Learning Multiple Representations with Inconsistency-Guided Detail Regularization for Mask-Guided Matting

Weihao Jiang, Zhaozhi Xie, Yuxiang Lu, *Student Member, IEEE,* Longjie Qi, Jingyong Cai, Hiroyuki Uchiyama, Bin Chen, Yue Ding, *Member, IEEE,* and Hongtao Lu, *Member, IEEE*

***Abstract*—Mask-guided matting networks have achieved significant improvements and have shown great potential in practical applications in recent years. However, simply learning matting representation from synthetic and lack-of-real-world-diversity matting data, these approaches tend to overfit low-level details in wrong regions, lack generalization to objects with complex structures and real-world scenes such as shadows, as well as suffer from interference of background lines or textures. To address these challenges, in this paper, we propose a novel auxiliary learning framework for mask-guided matting models, incorporating three auxiliary tasks: semantic segmentation, edge detection, and background line detection besides matting, to learn different and effective representations from different types of data and annotations. Our framework and model introduce the following key aspects: (1) to learn real-world adaptive semantic representation for objects with diverse and complex structures under real-world scenes, we introduce extra semantic segmentation and edge detection tasks on more diverse real-world data with segmentation annotations; (2) to avoid overfitting on low-level details, we propose a module to utilize the inconsistency between learned segmentation and matting representations to regularize detail refinement; (3) we propose a novel background line detection task into our auxiliary learning framework, to suppress interference of background lines or textures. In addition, we propose a high-quality matting benchmark, Plant-Mat, to evaluate matting methods on complex structures. Extensively quantitative and qualitative results show that our approach outperforms state-of-the-art mask-guided methods.**

***Index Terms*—Detail regularization, background line detection, mask-guided matting, dense prediction.**

I. Introduction

IMAGE alpha matting is an important computer vision task that predicts an alpha matte representing the opacity of foreground objects to precisely cut them out in an image. It has many applications in computational photography and image or video processing, editing, and compositing [1]–[6]. Alpha matting tasks usually model the natural image I as a convex combination of a foreground image F and a background image B at each pixel i, as shown below:

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i, \alpha_i \in [0, 1], \quad (1)$$

where α_i is the value of the alpha matte at pixel i.

Weihao Jiang, Zhaozhi Xie, Yuxiang Lu, Longjie Qi, Yue Ding, and Hongtao Lu are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (email: jiangweihao@sjtu.edu.cn; xiezhzh@sjtu.edu.cn; luyuxiang_2018@sjtu.edu.cn; qilongjie@sjtu.edu.cn; dingyue@sjtu.edu.cn; htlu@sjtu.edu.cn)
Jingyong Cai, Hiroyuki Uchiyama, and Bin Chen are with the OPPO Japan Research Center, Japan

Fig. 1. Qualitative comparisons between MGMatting [7] and Ours. From left to right, the input image and a binary guidance mask, MGMatting, Ours.

The Eq. 1 is highly ill-posed [8]. Therefore, many traditional methods [8]–[14] and deep learning based methods [15]–[27] utilize trimaps as guidance to reduce the solution space. In recent years, methods like MGMatting [7] and IGF [28] propose mask-guided matting frameworks, which only need an easily obtained coarse mask instead of a complex trimap. Since fine matting data is labor-intensive in data selection and annotation, deep matting methods composite finely annotated foreground on various background images to train models for objects under diverse scenes. However, these training samples still lack real-world diversity. Although MGMatting has an elaborate network and uses a strong training data augmentation like Context-aware matting [29] did on composited data to adapt the real-world application, and promote the mask-guided

Fig. 2. Visualization of the inconsistency between matting and segmentation masks, which points out important low-level details.

task, it still suffers from a few problems. Due to composited data or hard-to-attain and lack-of-real-world-diversity data, MGMatting is hard to generalize to real-world scenes such as shadows shown in Row 1, Fig. 1 and complex real-world foreground structures such as elongated cactus spines shown in Row 3, Fig 1. Due to overfitting on low-level details, MGMatting refines low-level details in the wrong regions (SARI's textures on a woman's body) instead of the correct sparse hairs in Row 2, Fig. 1. Last but not least, MGMatting also struggles to suppress interference of background lines or textures such as the two red boxes in Row 3, Fig. 1.

To address these challenges for mask-guided matting, we propose a novel auxiliary learning framework to properly learn real-world adaptive semantic representation and background line aware representation from different types of data (composited and real-world) and annotations (matting, segmentation, background line), and propose an inconsistency-guided detail regularization module to regularize detail refinement.

First, to adapt to diverse and complex object structures in real world, we introduce a real-world adaptive semantic representation to mask-guided networks through auxiliary semantic segmentation learning on diverse real-world data. Although the composited matting data provides precise alpha mattes strictly following Eq. 1 and can train matting models with detailed predictions, the real-world data provides real scenes such as shadows that can not be provided by the composited data. As the real-world data with coarse segmentation masks is much easier to attain than data with fine matting alpha mattes, it can provide training data with diverse and complex objects under real-world scenes. Different from matting alpha mattes, semantic segmentation masks focus more on representing the high-level semantic foreground regions instead of low-level details. Therefore, instead of naively using the segmentation data to supervise matting heads, we set an extra segmentation head on proper high-level feature maps in our matting network to learn the high-level semantics of real-world objects. In addition, we also add an extra edge detection head on proper high-resolution feature maps to learn real-world object boundaries and semantic contours. In this way, our network learns real-world adaptive semantic representation Se for diverse and complex real-world objects, besides the matting representation Ma with richer low-level details learned from the matting task, and adapts better to diverse and complex structures and real-world scenes.

Second, we propose an inconsistency-guided detail regularization (IGDR) module to regularize detail refinement and avoid overfitting low-level details. Generally, matting tasks represent a foreground object with a fine and soft mask, while segmentation tasks represent semantic foregrounds with a binary mask. As shown in Fig. 2, it is easy to observe that inconsistent parts between a matting alpha matte and segmentation mask highlight objects' fine low-level details which should be refined for matting. Based on this observation, and inspired by spatial sampling or alignment works [30]–[32], we generate a semantic representation feature map Se from a matting representation feature map Ma by eliminating their spatial inconsistency using a spatial sampling process. Then Ma and Se will be fed to a matting head and a segmentation head, respectively, for different supervision tasks. Then we obtain the inconsistency map IN as $IN = Ma - Se$, which indicates low-level details in proper regions, and our IGDR module uses IN to guide and enhance low-level details of objects in proper regions, preventing overfitting in wrong regions like MGMatting [7] in Row 2, Fig. 1.

Third, we propose a novel background line detection task into our auxiliary learning framework, to suppress interference of background lines or textures. Mask-guided matting networks simply learning a matting representation suffer from interference of background details such as lines and textures like MGMatting [7] in Row 3, Fig. 1. Therefore, we incorporate a novel background line detection task into our auxiliary learning framework to learn discriminative representation to better distinguish foreground objects from background lines or textures. For training data, we generate distance maps with LSD [33] for background images through a homography adaptation [34] as pseudo ground truth, then the background images will be composited with matting foregrounds based on their alpha matte. For the model, we set a background line detection head on proper high-resolution feature maps in our network decoder, and supervise it with the distance maps of background lines adapted to corresponding composited samples, to learn a background line aware representation. In this way, our network learns an effective representation to suppress interference of background lines or textures.

In addition, we propose a high-quality matting benchmark, Plant-Mat, to evaluate mask-guided matting methods on complex object structures for academic research.

Our contributions can be summarized as follows:

- We propose a real-world adaptive semantic representation (RASR) learned through auxiliary semantic segmentation and edge detection tasks on real-world segmentation data, to adapt our network to diverse and complex object structures and real-world scenes.
- To overcome the overfitting on low-level details of mask-guided approaches, we propose a novel inconsistency-guided detail regularization (IGDR) module in our network to regularize low-level detail refinement.
- We propose a novel background line detection task into our auxiliary learning framework, to suppress interference of background lines or textures for matting.
- Quantitative and qualitative results on the RWP [7], AIM-500 [35], AM-2k [36], PPM-100 [37], and the proposed Plant-Mat benchmarks demonstrate that our approach outperforms SOTA mask-guided methods.

II. Related Work

Deep trimap-based and trimap-free matting. Since Adobe [2] develops a training method on large-scale synthetic matting datasets that can generate large and diverse composited training matting data with ground-truth alpha strictly following Eq. 1, both trimap-based [2], [16], [27], [38] and trimap-free [35]–[37], [39]–[42] deep matting methods have been promoted significantly on natural matting. However, trimap-based approaches rely on complex trimaps, while trimap-free approaches lack user interaction or auxiliary inputs, so their performances rely on the distribution of training matting data and can not be improved by user guidance. These problems limit the application of deep matting.

Mask-guided matting. To extend the application of deep matting, MGMatting [7] and IGF [28] use accessible coarse guidance masks as auxiliary inputs. To utilize coarse masks, MGMatting designed training perturbation strategies, including dilation and erosion, on guidance masks. For real-world adaptation, similar to Context-aware matting [29], MGMatting applies strong data augmentation including re-JPEGing, gaussian blur, and gaussian noises to input images during training. While MG-Wild [43] adopts the Mean teacher [44] mechanism on the matting task across composited matting data and real-world data to achieve better generalization ability.

Spatial sampling and alignment are useful learnable techniques to create more flexible neural networks. Deformable convolution [30] combines convolution kernels with learnable offsets to sample or aggregate features with flexible receptive fields. For video processing tasks, [31] uses learnable spatial alignment as optical flow. SFNet [32] dynamically aligns feature maps in different resolutions with learnable offsets.

Line detection. Line detection methods can be classified into handcrafted methods and learning-based methods. Handcrafted methods [33], [45]–[47] are traditionally performed based on the image gradient. Deep line detection was first introduced through wireframe [48] parsing tasks. Several approaches estimate the structural lines of a scene by representing the line segments with two endpoints [49], attraction fields [50], graphs [51], etc. DeepLSD [34] combines deep learning methods with classical line extractors, which supervises deep networks with attraction fields generated by LSD [33] through the homography adaptation technique [52] and uses the prediction of deep networks to improve the results of the LSD detector.

III. Our Auxiliary Learning Framework

Overall framework. To generalize to objects with complex structures in real-world scenes, avoid overfitting on wrong details, and suppress background interference, we proposed a novel auxiliary learning framework with three auxiliary tasks: semantic segmentation, edge detection, and background line detection, and an inconsistency-guided detail regularization (IGDR) module as shown in Fig. 3. Task 1 uses fine matting data for the matting task to learn detailed matting representations. Tasks 2 and 3 use the real-world segmentation data for both segmentation and edge detection tasks to learn real-world adaptive semantic representations on diverse and complex objects. Since background line detection needs a background image for composition, Task 4 uses only synthetic matting data for background line detection and matting, to learn a discriminative representation for better distinguishing foreground objects from background lines or textures. Additionally, we propose an inconsistency-guided detail regularization (IGDR) module into our auxiliary learning network, utilizing the inconsistency between matting representation and semantic representation to regularize low-level detail refinement.

Network architecture. We adopt a ResNet34-UNet matting network proposed in [7] with three matting heads as the base network. According to our auxiliary learning framework, we attach our segmentation head, edge detection head, and background line detection head to the features at output stride (OS) 8, 1, and 1 respectively to learn extra representations. And our IGDR module is inserted between OS32 and OS8 features, as shown in Fig. 3.

A. Learning real-world adaptive semantic representation

Models trained with synthetic or less diverse matting data tend to fail at complex structures and real-world scenes such as shadows, due to the limitation of composited data and the hardness to acquire precise alpha mattes for diverse objects in diverse real scenes. Additionally, these models simply learn detailed matting representation and neglect the high-level semantics of real objects. Since real-world data with segmentation masks is easier to attain and can provide semantic supervision for a large number of diverse and complex objects in various real scenes, we introduce an effective auxiliary learning framework to learn a real-world adaptive semantic representation, enabling the network to adapt to complex structures and various real scenes.

As shown in Fig. 3, besides the matting task and its supervision $L_{MatData}$ using the L_1 regression loss and Laplacian loss as [7] on matting data, we implement supervisions of semantic segmentation and edge detection on real-world segmentation data. Instead of naively using the segmentation data with binary masks to supervise the original matting heads, we incorporate additional supervisions, introducing an extra segmentation head on high-level feature maps (OS8) in our matting network. This allows us to learn the high-level semantics of target objects. Additionally, we introduce an extra edge detection head on high-resolution feature maps (OS1), fused with low-level features, to capture real-world boundaries and object contours. The segmentation ground-truth mask is provided by real-world segmentation datasets, while we generate binary edge ground-truth masks from segmentation ground-truth masks like [53] did. Hence, the total loss of Tasks 2 and 3 on real-world segmentation data $L_{SegData}$ can be formulated as:

$$L_{SegData} = L_{Seg} + L_{Edge}, \tag{2}$$

where L_{Edge} denotes the weighted cross entropy loss [53] function between the edge map by our edge detection head and the binary edge GT mask; L_{Seg} denotes the binary cross entropy loss function between our segmentation output and the binary GT mask.

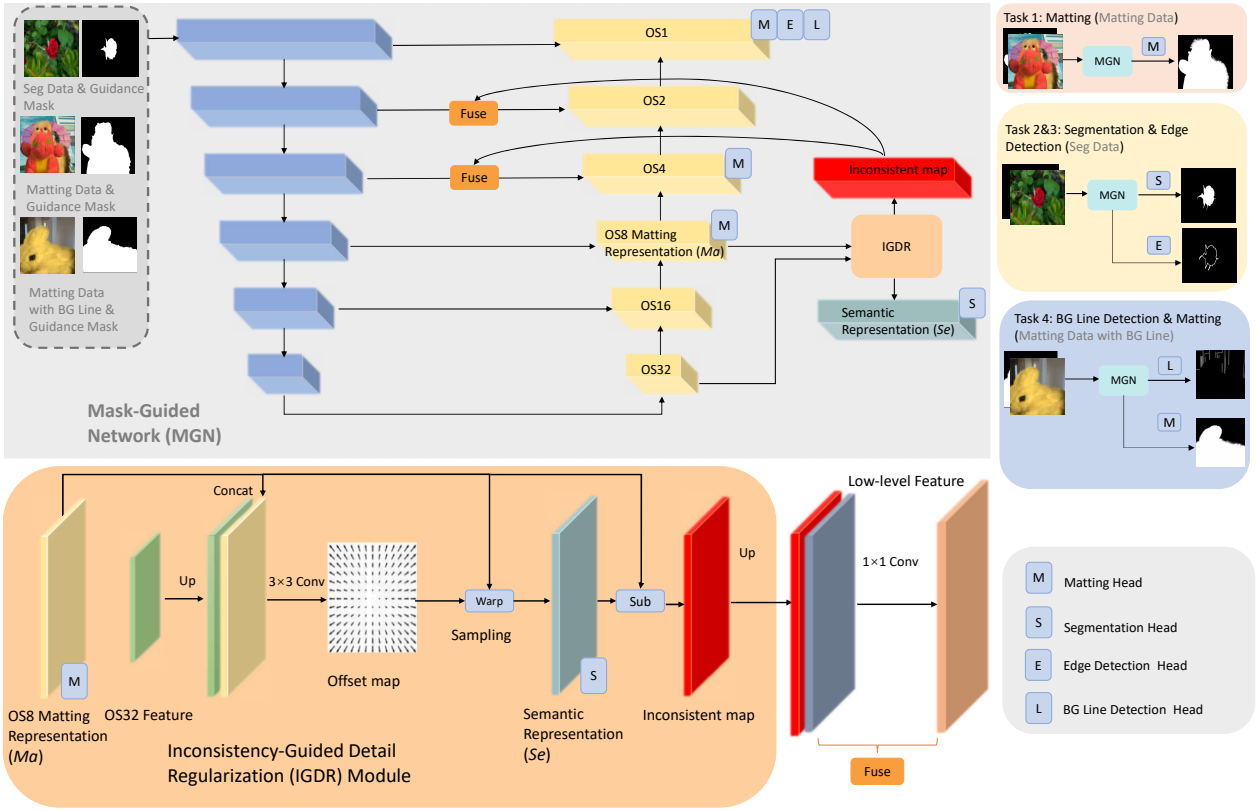

Fig. 3. Overview of our proposed auxiliary learning framework and our proposed network. The proposed network leans multiple representations from different types of data and annotations in our auxiliary learning framework. Our IGDR module uses the inconsistency between matting representation and real-world adaptive semantic representation to regularize refinement on low-level details.

In addition to learning the matting representation (Ma) from matting data, we leverage real-world segmentation data to learn a real-world adaptive semantic representation (Se). This auxiliary representation helps the network handle diverse and complex objects across various real-world scenes effectively.

B. Inconsistency-Guided detail regularization

Matting approaches focus on refining details of objects, but they are also prone to overfit low-level details in the wrong regions (SARI's textures on a woman's body) as MGMatting in Row 2, Fig. 1. As shown in Fig. 2, we use labels of matting and semantic segmentation to point out where the network should focus on, for matting and segmentation tasks, respectively. It's easy to observe that the segmentation mask can be treated as a warped alpha matte with fewer low-level details and their inconsistent regions highlight low-level details of correct regions that should be focused on. Based on this observation, and inspired by spatial sampling or alignment works [30]–[32] for spatial wrapping, we proposed an inconsistency-guided detail regularization module to guide and enhance low-level details of objects in proper regions and avoid overfitting in wrong regions.

As shown in Fig. 3, our IGDR generates semantic representation $Se \in R^{H\times W\times C}$ from matting representation $Ma \in R^{H\times W\times C}$ with learnable spatial wrapping and then acquires their inconsistency to guide detail regularization. We firstly introduce the high-level semantic information from the OS32 feature map by concatenating it with the matting representation Ma, and use a 3×3 convolution to generate an offset map $\Delta \in R^{H\times W\times 2}$ for spatial sampling. Then, to form Se, for every spatial point p in Se, the warp process in Fig. 3 bilinearly samples a point $p' = p + \Delta(p)$ in Ma as Eq 3,

$$Se(p) = \sum_{p_n \in \mathcal{N}(p+\Delta(p))} w_{p_n} Ma(p_n), \tag{3}$$

where $\mathcal{N}(p+\Delta(p))$ denotes neighbors of the warped point $p+\Delta(p)$ in Ma, and w_p denotes the bi-linear kernel weights calculated by the distance of warped grid. As shown in Fig. 3, Ma and Se are fed to a matting head and a segmentation head, respectively, to learn corresponding representations. Then, we generate the inconsistent map IN by $IN = Ma - Se$, which points out proper regions of important low-level details in the feature space. Subsequently, we use the inconsistent map IN from our IGDR module as guidance and fuse it with low-level feature maps, to refine low-level details in proper regions and prevent overfitting them in wrong regions.

C. Learning background line detection

Distinguishing target objects in detail and suppressing interference of background lines or textures is an important

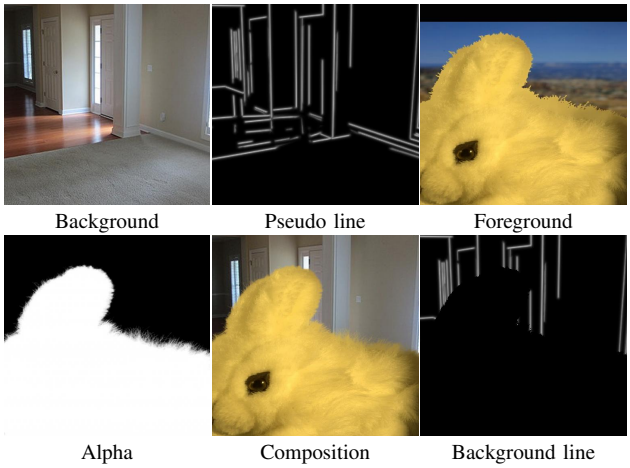

Fig. 4. Visualization of a training sample for background line detection.

challenge for matting. Previous mask-guided matting networks trained with detailed matting data also suffer from interference of background textures. Therefore, we proposed a novel auxiliary task, the background line detection in our framework, in order to learn discriminative representation to distinguish foreground objects from background lines or textures.

To generate training samples for this task, for a background image such as "Background" in Fig. 4, we first generate a representative pseudo distance field $D \in R^{H\times W}$ denoting the distance from every pixel to the nearest line, through the homography adaptation [34], [52]. In detail, given a single background image $I \in R^{H\times W}$, we warp it with 100 random homographies H_i to generate the warped images I_i, detect line segments in all the I_i using the LSD [33] line detector, then warp back the segments into I to get a set L_i of lines, then we generate a pseudo distance field $D_i \in R^{H\times W}$ for every set L_i of lines, and then we calculate the median value of every spatial location of D_i to get a more representative pseudo distance field D. Subsequently, we convert D to an activation map $Pl = e^{-\frac{D}{2}} \in (0, 1]$, visually represented as the "Pseudo Line" in Fig. 4. Subsequently, as shown in Fig. 4, a matting training image ("Composition") is generated by compositing the background image with a foreground image ("Foreground") using its alpha matte A ("Alpha"). To get background line GT for supervision, we zero out Pl in the unseen part of the background based on the corresponding alpha matte A, and we get background lines GT Bl visualized as "Background line" in Fig. 4 by

$$Bl = \begin{cases} Pl, A < 0.8, \\ ignore, 0.8 \le A < 1, \\ 0, A = 1, \end{cases} , \tag{4}$$

we assign the value as ignore, when $0.8 \le A < 1$, to prevent step of Bl, and if the value of a pixel i in Bl is $ignore$, the loss function will not be calculated at i.

As shown in Fig. 3, we place the background line detection head on the high-resolution feature map (OS1). For Task 4, we use the pseudo GT Bl to supervise the output of background line detection head $\hat{Bl}$, and use GT alpha matte A to supervise the output of matting head $\hat{A}$, using L_1 regression loss in the neighborhood of lines in a background image based on the distance field D. The distance threshold for background line detection and matting in Task 4 are 13 and 3, respectively. The total loss L_{BG} for Task 4 can be formulated as:

$$L_{BG} = L_1^{Line}(D \le 13) + L_1^{Mat}(D \le 3), \tag{5}$$

where L_1^{Line} is the L_1 regression loss between $\hat{Bl}$ and Bl, and L_1^{Mat} is the L_1 regression loss between $\hat{A}$ and A. With this novel auxiliary task and supervision of pseudo background line, our network learns a discriminative representation to suppress background interference of lines or textures.

Finally, we establish the total matting framework with all our auxiliary tasks, and the total loss is formulated as:

$$L_{total} = L_{MatData} + L_{SegData} + L_{BG}. \tag{6}$$

IV. Our Plant-Mat Benchmark

To evaluate mask-guided matting for complex objects under proper and clear backgrounds or scenes, we propose a plant matting test dataset, the Plant-Mat, containing 130 plant images and ground-truth alpha mattes with complex object structures, high-quality annotations, and look-natural composition. Unlike portraits and animals, plant objects usually have a large number of holes and elongated branches, as well as complex shadows and reflections on leaves, which is hard to annotate manually in high quality. Therefore, we capture the image of diverse plants in a blue screen studio and utilize a blue screen matting technique [54] to generate preliminary foregrounds and alpha mattes. Then we denoise the alpha mattes with filters and manually refine the remaining defects. In this way, we can generate high-quality matting annotations for plants. Then we carefully composite the foregrounds of plants on proper clear background images by considering whether scenes or relationships between plants and background objects are proper. Finally, we get our high-quality Plant-Mat benchmark containing various plants with diverse and complex structures under clear backgrounds.

V. Experiments

A. Implementation details

Training data and annotations. We implement our auxiliary learning framework on different types of data and different types of annotation. For fine matting data, we adopt synthetic matting datasets including a subset of Adobe [2] with 269 foreground images like [7] did, Human-2k [55] and Animal-2k (AM-2k) [36], and a real-world portrait matting dataset P3M-10k [56] with about 10k images. To learn real-world semantic representation from diverse and complex real-world data, we introduce 2 high-resolution segmentation datasets, including UHRSD [57] and HRSOD [58]. As for unlabeled background images for composited training data, we adopt COCO [59] and Wireframe [48]. We generate a distance field of lines through a homography adaptation [34] using [33] for a background image, and then generate pseudo GT for background line detection based on the corresponding composited foreground.

Training data augmentation. We follow the strong data augmentation on synthetic training images and guidance perturbation in the real-world setting of MGMatting [7] for all

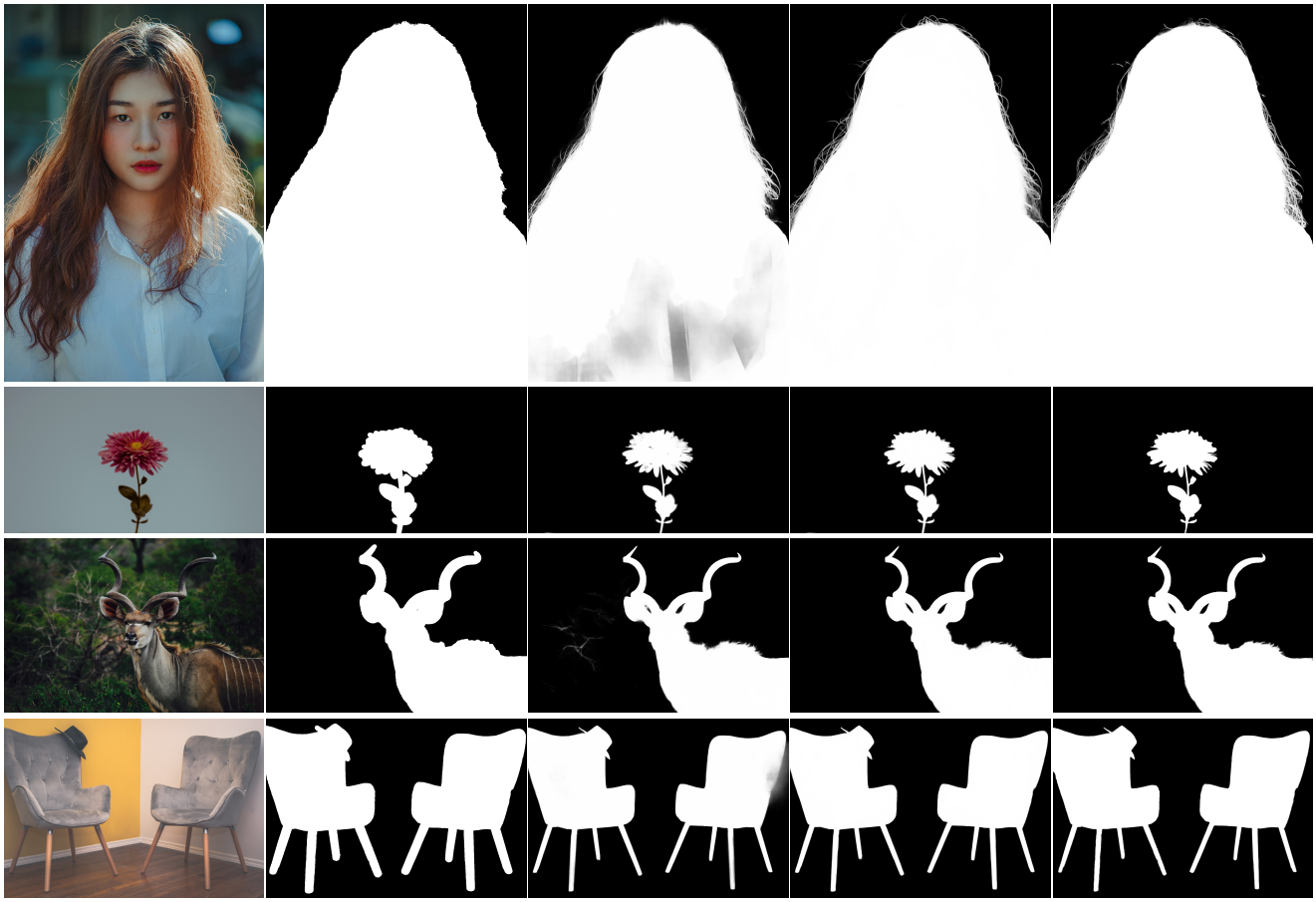

Fig. 5. The qualitative comparisons between MGMatting [7] and ours on various real-world images from test sets [7], [35], [36].

TABLE I
RESULTS ON RWP [7] BENCHMARK.

Model	Whole Image		Detail	
	SAD	MSE (10^{-3})	SAD	MSE (10^{-3})
DIM [2]	28.5	11.7	19.1	74.6
GCA [16]	29.2	12.7	19.7	82.3
Index [38]	28.5	11.5	18.8	72.7
LFM [39]	78.6	39.8	24.3	88.3
MODNet [37]	35.9	14.6	67.7	145.7
P3MNet [40]	36.5	18.6	19.3	80.7
MGMatting(official weight)	28.6	9.39	17.0	55.6
MGMatting(matting only)	27.9	9.55	16.8	58.0
Ours	**24.6**	**9.26**	**16.1**	**55.3**

TABLE II
RESULTS ON AIM-500 [35] BENCHMARK.

Model	SAD	MSE(10^{-3})	GRAD	CONN
Context-Aware [29]	32.2	38.8	30.3	31.0
GFM [36]	52.7	21.3	46.1	52.7
AIMNet [35]	43.9	16.1	33.1	43.2
MGMatting [7]	26.2	5.60	15.8	14.5
MG-Wild [43]	16.7	3.00	14.7	12.0
Ours	**14.3**	**2.67**	**12.4**	**11.2**

TABLE III
RESULTS ON AM-2K [36] BENCHMARK.

Model	SAD	MSE(10^{-4})	GRAD	CONN
AIMNet [35]	27.5	10.0	17.9	12.2
SHM [41]	17.8	68.0	12.5	17.0
GFM [36]	10.3	29.0	8.82	9.57
MGMatting [7]	10.1	10.4	5.58	7.11
MGMatting (matting only)	8.35	8.07	4.92	6.58
Ours	**5.86**	**5.95**	**3.67**	**4.55**

experiments. Especially, for Task 4 in our framework, we add binary lines on binary guidance masks with random widths (from 2 to 8) based on the distance field D, with a probability of 0.2, to perturb the guidance.

B. Benchmarks

We evaluate our mask-guided method on the RWP [7], PPM-100 [37], AM-2k [36], AIM-500 [35], and the proposed Plant-Mat benchmarks. The RWP benchmark officially and publicly provides 636 real-world portraits with matting ground truth and coarse guidance masks in various scenes. We adhere to its original evaluation protocol, which includes metrics for both the entire image and detailed regions. The PPM-100 [37] provides 100 high-resolution real-world portraits with matting ground truth for evaluation. The AM-2k [36] provides 200 real-world animal images with matting ground truth for evaluation. The AIM-500 [35] provides 500 images with matting ground truth for evaluation, containing various real-world objects. We follow the evaluation protocol in the mask-guided

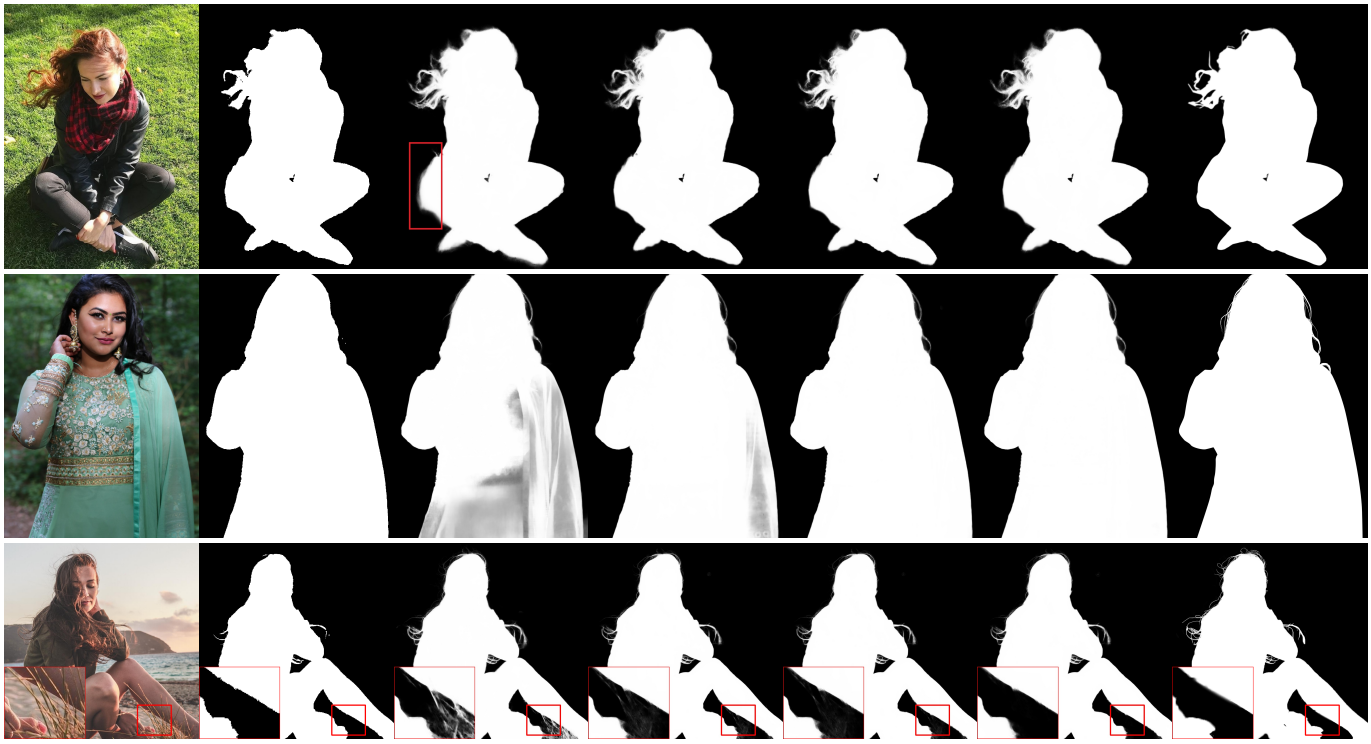

Fig. 6. The visual comparison results among different methods on real-world images. RASR: real-world adaptive semantic representation. IG: with our IGDR module. LD: auxiliary learning with background line detection as Task 4.

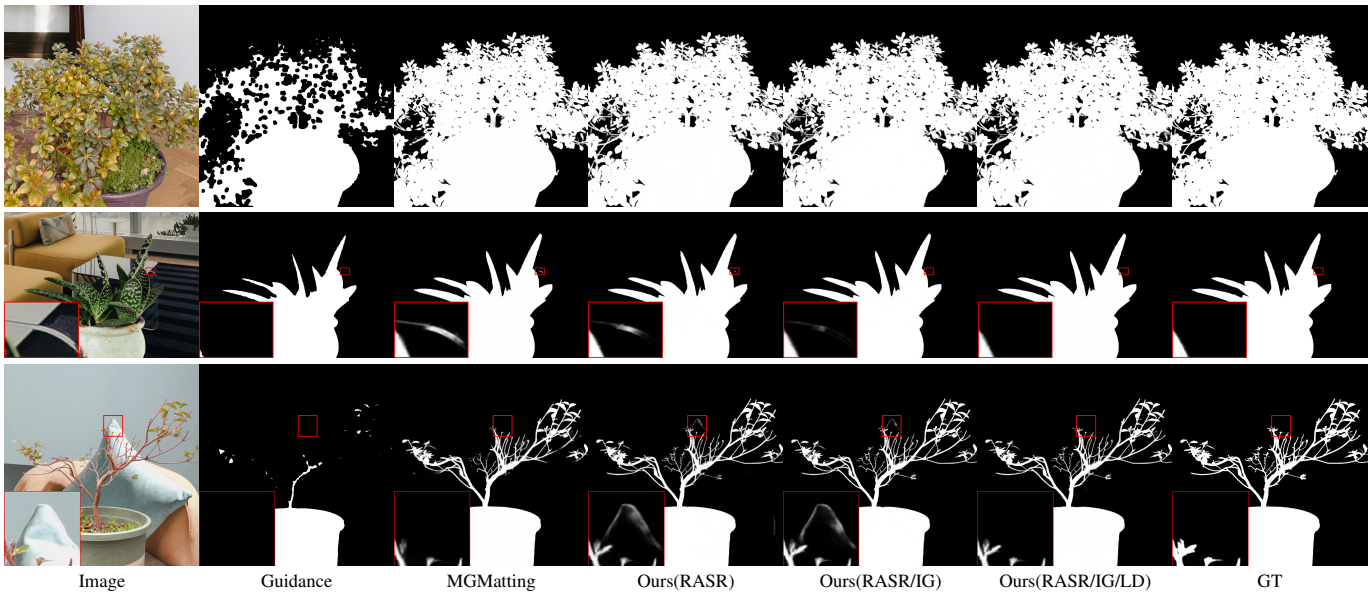

Fig. 7. The visual comparison results among different methods on Plant-Mat. RASR: real-world adaptive semantic representation. IG: with our IGDR module. LD: auxiliary learning with background line detection as Task 4.

method MG-wild [43] on AIM-500 [35], AM-2k [36], and PPM-100 [37]. Our Plant-Mat benchmark provides 130 plant images, which can evaluate mask-guided methods for objects with diverse and complex structures under clear backgrounds. To generate the coarse mask guidance for evaluation, we binarize the alpha matte of Plant-Mat with a threshold of 0.95 and then erode the binary mask with a 20×20 kernel.

Evaluation. We follow previous mask-guided methods to evaluate the results by Sum of Absolute Differences (SAD), Mean Squared Error (MSE), Gradient error (Grad), and Connectivity error (Conn) using the official evaluation code [7]. Since mask-guided matting has shown great practicality compared to traditional trimap-based or guidance-free methods, we focus on the comparison of mask-guided methods like MGMatting [7] or MG-Wild [43], but also report metrics for trimap-based methods [2], [16], [29], [38] and a trimap-free method [36], [37], [39]–[42] as a reference. We denote our baseline model trained with only matting data as "Matting

TABLE IV
RESULTS ON PPM-100 [37] BENCHMARK.

Model	SAD	MSE(10^{-4})	GRAD	CONN
AIMNet [35]	201.9	185.9	79.63	68.21
P3MNet [40]	130.8	128.6	56.37	130.4
MODNet [37]	95.1	44.7	64.26	80.82
RVM [42]	108.2	65.3	63.13	105.2
MGMatting	67.6	18.9	37.46	29.48
MGMatting(matting only)	40.0	8.80	36.17	28.50
Ours	**30.9**	**7.79**	**32.76**	**21.74**

TABLE V
ABLATION STUDY ON PPM-100 [37] BENCHMARK.

Model	SAD	MSE(10^{-4})	GRAD	CONN
Matting only	40.0	8.80	36.17	28.50
Ours(RASR)	35.4	8.64	33.45	23.88
Ours(RASR+IG)	32.3	8.00	32.95	21.97
Ours(RASR/IG/LD)	**30.9**	**7.79**	**32.76**	**21.74**

only". The model learning real-world adaptive semantic representation from Tasks 2 and 3 is denoted as "RASR". Our inconsistency-guided detail regularization (IGDR) module is denoted as "IG". The model trained with background line detection is denoted as "LD".

C. Qualitative and quantitative comparisons

As is shown in Tab. I, II, III, and IV, our method outperforms SOTA mask-guided methods such as MGMatting [7] and MG-Wild [43] on real-world matting benchmarks significantly. Specifically, on AIM-500 [35] with diverse real-world objects and scenes, our method, learning multiple auxiliary representations, outperforms the mask-guided method MG-Wild that solely learns its matting representation with the Mean Teacher mechanism for real-world generalization. Furthermore, our method significantly outperforms the SOTA MGMatting employing strong data augmentation proposed in [29] for real-world adaptation, on other high-quality real-world benchmarks including PPM-100 [37], RWP [7], and AM-2k [36]. As for our Plant-Mat benchmark with complex object structures and clear backgrounds, our method outperforms MGMatting by a large margin. For qualitative comparison in Fig. 1 and 5, our method adapts to diverse and complex real-world scenes and objects, performs better for objects with real shadows (Row 1, Fig. 1 and Row 4, Fig. 5) and complex structures like elongated cactus spines (Row 3, Fig. 1), regularizes low-level detail refinement (Row 2, Fig. 1), and suppresses interference of background lines or textures better (Row 3, Fig. 1 and Row 3, Fig. 5).

TABLE VI
RESULTS OF MASK-GUIDED MATTING METHODS ON OUR PLANT-MAT BENCHMARK.

Methods	SAD	MSE(10^{-3})	GRAD	CONN
MGMatting [7]	97.4	3.36	49.8	19.5
Matting only	60.0	2.17	48.0	18.4
Ours(RASR)	45.7	1.65	45.6	18.1
Ours(RASR/IG)	31.2	1.42	42.5	17.5
Ours(RASR/IG/LD)	**24.0**	**1.34**	**39.3**	**17.1**

D. Ablation analysis

Effectiveness of "RASR" and our IGDR module. With our real-world adaptive semantic representation learned from more diverse and complex real-world segmentation data, "RASR" achieves 4.6 and 14.3 SAD improvements on "Matting only" for PPM-100 (Tab. V) and Plant-Mat (Tab. VI), respectively. Besides metric improvements, "RASR" predicts better on real-world shadows (the right leg in Row 1, Fig. 6) and complex structures like elongated and irregular branches (Row 1 and 3, Fig. 7). Utilizing inconsistency between matting and semantic representations learned from "RASR", our inconsistency-guided detail regularization "IG" achieves 3.1 and 14.5 SAD improvements for PPM-100 and Plant-Mat, respectively. As shown in Row 2, Fig. 6, "IG" regularizes detail refinement and avoids overfitting on low-level details in the transparent SARI on a completely opaque body.

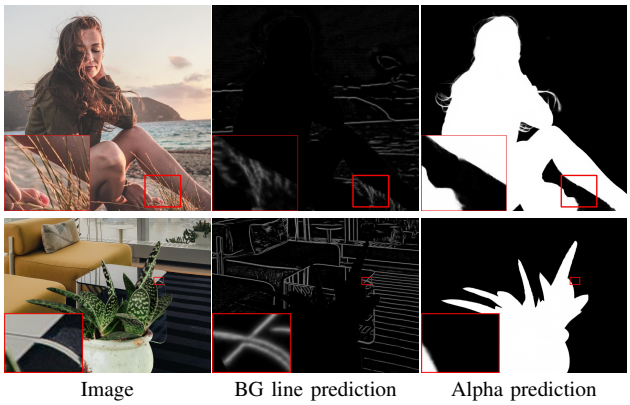

Fig. 8. Visualization for predictions of the background line and matting heads in our multi-task model. The auxiliary background line detection task learns representations that are aware of background lines, contributing to our matting predictions by suppressing background interference.

Effectiveness of our background line detection auxiliary task. Learning with our novel background line detection auxiliary task, our model learns a discriminative representation to suppress background interference. With our background line detection task, "LD" improves 1.4 SAD and achieves 30.9 SAD on the PPM-100 benchmark (Tab. V). For Plant-Mat with complex foreground structures over clear background lines, our "LD" performs better and achieves 7.2 SAD improvements. For qualitative results, we use red boxes in Fig. 6 and 7 to zoom in on the effect of one place of background lines and textures in an image. For a real-world portrait in Row 3, Fig. 6, both MGMatting and our models without "LD" suffer from interference from background lines or textures of weeds, while our model with "LD" suppresses the background interference and predicts the foreground well. For a plant in Row 2, Fig. 7, both MGMatting and our models without "LD" suffer from background interference from the glass of a table, while "LD" also suppresses the interference. We visualize predictions of corresponding background lines in Fig. 8. The predictions of background lines in Fig. 8 indicate that "LD" learns better representation to distinguish the background lines or textures, which helps our model suppress background interference. For plants in Row 3, Fig. 7, our model with "LD" simultaneously predicts the complex foreground structures and suppresses the

interference from clear background textures well. More details can be found in the supplemental material.

VI. Conclusion

In this paper, we propose a novel matting framework and model that learn different and effective representations through auxiliary learning and adopt a novel inconsistency-guided detail regularization, to address challenges in mask-guided matting. By introducing auxiliary semantic segmentation and edge detection tasks and leveraging more accessible coarse segmentation annotations on real-world data, our model acquires a superior real-world adaptive semantic representation alongside matting representation, enabling it to adapt to complex real-world objects and scenes. Utilizing the inconsistency between matting representation and semantic representation, our IGDR module regularizes the refinement of low-level details effectively. With our novel background line detection auxiliary task, our model learns discriminative representation to suppress background interference. In addition, we propose a plant matting dataset with complex object structures under proper and clear backgrounds, high-quality annotations, and look-natural composition to evaluate mask-guided matting methods. The quantitative and qualitative results on both established real-world matting benchmarks and our Plant-Mat demonstrate the superiority of our proposed method.

Acknowledgments

This paper is supported by NSFC, China (No. 62176155), Shanghai Municipal Science and Technology Major Project, China (2021SHZDZX0102), and OPPO Research Fund.

References

[1] J. Wang and M. F. Cohen, *Image and video matting: a survey*. Now Publishers Inc, 2008.

[2] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2970–2979.

[3] A. Rao *et al.*, "A coarse-to-fine framework for automatic video unscreen," *IEEE Transactions on Multimedia*, 2022.

[4] K.-T. Ng, Z.-Y. Zhu, C. Wang, S.-C. Chan, and H.-Y. Shum, "A multi-camera approach to image-based rendering and 3-d/multiview display of ancient chinese artifacts," *IEEE transactions on multimedia*, vol. 14, no. 6, pp. 1631–1641, 2012.

[5] F.-L. Zhang, M. Wang, and S.-M. Hu, "Aesthetic image enhancement by dependence-aware object recomposition," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1480–1490, 2013.

[6] C. L. Zitnick and S. B. Kang, "Stereo for image-based rendering using image over-segmentation," *International Journal of Computer Vision*, vol. 75, no. 1, pp. 49–65, 2007.

[7] Q. Yu *et al.*, "Mask guided matting via progressive refinement network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1154–1163.

[8] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 228–242, 2007.

[9] Y.-Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski, "A bayesian approach to digital matting," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 2. IEEE, 2001, pp. II–II.

[10] J. Wang and M. F. Cohen, "Optimized color sampling for robust matting," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[11] K. He, C. Rhemann, C. Rother, X. Tang, and J. Sun, "A global sampling method for alpha matting," in *CVPR 2011*. IEEE, 2011, pp. 2049–2056.

[12] E. S. Gastal and M. M. Oliveira, "Shared sampling for real-time alpha matting," in *Computer Graphics Forum*, vol. 29, no. 2. Wiley Online Library, 2010, pp. 575–584.

[13] Q. Chen, D. Li, and C.-K. Tang, "Knn matting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 9, pp. 2175–2188, 2013.

[14] K. He, J. Sun, and X. Tang, "Fast matting using large kernel matting laplacian matrices," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2165–2172.

[15] Y. Liu, J. Xie, Y. Qiao, Y. Tang, and X. Yang, "Prior-induced information alignment for image matting," *IEEE Transactions on Multimedia*, 2021.

[16] Y. Li and H. Lu, "Natural image matting via guided contextual attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 450–11 457.

[17] Y. Dai, H. Lu, and C. Shen, "Learning affinity-aware upsampling for deep image matting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6841–6850.

[18] G. Park, S. Son, J. Yoo, S. Kim, and N. Kwak, "Matteformer: Transformer-based image matting via prior-tokens," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 696–11 706.

[19] Y. Sun, C.-K. Tang, and Y.-W. Tai, "Semantic image matting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 120–11 129.

[20] S. Cai *et al.*, "Disentangled image matting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8819–8828.

[21] M. Forte and F. Pitié, "f, b, alpha matting," *arXiv preprint arXiv:2003.07711*, 2020.

[22] Q. Liu, H. Xie, S. Zhang, B. Zhong, and R. Ji, "Long-range feature propagating for natural image matting," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 526–534.

[23] Y. Sun, C.-K. Tang, and Y.-W. Tai, "Semantic image matting: General and specific semantics," *International Journal of Computer Vision*, pp. 1–21, 2023.

[24] H. Yu, N. Xu, Z. Huang, Y. Zhou, and H. Shi, "High-resolution deep image matting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3217–3224.

[25] J. Tang, Y. Aksoy, C. Oztireli, M. Gross, and T. O. Aydin, "Learning-based sampling for natural image matting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3055–3063.

[26] S. Lutz, K. Amplianitis, and A. Smolic, "Alphagan: Generative adversarial networks for natural image matting," in *BMVC*, 2018.

[27] Y. Dai, B. Price, H. Zhang, and C. Shen, "Boosting robustness of image matting with context assembling and strong data augmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 707–11 716.

[28] Y. Li, J. Zhang, W. Zhao, W. Jiang, and H. Lu, "Inductive guided filter: Real-time deep matting with weakly annotated masks on mobile devices," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[29] Q. Hou and F. Liu, "Context-aware image matting for simultaneous foreground and alpha estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4130–4139.

[30] J. Dai *et al.*, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[31] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *CVPR*, July 2017.

[32] X. Li *et al.*, "Semantic flow for fast and accurate scene parsing," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 775–793.

[33] R. G. Von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "Lsd: A fast line segment detector with a false detection control," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 4, pp. 722–732, 2008.

[34] R. Pautrat, D. Barath, V. Larsson, M. R. Oswald, and M. Pollefeys, "Deeplsd: Line segment detection and refinement with deep image gradients," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 327–17 336.

[35] J. Li, J. Zhang, and D. Tao, "Deep automatic natural image matting," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Z.-H. Zhou, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2021, pp. 800–806, main Track.

[36] J. Li, J. Zhang, S. J. Maybank, and D. Tao, "Bridging composite and real: towards end-to-end deep image matting," *International Journal of Computer Vision*, vol. 130, no. 2, pp. 246–266, 2022.
[37] Z. Ke, J. Sun, K. Li, Q. Yan, and R. W. Lau, "Modnet: Real-time trimap-free portrait matting via objective decomposition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 1140–1147.
[38] H. Lu, Y. Dai, C. Shen, and S. Xu, "Indices matter: Learning to index for deep image matting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3266–3275.
[39] Y. Zhang *et al.*, "A late fusion cnn for digital matting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7469–7478.
[40] S. Ma, J. Li, J. Zhang, H. Zhang, and D. Tao, "Rethinking portrait matting with pirvacy preserving," *International Journal of Computer Vision*, 2023.
[41] Q. Chen, T. Ge, Y. Xu, Z. Zhang, X. Yang, and K. Gai, "Semantic human matting," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 618–626.
[42] S. Lin, L. Yang, I. Saleemi, and S. Sengupta, "Robust high-resolution video matting with temporal guidance," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 238–247.
[43] K. Park, S. Woo, S. W. Oh, I. S. Kweon, and J.-Y. Lee, "Mask-guided matting in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1992–2001.
[44] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.
[45] C. Akinlar and C. Topal, "Edlines: Real-time line segment detection by edge drawing (ed)," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 2837–2840.
[46] Y. Salaün, R. Marlet, and P. Monasse, "Multiscale line segment detector for robust and accurate sfm," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 2000–2005.
[47] I. Suárez, J. M. Buenaposada, and L. Baumela, "Elsed: Enhanced line segment drawing," *Pattern Recognition*, vol. 127, p. 108619, 2022.
[48] K. Huang, Y. Wang, Z. Zhou, T. Ding, S. Gao, and Y. Ma, "Learning to parse wireframes in images of man-made environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 626–635.
[49] Y. Zhou, H. Qi, and Y. Ma, "End-to-end wireframe parsing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 962–971.
[50] N. Xue, S. Bai, F. Wang, G.-S. Xia, T. Wu, and L. Zhang, "Learning attraction field representation for robust line segment detection," 2019.
[51] Z. Zhang *et al.*, "Ppgnet: Learning point-pair graph for line segment detection," 2019.
[52] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.
[53] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, and Y. Zhao, "Devil in the details: Towards accurate single and multiple human parsing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4814–4821.
[54] A. R. Smith and J. F. Blinn, "Blue screen matting," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 259–268.
[55] Y. Liu *et al.*, "Tripartite information mining and integration for image matting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7555–7564.
[56] J. Li, S. Ma, J. Zhang, and D. Tao, "Privacy-preserving portrait matting," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3501–3509.
[57] C. Xie, C. Xia, M. Ma, Z. Zhao, X. Chen, and J. Li, "Pyramid grafting network for one-stage high resolution saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 717–11 726.
[58] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, and H. Lu, "Towards high-resolution salient object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7234–7243.
[59] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.