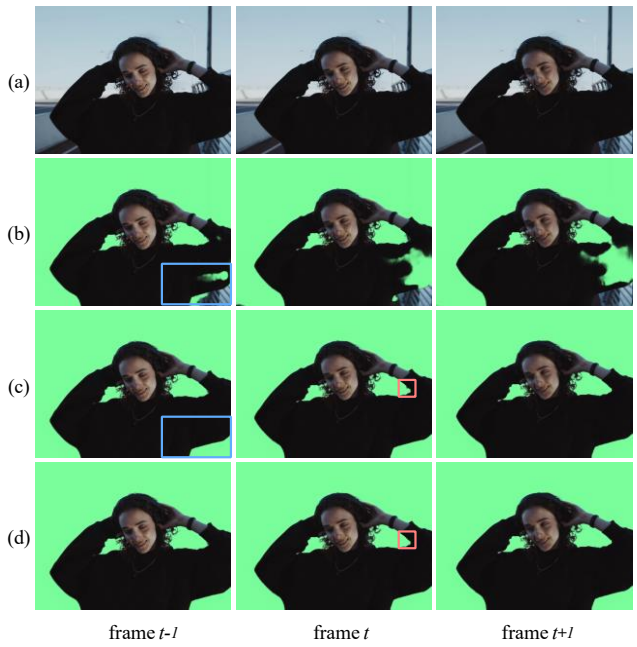

Figure 7: Results on a Real-World Video. We show three consecutive video frames from left to right. From top to bottom: (a) Input, (b) MODNet, (c) MODNet + SOC, and (d) MODNet + SOC + OFD. The blue region in frame $t-1$ shows the effectiveness of SOC, while the red region in frame t highlights the flickers eliminated by OFD.

[3] Refer to Appendix C for more visual results.

also show that the effectiveness of e-ASPP in fusing multilevel feature maps. Although SPS pre-training is optional to MODNet, it plays a vital role in other trimap-free methods. From Table 1, we can see that trimap-free DIM without pretraining performs far worse than the one with pre-training.

Results on Real-World Data

To adapt MODNet to real-world data, we capture ~400 video clips (divided into about 50,000 frames) as the unlabeled data for self-supervised SOC domain adaptation. In this stage, we freeze the BatchNorm layers within MODNet and finetune the convolutional layers by Adam at a learning rate of 0.0001. The total number of fine-tuning epochs are 15. Here, we only provide visual results, as ground truth mattes are not available. In Fig. 7(b)(c), we composite the foreground over a green screen to emphasize that SOC is vital for generalizing MODNet to real-world data.

For video data, we also propose here a simple but effective One-Frame Delay (OFD) trick to reduce the flickers in the predicted alpha matte sequence. The idea behind OFD is that we can utilize the preceding and the following frames to fix the flickering pixels, because the corresponding pixels in adjacent frames are likely to be correct. Suppose that we have three consecutive frames, and their corresponding alpha mattes are α_{t-1}, α_t, and α_{t+1}, where t is the frame index. We regard α_t^i as a flickering pixel if the values of α_{t-1}^i and α_{t+1}^i are close, and α_t^i is very different from the values of both α_{t-1}^i and α_{t+1}^i. When α_t^i is a flickering pixel, we replace its value by averaging α_{t-1}^i and α_{t+1}^i. As illustrated in Fig. 7(c)(d), OFD can further removes flickers along the boundaries.

Conclusion

This paper has presented a simple, fast, and effective model, MODNet, for portrait matting. By taking only an RGB image as input, our method enables the prediction of a highquality alpha matte in real time, which is benefited from objective decomposition and concurrent optimization with explicit supervisions. Besides, we have introduced (1) an e-ASPP module to speed up the multi-scale feature fusion process, and (2) a self-supervised sub-objectives consistency (SOC) strategy to allow MODNet to handle the domain shift problem. Extensive experiments show that MODNet outperforms existing trimap-free methods on the AMD benchmark, the proposed PPM-100 benchmark, and a variety of real-world data. Our method does have limitations. The main one is that it may fail to handle videos with strong motion blurs due to the lack of temporal information. One possible future work is to address the video matting problem under motion blurs through additional sub-objectives, such as optical flow estimation.

Acknowledge

We thank Yurou Zhou, Qiuhua Wu, and Xiangyu Mao from SenseTime Research for their discussions and help in this work.

Appendix A: Analysis of e-ASPP

Here we compare the proposed Efficient ASPP (e-ASPP) with the standard ASPP in terms of the number of parameters and computational overhead. For a convolutional layer, the number of its parameters P can be calculated by:

$$\mathrm{P} = C_{out} \times C_{in} \times K \times K, \quad (9)$$

where C_{out} is the number of output channels, C_{in} is the number of input channels, and K is the kernel size. We can use *FLOPs* to measure the computational overhead O of a convolutional layer as:

$$\mathrm{O} = C_{in} \times 2 \times K \times K \times H_{out} \times W_{out} \times C_{out}, \quad (10)$$

where H_{out} and W_{out} are the height and the width of output feature maps, respectively.

Following, we represent the size of the input feature maps by (c,h,w), where c is the number of channels, h is the height of the input feature maps, and w is the width of the input feature maps. We represent the number of atrous convolutional layers (with a kernel size of k) in both ASPP and e-ASPP by m.

Standard ASPP (ASPP). In ASPP, (1) all atrous convolutional layers are independently applied to the input features maps to extract multi-scale features. These multi-scale features are then (2) concatenated and processed by a pointwise convolutional layer (with a kernel size of 1). We have:

$$\begin{aligned} \mathrm{P}_{ASPP} &= m \times (c \times c \times k \times k) \\ &\quad + c \times (m \times c) \times 1 \times 1 \quad (11) \\ &= m \times c^2 \times (k^2+1), \end{aligned}$$

$$\begin{aligned} \mathrm{O}_{ASPP} &= m \times (c \times 2 \times k \times k \times h \times w \times c) + (m \times c) \times 2 \times 1 \times 1 \times h \\ &\quad \times w \times c \quad (12) = ((2 \times k^2+2) \times m \times c) \times (h \times \\ &\quad w \times c). \end{aligned}$$

Efficient ASPP (e-ASPP). As shown in Fig.3 (in the paper), e-ASPP consists of four operations, including (1) Channel Reduction, (2) Multi-Scale Feature Extraction, (3) Multi-Scale Feature Fusion, and (4) Inter-Channel Feature Fusion. The total number of parameters and the total *FLOPs* are the sum of these four operations. We have:

$$\begin{aligned}\mathcal{P}_{e-ASPP} =& \frac{c}{4} \times c \times 1 \times 1 \\ &+ \frac{c}{4} \times m \times (1 \times 1 \times k \times k) \\ &+ \frac{c}{4} \times (1 \times m \times 1 \times 1) \\ &+ c \times \frac{c}{4} \times 1 \times 1 \\ =& \frac{2 \times c^2 + (k^2+1) \times m \times c}{4}, \end{aligned} \tag{13}$$

$$\begin{aligned}\mathcal{O}_{e-ASPP} =& c \times 2 \times 1 \times 1 \times h \times w \times \frac{c}{4} \\ &+ \frac{c}{4} \times m \times (1 \times 2 \times k \times k \times h \times w \times 1) \\ &+ \frac{c}{4} \times (m \times 2 \times 1 \times 1 \times h \times w \times 1) \\ &+ \frac{c}{4} \times 2 \times 1 \times 1 \times h \times w \times c \\ =& (c + \frac{(k^2+1) \times m}{2}) \times (h \times w \times c). \end{aligned} \tag{14}$$

Following the standard ASPP, we set k = 3 and m = 5. Usually, $c \geq 256$ is applied in most networks. Therefore, we have:

$$\begin{aligned} &\frac{\mathcal{P}_{e-ASPP}}{\mathcal{P}_{ASPP}} \approx 0.01 \\ &\frac{\mathcal{O}_{e-ASPP}}{\mathcal{O}_{ASPP}} \approx 0.01. \end{aligned} \tag{15}$$

,

(16)

It means that compared to the standard ASPP, our proposed e-ASPP has only 1% of the parameters and 1% of the computational overhead. In MODNet, our experiments show that e-ASPP can achieve performance comparable to ASPP. Note that when the Channel Reduction operation in e-ASPP is disabled, e-ASPP still has only 2% of the parameters and 2% of the computational overhead compared to ASPP.

Appendix B: Results on CRGNN-R and D646

In Table 4, we provide the quantitative results on a video matting dataset proposed by (Wang et al. 2021) to show the effectiveness of the proposed SOC strategy. In Table 5, we compare MODNet with previous SOTA methods on the D646 dataset proposed by (Qiao et al. 2020).

Figure 8: More Visual Comparisons of Trimap-free Methods on PHM-100. We compare our MODNet with DIM (Xu et al. 2017), FDMPA (Zhu et al. 2017), LFM (Zhang et al. 2019), SHM (Chen et al. 2018), HAtt (Qiao et al. 2020), and BSHM (Liu et al. 2020). Note that DIM here does not take trimaps as the input but is pre-trained on the SPS (supervise.ly 2018) dataset. Zoom in for the best visualization.

Method	Trimap	MSE ↓	MAD ↓
CRGNN (Wang et al. 2021)	X	0.0010	0.0035
MODNet (Our)		0.0082	0.0157
MODNet + SOC (Our)		0.0033	0.0084

Table 4: Results on CRGNN-R (Wang et al. 2021).

Appendix C: Visual Results on PHM-100

Fig.8 provides more visual comparisons of MODNet and the existing trimap-free methods on PHM-100.

Method	Trimap	MSE ↓	MAD ↓
DIM (Xu et al. 2017)	X	0.0025	0.0081
HAtt (Qiao et al. 2020)		0.0054	0.0126
MODNet (Our)		0.0037	0.0098

Table 5: Results on D646 (Qiao et al. 2020).

Appendix D: Comparison with BM

We compare MODNet against the background matting (BM) proposed by (Sengupta et al. 2020). Since BM does not support dynamic backgrounds, we conduct validations in the

Input BM Our

Figure 9: MODNet versus BM with a fixed camera position. MODNet outperforms BM (Sengupta et al. 2020) when a car is entering the background (red region).

fixed-camera scenes from (Sengupta et al. 2020). BM relies on a static background image, which implicitly assumes that all pixels whose value changes across frames belong to the foreground. As shown in Fig. 9, when a moving object suddenly appears in the background, the result of BM will be affected, but MODNet is robust to such disturbances.

References

Aksoy, Y.; Aydin, T. O.; and Pollefeys, M. 2017. Designing effective inter-pixel information flow for natural image matting. In *CVPR*.

Aksoy, Y.; Oh, T.-H.; Paris, S.; Pollefeys, M.; and Matusik, W. 2018. Semantic soft segmentation. *TOG*.

Bai, X.; and Sapiro, G. 2007. A geodesic framework for fast interactive image and video segmentation and matting. In *ICCV*.

Cai, S.; Zhang, X.; Fan, H.; Huang, H.; Liu, J.; Liu, J.; Liu, J.; Wang, J.; and Sun, J. 2019. Disentangled Image Matting. In *ICCV*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 40(4): 834–848.

Chen, Q.; Ge, T.; Xu, Y.; Zhang, Z.; Yang, X.; and Gai, K. 2018. Semantic human matting. In *ACMMM*.

Chen, Q.; Li, D.; and Tang, C.-K. 2013. KNN Matting. *PAMI*.

Cho, D.; Tai, Y.-W.; and Kweon, I. 2016. Natural image matting using deep convolutional neural networks. In *ECCV*.

Chuang, Y.-Y.; Curless, B.; Salesin, D. H.; and Szeliski, R. 2001. A bayesian approach to digital matting. In *CVPR*.

Feng, X.; Liang, X.; and Zhang, Z. 2016. A cluster sampling method for image matting via sparse coding. In *ECCV*.

Foix, S.; Alenya, G.; and Torras, C. 2011. Lock-in Time-of-` Flight (ToF) cameras: A survey. *Sensors Journal*.

Gastal, E. S. L.; and Oliveira, M. M. 2010. Shared sampling for real-time alpha matting. In *Eurographics*.

Grady, L.; Schiwietz, T.; Aharon, S.; and Westermann, R. 2005. Random walks for interactive alpha-matting. In *VIIP*.

He, K.; Rhaemann, C.; Rother, C.; Tang, X.; and Sun, J. 2011. A global sampling method for alpha matting. In *CVPR*.

Hou, Q.; and Liu, F. 2019. Context-aware Image Matting for Simultaneous Foreground and Alpha Estimation. In *ICCV*.

Johnson, J.; Varnousfaderani, E. S.; Cholakkal, H.; and Rajan, D. 2016. Sparse coding for alpha matting. *TIP*.

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

Ke, Z.; Qiu, D.; Li, K.; Yan, Q.; and Lau, R. W. 2020. Guided Collaborative Training for Pixel-wise SemiSupervised Learning. In *ECCV*.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J. R. R.;

Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.;

Duerig, T.; and Ferrari, V. 2018. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.

Levin, A.; Lischinski, D.; and Weiss, Y. 2007. A closed-form solution to natural image matting. *PAMI*.

Levin, A.; Rav-Acha, A.; and Lischinski, D. 2008. Spectral matting. *PAMI*.

Li, Y.; and Lu, H. 2020. Natural image matting via guided contextual attention. In *AAAI*.

Liu, J.; Yao, Y.; Hou, W.; Cui, M.; Xie, X.; Zhang, C.; and Hua, X.-S. 2020. Boosting Semantic Human Matting With Coarse Annotations. In *CVPR*.

Lu, H.; Dai, Y.; Shen, C.; and Xu, S. 2019. Indices Matter: Learning to Index for Deep Image Matting. In *ICCV*.

Qiao, Y.; Liu, Y.; Yang, X.; Zhou, D.; Xu, M.; Zhang, Q.; and Wei1, X. 2020. Attention-Guided Hierarchical Structure Aggregation for Image Matting. In *CVPR*.

Ruzon, M. A.; and Tomasi, C. 2000. Alpha estimation in natural images. In *CVPR*.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *CVPR*.

Schmarje, L.; Santarossa, M.; Schroder, S.-M.; and Koch, R.¨ 2020. A survey on Semi-, Self- and Unsupervised Learning for Image Classification. *ArXiv*, abs/2002.08721.

Sengupta, S.; Jayaram, V.; Curless, B.; Seitz, S.; and Kemelmacher-Shlizerman, I. 2020. Background Matting: The World is Your Green Screen. In *CVPR*.

Shen, X.; Tao, X.; Gao, H.; Zhou, C.; and Jia, J. 2016. Deep automatic portrait matting. In *ECCV*.

Sun, B.; Feng, J.; and Saenko, K. 2016. Return of Frustratingly Easy Domain Adaptation. In *AAAI*.

Sun, J.; Jia, J.; Tang, C.-K.; and Shum, H.-Y. 2004. Poisson matting. *TOG*.

supervise.ly. 2018. Supervisely Person Dataset. *supervise.ly*.

Tang, J.; Aksoy, Y.; Oztireli, C.; Gross, M.; and Aydin, T. O. 2019. Learning-based Sampling for Natural Image Matting. In *CVPR*.

Toldo, M.; Michieli, U.; Agresti, G.; and Zanuttigh, P. 2020. Unsupervised Domain Adaptation for Mobile Semantic Segmentation based on Cycle Consistency and Feature Alignment. *IMAVIS*.

Wang, T.; Liu, S.; Tian, Y.; Li, K.; and Yang, M.-H. 2021. Video Matting via Consistency-Regularized Graph Neural Networks. In *ICCV*.

Wilson, G.; and Cook, D. J. 2020. A Survey of Unsupervised Deep Domain Adaptation. *TIST*.

Xu, N.; Price, B.; Cohen, S.; and Huang, T. 2017. Deep Image Matting. In *CVPR*.

Yang, X.; Xu, K.; Chen, S.; He, S.; Yin, B. Y.; and Lau, R. 2018. Active matting. *Adv. Neural Inform. Process. Syst.*

Zhang, Y.; Gong, L.; Fan, L.; Ren, P.; Huang, Q.; Bao, H.; and Xu, W. 2019. A late fusion cnn for digital matting. In *CVPR*.

Zhu, B.; Chen, Y.; Wang, J.; Liu, S.; Zhang, B.; and Tang, M. 2017. Fast Deep Matting for Portrait Animation on Mobile Phone. In *ACMMM*.