Privileged Prior Information Distillation for Image Matting

Cheng Lyu[1*], Jiake Xie[2*], Bo Xu[3*†], Cheng Lu[3], Han Huang[4], Xin Huang[5], Ming Wu[1], Chuang Zhang[1], Yong Tang[2]

[1]Beijing University of Posts and Telecommunications
[2]PicUp.Ai
[3]Xpeng
[4]AI[2] Robotics
[5]Towson University
{lvcheng, wuming, zhangchuang}@bupt.edu.cn, {jxie,ty}@road.win, Xhuang@towson.edu

Abstract

Performance of trimap-free image matting methods is limited when trying to decouple the deterministic and undetermined regions, especially in the scenes where foregrounds are semantically ambiguous, chromaless, or high transmittance. In this paper, we propose a novel framework named Privileged Prior Information Distillation for Image Matting (PPID-IM) that can effectively transfer privileged prior environment-aware information to improve the performance of trimap-free students in solving hard foregrounds. The prior information of trimap regulates only the teacher model during the training stage, while not being fed into the student network during actual inference. To achieve effective privileged cross-modality ($i.e.$ trimap and RGB) information distillation, we introduce a Cross-Level Semantic Distillation (CLSD) module that reinforces the students with more knowledgeable semantic representations and environment-aware information. We also propose an Attention-Guided Local Distillation module that efficiently transfers privileged local attributes from the trimap-based teacher to trimap-free students for the guidance of local-region optimization. Extensive experiments demonstrate the effectiveness and superiority of our PPID on image matting. The code will be released soon.

1 Introduction

Image matting is one of the most fundamental computer vision tasks, which aims to separate the foreground objects from a single image or video stream. It has tremendous practical value for background replacement task in multimedia applications such as image/video entertainment creation, special-effect film-making and live video. To accurately estimate the opacity of each pixel inside foreground regions, the matting is generally formulated as an image/video frame composite problem, which solves the 7 unknown variables per pixel from only 3 known values:

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

where I_i refers to known 3-dimensional RGB color at pixel i, while foreground RGB color F_i, background RGB color B_i, and alpha matte estimation α_i are unknown. According to the above equation, $\alpha_i = 1$, $\alpha_i = 0$, and $\alpha_i \in (0, 1)$ represent the deterministic foreground, background, and undetermined regions, respectively.

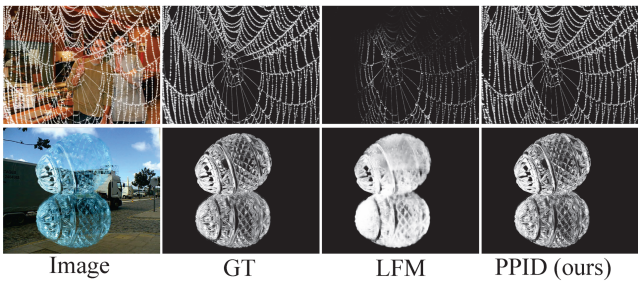

Figure 1: Given challenging images of chromaless objects and semantically ambiguous scenes, one SOTA trimap-free matting approach - LFM (Zhang et al. 2019), fails to decouple the deterministic and undetermined regions. However, our PPID-guided model can perform more accurate and fine-grained region decoupling for these hard foregrounds.

To solve this highly challenging problem, the typical methods (Xu et al. 2017; Hou and Liu 2019; Lu et al. 2019; Li and Lu 2020; Liu et al. 2021) utilize trimap as a piece of environment-aware priori that marks the foreground, background and transition regions to locate the targets and reduce the solution spaces. Unfortunately, a high-quality trimap can not be obtained without tedious manual annotation effort and significant time costs, which limits its practical application in low-cost consumer products. Some trimap-free methods (Zhang et al. 2019; Qiao et al. 2020) are proposed to utilize the typical encoder-to-decoder structures that stem from segmentation and detection, etc, for alpha prediction without auxiliary cues. Although environmental saliency can be roughly predicted by borrowing such transitional network structure and its pre-trained feature, there still remain two tricky challenges in the trimap-free matting setting. First, previous trimap-free matting methods may fail to identify certain attributes when the foreground is semantically ambiguous, chromaless, or with high transmittance. For example in Row 2 of Fig. 1, due to the lack of strong opaqueness and color hue difference, the previous trimap-free net-

*These authors contributed equally.
†Bo Xu is the corresponding author.

work (Zhang et al. 2019) fails to decouple the deterministic and undetermined regions of this highly transmissive 'glass ball'. That is especially obvious in those upper pixels where the networks are confused by the background (sky) with similar colors.

Second, for most of the existing trimap-free models, the learning of semantic mining often stems from upstream tasks such as segmentation. However, they may fail to identify the complete and meaningful foregrounds when such foregrounds are semantically ambiguous, rare, or spatially sparse, $i.e.$ the cobweb in Row 1 of Fig. 1. Although several current methods try to construct pseudo-trimap or implicitly learn the transition region distribution as a decoupling effort by imposing local supervision on RGB features, they may fail to balance global and local matting quality ($i.e.$ texture similarity, location correlation, $etc.$) which leads to incomplete information mining.

To address the above challenges, we propose a novel Privileged Prior Information Distillation framework for Image Matting (PPID-IM). Our PPID-IM framework can effectively transfer privileged environment-aware prior information to improve the trimap-free models, leading to 1) more accurate region decoupling for hard foregrounds which are chromaless or with high transmittance 2) better balance between global and local matting quality. We rethink trimap-free matting via privileged information distillation, with the following motivations: a) Trimap provides critical environment-aware information which the trimap-free models may lack and focuses on the undetermined regions with semantic ambiguity and poor color hue. Consequently, it can be formulated as a privileged modality in the knowledge distillation framework. b) Due to the successes of privileged information ($i.e.$ features, extra modalities, $etc.$) distillation in multiple tasks, $i.e.$ image classification (Lopez-Paz et al. 2015), object detection (Yang et al. 2022), and image super-resolution (Lee et al. 2020), action detection (Zhao et al. 2020), etc, we can borrow a similar idea to introduce trimap as privileged information for knowledge distillation that can improve the performance of alpha matte prediction in a trimap-free setting.

To showcase this new matting framework, we employ multiple models from both original trimap-based methods and the trimap-based variants of state-of-the-art (SOTA) trimap-free methods as teachers to guide the training of their corresponding trimap-free students, which aims to demonstrate the generalization of our PPID for image matting. Each paired teacher and student share the same structure except for the inputs, where both RGB images and trimaps are given to the teacher and only RGB images for the student. To leverage effective privileged features for environment-aware information distillation, we introduce a Cross-Level Semantic Distillation (CLSD) module that guides each student layer to learn from more knowledgeable privileged features of the teacher model in addition to the corresponding layer distillation, for mining more effective environment-aware information. In addition, we propose an Attention-Guided Local Distillation module that efficiently transfers privileged local attributes from the trimap-based teacher to guide local region optimization for the trimap-free student.

Overall, the contributions of this paper are as follows:

- We propose a novel Privileged Prior Information Distillation framework for Image Matting that can effectively transfer privileged prior information to improve the trimap-free models, especially in scenes with chromaless, semantic ambiguity, or irregular objects.
- We introduce a Cross-Level Semantic Distillation (CLSD) module that complements the student networks with more sufficient environmental awareness and higher-level semantic feature representations.
- We also propose an Attention-Guided Local Distillation module to guide local region optimization of the trimap-free student by borrowing privileged local attributes from the trimap-based teacher.
- Extensive experiments on public datasets demonstrate the effectiveness and superiority of our proposed framework, outperforming the SOTA approaches on both synthetic and real-world images by a large margin.

2 Related Works

2.1 Classic Methods

Classic foreground matting methods can be generally categorized into two approaches: sampling-based and propagation-based. Sampling-based methods (Chen et al. 2013; Jue and Michael F 2007; Kaiming et al. 2011; Yung-Yu et al. 2001) 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, $e.g.$ Bayesian matting (Yung-Yu et al. 2001), optimized color sampling (Jue and Michael F 2007), global sampling method (Kaiming et al. 2011), and comprehensive sampling (Ehsan et al. 2013). Propagation-based methods (Chen, Li, and Tang 2013; He, Sun, and Tang 2010; Levin, Lischinski, and Weiss 2007) 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. Classic matting 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.

2.2 Deep Learning-Based Methods

Trimap-Based Methods. Initially, some attempts are made to combine deep learning networks with classic matting techniques, $e.g.$ closed-form matting (Levin, Lischinski, and Weiss 2007) and KNN matting (Chen, Li, and Tang 2013). Cho $et\ al.$ (Cho, Tai, and Kweon 2016) employ a deep neural network to improve the results of the closed-form matting and KNN matting. For facilitating the end-to-end training, Xu $et\ al.$ (Xu et al. 2017) propose a two-stage deep neural network (Deep Image Matting) based on SegNet (Badrinarayanan, Kendall, and Cipolla 2017) for alpha matte estimation and contribute a large-scale composition image matting dataset (Adobe dataset) with ground truth foreground (alpha) matte. Lutz $et\ al.$ (Lutz, Amplianitis, and Smolic 2018) introduce a generative adversarial network (GAN) for

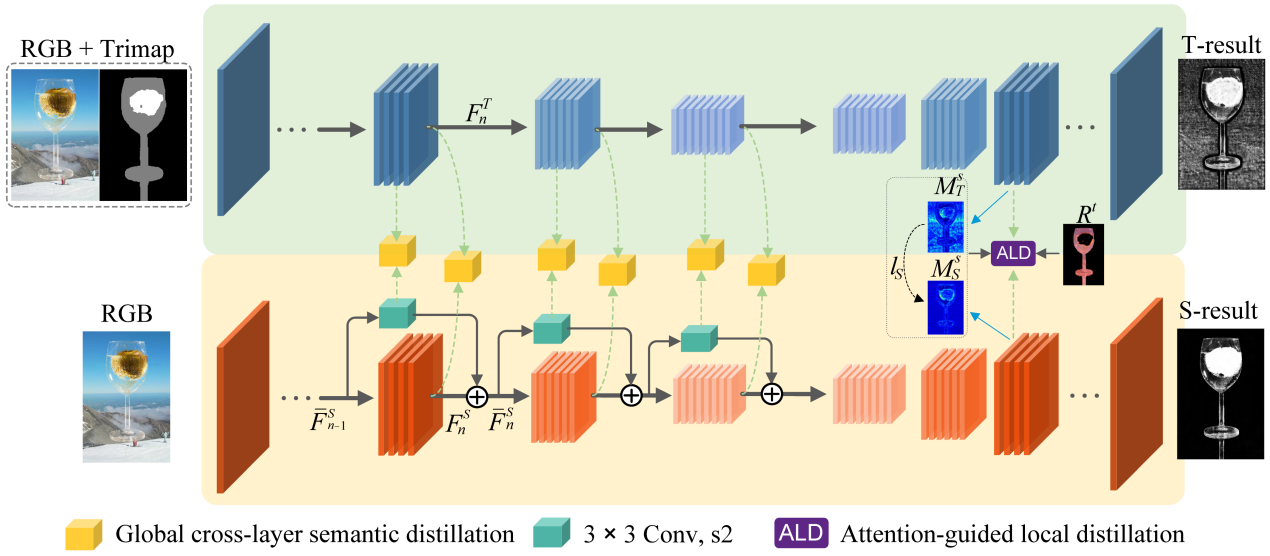

Figure 2: Architecture of the Privileged Prior Information Distillation for Image Matting (PPID-IM). The privileged information distillation is operated between the trimap-based teacher and its trimap-free student variant by an efficient Cross-Layer Semantic Distillation (CLSD) module. Subsequently, the privileged local attributes of the teacher network are further transferred to the student via an Attention-guided Local Distillation (ALD) module, for facilitating the student's local optimization. T-result and S-result denote the outputs of the teacher and student respectively during training.

natural image matting and improve the results of Deep Image Matting (Xu et al. 2017). Subsequently, a range of methods (Hou and Liu 2019; Lu et al. 2019; Li and Lu 2020; Cai et al. 2019) have introduced different theoretical developments for matting performance improvements.

Trimap-Free Methods. Fully automatic matting without any auxiliary additional constraints (*e.g.* user-annotated trimap) has also been studied. (Chen et al. 2018) tries to predict the trimap first, followed by an alpha matting network. Zhang *et al.* (Zhang et al. 2019) propose a dual-decoder network for foreground and background classification, followed by a fusion branch to integrate the dual results. Ke *et al.* (Ke et al. 2022) present a lightweight matting objective decomposition network (MODNet) and introduce an e-ASPP module for efficient multi-scale feature fusion. Lin *et al.* (Lin et al. 2022) propose a robust real-time matting method (RVM) training strategy that optimizes the network on both matting and segmentation tasks.

2.3 Privileged Information Distillation Methods

Lopez-Paz *et al.* (Lopez-Paz et al. 2015) first combine Distillation (Hinton et al. 2015) and privileged information (Vapnik, Izmailov et al. 2015) into generalized distillation, and extended it to unsupervised, semi-supervised and multi-task learning scenarios. Li *et al.* (Li et al. 2022a) propose a cross-modality knowledge distillation model that leverages the additionally privileged depth to guide the training of the monocular visual odometry network. Wang *et al.* (Wang and Chen 2020) propose a Privileged Modality Distillation Network that improves the RGB-based hand pose estimation by excavating the privileged information from depth prior during training. Zhao *et al.* (Zhao et al. 2020) consider future frames from the off-line teacher as privileged information to guide the online student for action detection, by knowledge distillation.

3 Methodology

Our Privileged Prior Information Distillation for image matting (PPID-IM) is designed to effectively transfer the privileged prior information that can guide trimap-free students in capturing sufficient environmental awareness information and performing more accurate domain decoupling. The overall architecture of the PPID-IM is shown in Fig. 2, PPID consists of dual key components: cross-level semantic distillation and attention-guided local distillation modules.

3.1 Cross-layer Semantic Distillation

To effectively leverage privileged prior features for environment-aware information distillation, we propose a Cross-Level Semantic Distillation (CLSD) module that guides each student layer to learn higher-level privileged feature representations from the teacher in addition to the corresponding layer distillation.

Inspired by multiple tasks (Hu et al. 2018; Hu, Shen, and Sun 2018; Wang et al. 2018; Zhang et al. 2018; Cao et al. 2019) that need a global understanding of a visual scene, we believe that extracting the associations between pixels both within a region (*i.e.* known, transition, or background regions) and across regions can lead to capturing more sufficient environment-aware information. And then, we employ the Global Context block (GCblock) (Cao et al. 2019) to model the global association context of the teacher's feature representations. Due to the motivation that learning higher-level semantic information can enhance the students'

perception of more comprehensive information, we further transfer the teacher's representations to both the corresponding layer and neighbor layer of the student network for effective privileged information distillation. The GCblock ($Gc(F)$) (Cao et al. 2019) is formulated as follows:

$$Gc(F) = F + W_{v2}\big(ReLU\big(LN\big(W_{v1} \big(\sum_{j=1}^{N_p} \frac{e^{W_k F_j}}{\sum_{m=1}^{N_p} e^{W_k F_m}} F_j\big)\big)\big)\big) \tag{2}$$

where W_k, W_{v1} and W_{v2} denote linear transformation matrices, LN denotes the layer normalization, N_p denotes the number of positions in the feature map($i.e.$ $N_p = H \cdot W$). The global feature transfer of CLSD can be formulated as:

$$L_{CLSD} = \lambda_1 \cdot \sum \Big(Gc(F_n^T) - Gc(F_n^S)\Big)^2 + \lambda_2 \cdot \sum \Big(Gc(F_n^T) - Gc\big(f^{3\times3}(\bar{F}_{n-1}^S)\big)\Big)^2 \tag{3}$$

$$\bar{F}_n^S = F_n^S + f^{3\times3}(\bar{F}_{n-1}^S) \tag{4}$$

where $Gc(F)$ represents the GCblock (Cao et al. 2019). F_n^T and F_n^S denote the feature maps of the teacher and student at the n-th layer, respectively. $f^{3\times3}$ is a 3×3 convolutional operation (stride=2). λ_1 and λ_2 are hyper-parameters for balancing the loss. We further update F_n^S to $\bar{F}_n^S$ by merging it with the cross-layer distilled feature $\bar{F}_{n-1}^S$ at $(n-1)$-th layer.

3.2 Attention-Guided Local Distillation

Most previous trimap-free methods are limited in terms of domain decoupling on local regions, due to the lack of prior constraints on the solution space. To address this situation, we propose an Attention-Guided Local Distillation (ALD) module that efficiently transfers privileged local attributes from the trimap-based teacher to guide the local region optimization for the trimap-free student. For efficient locally privileged information distillation, we generate a region mask to guide the transfer of effective pixel-level representations to the student, by the following formula:

$$R_{i,j}^t = \begin{cases} 1, & if\,(i,j) \in t \\ 0, & otherwise \end{cases} \tag{5}$$

where t denotes the transition region, and i,j refers to the pixel position in the feature map.

Specifically, we also introduce a spatial attention mask to further emphasize crucial information and suppress disturbing information about the target local regions. We follow CBAM (Woo et al. 2018) to respectively apply average-pooling (AP) and max-pooling (MP) along the channel and then feed the concatenated feature into a 7×7 convolution layer to generate the teacher's attention mask, as follows:

$$M^s(F) = \sigma(f^{7\times7}\left(\frac{[AP(F); MP(F)]}{T}\right)) \tag{6}$$

where σ denotes the Softmax function and $[\cdot\,;\cdot]$ is a concatenation operation along the channel axis. T is the temperature

Dataset	Volume	Natural	UHD	Trans.	Non-trans.	Category
DAPM	200	-	-	0	200	Portrait
Adobe	50	-	-	12	38	Multiple
Dist-646	50	-	-	10	40	Multiple
AM-2k	200	✓	-	0	200	Animal
AIM	500	✓	-	43	457	Multiple
RWP-636	636	✓	-	0	636	Portrait
Real-1K	1000	✓	✓	396	604	Multiple

Table 1: The configurations of image matting test sets.

hyper-parameter (Hinton et al. 2015) for distribution adjustment. Then the feature loss of the ALD module can be formulated as follows:

$$L_{ALD}^f = \sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W} R_{ij}^t M_{i,j}^s \left(F_{k,i,j}^T - F_{k,i,j}^S\right)^2 \tag{7}$$

where M^s denotes the teacher's attention map. F^T and F^S refer to the feature of the teacher and student, respectively. C, H, and W denote the channel, height, and width of the given feature, respectively. Further, we introduce an attention distillation function L_{ALD}^a that can guide the student to learn crucial information distribution of the transition regions from the teacher:

$$L_{ALD}^a = l_s\left(R^t M_T^s, R^t M_S^s\right) \tag{8}$$

where M_T^s and M_S^s denotes the attention maps of the teacher and student, respectively. l_s is a L1 loss. Overall, the loss function of the ALD module (L_{ALD}) can be formulated by the combination of L_{ALD}^f and L_{ALD}^a:

$$L_{ALD} = \alpha L_{ALD}^f + \beta L_{ALD}^a \tag{9}$$

where α and β are the hyper-parameters to balance the feature loss and attention loss. And then, the final loss function of the privileged prior information distillation is computed as:

$$L_{distn} = L_{CLSD} + L_{ALD} \tag{10}$$

3.3 Loss Function of Alpha Prediction

According to Eq. 1, one image can be divided into three solution regions, $i.e.$ known foreground, background, and transition, where the scale of each region is commonly different. The large-scale regions tend to receive more attention in terms of loss, while the supervision in small-scale ones may be weakened due to their low pixel proportion, which will lead to an unbalance in alpha predictions of multiple regions. To address this, we introduce a regional scaling mask to balance the supervision of each region, as follows:

$$S_{ij} = \begin{cases} \frac{1}{N_f}, & if\,(i,j) \in f \\ \frac{1}{N_b}, & if\,(i,j) \in b \\ \frac{1}{N_t}, & if\,(i,j) \in t \end{cases} \tag{11}$$

where i,j denotes the pixel position in the feature map, N_f, N_b and N_t represent the total number of pixels in the foreground, background, and transition region, respectively. For the supervision of alpha prediction, we employ L1 loss (L_1),

cross-entropy (CE) loss (L_{ce}), and gradient loss (L_{grad}) to reinforce different predictive attributes of the alpha.

$$L_{alpha} = \sum_{i=1}^{H_\alpha}\sum_{j=1}^{W_\alpha} w_{ij}^{L1} S_{ij} L_1^i + \sum_{i=1}^{H_\alpha}\sum_{j=1}^{W_\alpha} w_{ij}^{ce} S_{ij} L_{ce}^i + \sum_{i=1}^{H_\alpha}\sum_{j=1}^{W_\alpha} L_{grad}^i \tag{12}$$

where H_α and W_α are the height and width of the final alpha prediction. w_{ij}^{L1} and W_{ij}^{ce} denote the weight of L_1 and L_{ce} respectively and are set as follows:

$$w_{ij}^{L1} = \begin{cases} 1, & if\,(i,j) \in f \bigcup b \\ 2, & if\,(i,j) \in t \end{cases} \tag{13}$$

$$w_{ij}^{ce} = \begin{cases} 1, & if\,(i,j) \in f \bigcup b \\ 0.5, & if\,(i,j) \in t \end{cases} \tag{14}$$

where f, b, and t denote the foreground, background and transition region, respectively. The L_1 loss is the absolute difference between the predicted and the ground truth alpha matte, and can recover both the global and local detail alpha values:

$$L_1^i = \|\alpha_i - \alpha_i^*\|_1 \tag{15}$$

where α_i and α_i^* denote the student output and the ground truth alpha values at pixel i, respectively.

We follow (Zhang et al. 2019) to introduce the L_{ce} loss to accelerate the convergence of the pixels in foreground and background regions towards their targets, as follows:

$$L_{ce}^i = -[\alpha_i^* \cdot log\,(\alpha_i) + (1-\alpha_i^*) \cdot log\,(1-\alpha_i)] \tag{16}$$

Similar to (Zhang et al. 2019), we set a small weight for L_{ce} and combine it with L_1 to supervise alpha prediction in the transition region. In addition, we utilize the gradient loss L_{grad} to reduce the over-blurred alpha results, as follows:

$$L_{grad}^i = |\nabla\alpha_i - \nabla\alpha_i^*| \tag{17}$$

where ∇ denotes the calculation of the gradient magnitude.

3.4 Overall Training Loss

Overall, we train the student network with the total loss as follows:

$$L = L_{alpha} + \gamma L_{distn} \tag{18}$$

where γ is the hyper-parameter to balance the alpha-prediction loss L_{alpha} and the the privileged prior information distillation loss L_{distn}.

4 Experiments

4.1 Benchmark: Real-1K

We propose the first large-scale UHD (Ultra High Definition) natural image matting test set Real-World Image Matting-1K (Real-1K), which contains 1000 ultra high-resolution (from 4K to 8K) real-world natural samples of transparent and non-transparent attributes. Table 1 shows comparisons between some existing image matting datasets (DAPM (Shen et al. 2016), Adobe (Xu et al. 2017), Dist-646 (Qiao et al. 2020), AM-2k (Li et al. 2022b), AIM (Li, Zhang, and Tao 2021), and RWP-636 (Yu et al. 2021)) with ours. Real-1K has large-scale image data and multiple object classes including portrait, liquid, glass, mesh, animal, plant, fruit, and so on. There are also mixed categories in the same images from Real-1K. Our Real-1K could serve as a new challenging benchmark in the image matting area. We propose the first large-scale UHD (Ultra High Definition) natural image matting test set Real-World Image Matting-1K (Real-1K), which contains 1000 ultra high-resolution (from 4K to 8K) real-world natural samples of transparent and non-transparent attributes. Table 1 shows comparisons between some existing image matting datasets (DAPM (Shen et al. 2016), Adobe (Xu et al. 2017),) with ours. Real-1K has large-scale image data and multiple object classes including portrait, liquid, glass, mesh, animal, plant, fruit, and so on. There are also mixed categories in the same images from Real-1K. Our Real-1K could serve as a new challenging benchmark in the image matting area.

4.2 Composition Datasets

Adobe Matting Dataset (Xu et al. 2017). The training set consists of 431 foreground objects and each of them is composited over 100 random COCO (Lin et al. 2014) images to produce 43.1k composited training images. For the test set, we first composite each foreground from the test set with 20 random VOC (Everingham et al. 2010) images to produce 1k composited testing images (Composition-1K). Then we split Composition-1K into two groups (240 and 760 images, respectively) based on the critical attributes of transparent and non-transparent.

Distinction-646 (Qiao et al. 2020). It includes 596 and 50 foreground objects in training and test sets, respectively. We enforce the same rule, composited ratio, and grouping style with the AIM datasets that split the test set into two groups consisting of 200 and 800 images, respectively.

For testing on Real-1K, we train all models on the combined training set of Adobe (Xu et al. 2017) and Distinction-646 (Qiao et al. 2020).

4.3 Implementation Details

We implement the privileged prior information distillation on different matting models, including both trimap-based and trimap-free ones to evaluate the general applicability of our PPID framework. The teacher and the student use the same models but differ only in the inputs. All the experiments are conducted with Pytorch(Paszke et al. 2019).

4.4 Comparative Study

We conduct a comparative study on Real-1K and two composition benchmarks: Adobe Image Matting (Xu et al. 2017) and Distinction-646 (Qiao et al. 2020). We report mean square error (MSE), sum of the absolute difference (SAD), spatial-gradient (Grad), and connectivity (Conn) between predicted and ground truth alpha mattes. Lower values of these metrics indicate better-estimated alpha matte. To fairly compare, the metrics are computed on the entire image.

Attribute			Transparent				Non-transparent				Whole Test set			
Methods	**Trimap**	**Param**	SAD	MSE	Grad	Conn	SAD	MSE	Grad	Conn	SAD	MSE	Grad	Conn
DIM	✓	25.58M	146.29	0.032	90.37	68.91	47.68	0.012	44.35	31.62	86.73	0.020	62.57	46.39
INet	✓	5.95M	112.80	0.025	82.76	57.90	42.16	0.008	40.05	28.66	70.13	0.015	56.96	40.24
GCA	✓	25.13M	103.29	0.021	79.36	51.25	40.39	0.007	37.95	25.70	65.30	0.013	54.35	35.82
M-F_b	✓	44.80M	95.32	0.018	65.17	40.62	32.62	0.005	31.13	20.47	57.45	0.010	44.61	28.45
LFM	✓	37.91M	104.71	0.022	79.92	55.68	45.63	0.010	43.55	29.46	69.03	0.015	57.95	39.84
Hatt	✓	34.26M	107.58	0.024	81.93	58.21	38.24	0.007	36.65	23.53	65.70	0.014	54.58	37.26
AIM	✓	76.06M	102.37	0.021	75.38	55.27	35.71	0.006	34.49	21.88	62.11	0.012	50.68	35.10
DIM	-	25.58M	206.52	0.072	172.50	103.24	135.29	0.030	88.93	61.85	163.50	0.047	122.02	78.24
INet	-	5.95M	182.63	0.068	159.92	91.04	109.43	0.025	80.51	58.67	138.42	0.041	111.96	71.49
GCA	-	25.13M	190.05	0.068	162.33	98.37	104.96	0.023	80.22	56.31	138.66	0.041	112.74	72.97
M-F_b	-	44.80M	182.41	0.059	138.72	87.86	63.12	0.015	56.72	36.11	110.35	0.032	89.19	56.60
LFM	-	37.91M	181.37	0.058	126.09	87.23	68.23	0.016	60.38	38.78	113.03	0.033	86.40	57.97
Hatt	-	34.26M	175.37	0.053	120.75	83.98	58.35	0.013	50.62	34.77	104.69	0.029	78.39	54.26
AIM	-	76.06M	157.68	0.047	112.61	75.41	52.37	0.012	46.75	31.20	94.07	0.026	72.83	48.71
DIM-PPID	-	29.12M	152.35	0.039	98.62	57.22	70.39	0.017	62.49	42.41	102.85	0.026	76.80	48.27
INet-PPID	-	6.47M	114.78	0.026	82.69	60.43	44.76	0.009	42.68	29.09	72.49	0.016	58.52	41.50
GCA-PPID	-	26.60M	147.58	0.033	96.24	68.03	46.12	0.010	42.95	30.80	86.30	0.019	64.05	45.54
M-F_b-PPID	-	49.24M	106.50	0.023	80.91	57.09	41.94	0.008	39.25	26.45	67.50	0.014	**55.75**	**38.58**
LFM-PPID	-	40.68M	141.63	0.032	88.07	65.92	52.65	0.012	44.72	32.31	87.89	0.020	61.89	45.62
Hatt-PPID	-	38.01M	116.64	0.027	83.26	60.72	**39.25**	**0.008**	**38.28**	**25.89**	69.90	0.016	56.09	39.68
AIM-PPID	-	82.43M	**105.33**	**0.023**	**76.42**	**55.90**	45.37	0.009	43.01	30.60	**69.11**	**0.015**	56.24	40.62

Table 2: The quantitative results on our Real-1K.

Baselines	SD	$CLSD$	SAD	MSE	Grad	Conn
DIM			96.92	0.019	58.87	75.24
	✓		92.61	0.016	57.69	70.42
		✓	93.80	0.014	53.46	69.70
IndexNet			67.12	0.012	38.28	60.31
	✓		59.37	0.009	34.56	56.69
		✓	48.35	0.006	27.39	41.70
GCA			72.02	0.012	43.25	54.49
	✓		62.64	0.011	41.04	54.31
		✓	57.27	0.009	36.90	53.83

Table 3: Ablation study of the cross-layer semantic distillation (CLSD) module on Adobe dataset (Xu et al. 2017).

Baselines	LD	ALD	SAD	MSE	Grad	Conn
DIM			120.08	0.027	80.76	109.25
	✓		60.19	0.020	45.56	66.24
		✓	51.68	0.015	30.40	51.75
IndexNet			70.59	0.025	69.88	70.56
	✓		58.72	0.021	43.65	61.95
		✓	55.39	0.017	39.02	56.11
GCA			114.37	0.027	78.14	108.98
	✓		56.02	0.018	37.25	57.40
		✓	51.67	0.014	28.62	49.28

Table 4: Ablation study of the attention-guided local distillation (ALD) module on Adobe dataset (Xu et al. 2017).

We summarize the performance comparison in the above two attribute groups among our trimap-free models (with or w/o the privileged prior information distillation) and their trimap-based teachers ($i.e.$ DIM, INet (Lu et al. 2019), GCA (Li and Lu 2020), MatteFormer Baseline (M-F_b) (Park et al. 2022), LFM (Zhang et al. 2019), Hatt (Qiao et al. 2020), and AIM (Li, Zhang, and Tao 2021)). The trimap-based variants of LFM, Hatt, and AIM share the same network structures with their original models and only replace the input with image plus trimap. We follow the common setting of trimap-based methods to train these variants. For methods without publicly available codes, we follow their papers to reproduce the results with due diligence.

Table 2 shows the quantitative results of trimap-based teachers, trimap-free baselines, and our PPID-guided models on our Real-1K. We notice that our PPID framework significantly improves the performance of trimap-free baseline models in region decoupling, especially on the transparent attribute of both natural real-world and composited images. Although training on a limited amount of composited data, our PPID-guided models can further maintain a slight margin with the trimap-based teachers on real-world images, while the original trimap-free methods may not. It demonstrates that the PPID framework can help students mine more environment-aware information similar to that obtained from environmental priors. PPID can also reduce the weak generalization of trimap-free methods across data domains caused by the gap in data distribution between composited and natural images.

Compared to the existing SOTA methods, we observe that lighter-weighted trimap-free baseline models can achieve more significant performance improvement based on effective environment-aware information complements and guided local attribute optimization from our PPID. Particularly, our IndexNet-PPID gets the most significant performance boost among all competing baseline models, both on transparent ($e.g.$ MSE $61.8\% \downarrow$ in Table 2) and non-transparent ($e.g.$ MSE $64.0\% \downarrow$ in Table 2) attributes.

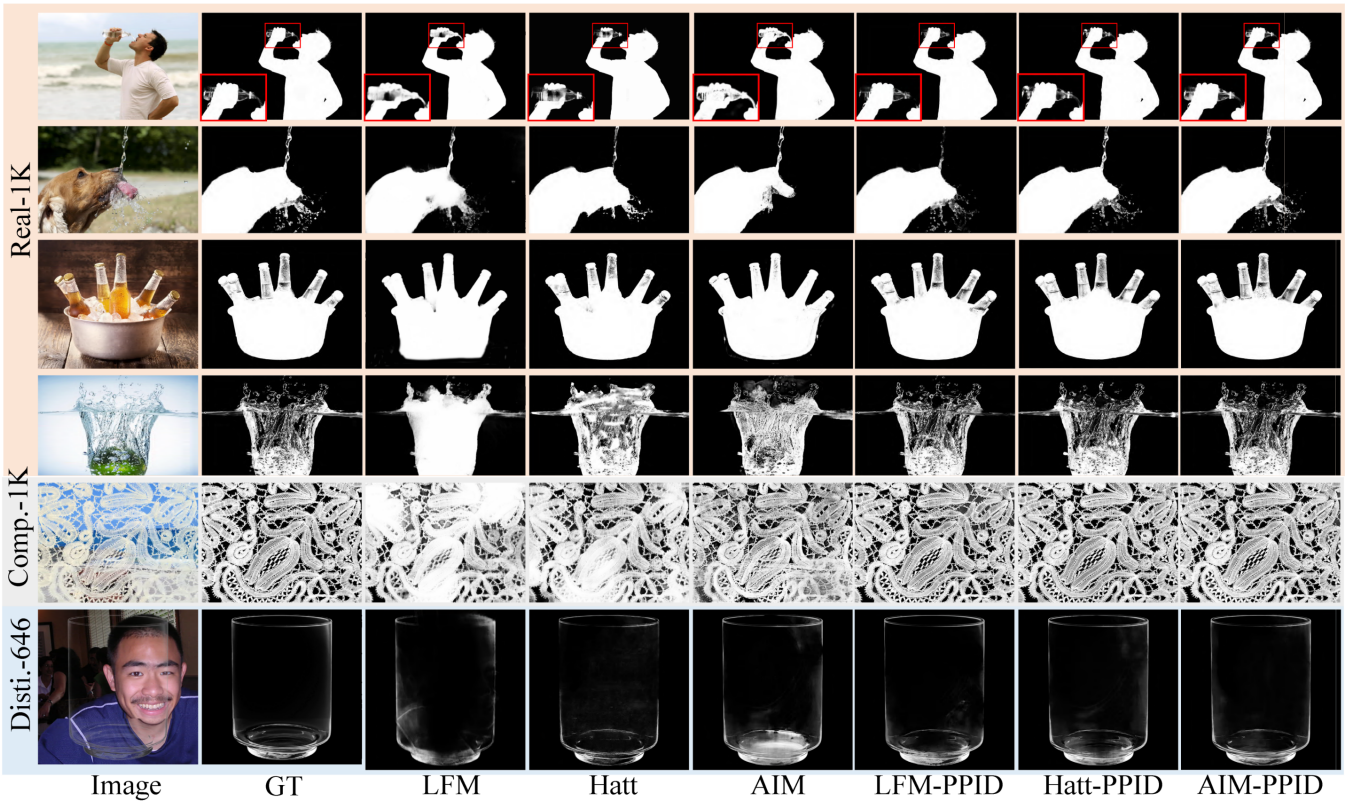

Figure 3: Visual comparisons of trimap-free baselines (LFM, Hatt, AIM) and PPID-guided models (LFM-PPID, Hatt-PPID, and AIM-PPID) on our Real-1K and public composition datasets (Comp.-1K and Dist.-646).

4.5 Ablation Study

To validate the effectiveness of our key components in privileged information distillation, we conduct ablation study under the following settings: (a) $CLSD$: cross-layer semantic distillation (CLSD); (b) SD: semantic distillation is only performed between the same layers; (c) ALD: attention-guided local distillation; (d) LD: local distillation w/o attention guidance.

CLSD $vs.$ **SD.** We report quantitative comparison results of our privileged information distillation on both transparent and non-transparent groups of the Composition-1K (Xu et al. 2017) dataset, with and without the privileged semantic distillation components (CLSD or SD). As summarized in Table 3, either CLSD or SD can significantly contribute to environmental awareness enhancement of the trimap-free baselines, and the alpha prediction performance is further improved on both transparent and non-transparent groups by introducing the CLSD module. That is because the CLSD mechanism can complement more sufficient environment-aware information by guiding each student layer to additionally mine higher-level semantic context.

ALD $vs.$ **LD.** We perform quantitative comparisons on our PPID-guided models with or without the ALD module under three settings, $i.e.$ w/o ALD, with LD, and with ALD. To evaluate the effectiveness of the ALD module on local attribute optimization, we compute the metrics on the transition region of each image in AIM (Xu et al. 2017). As shown in Table 4, the proposed local distillation modules (ALD and LD) improve the performance in local detail predictions, it offers further gain after introducing the spatial attention guidance that forces the students to focus on the crucial pixels. Additionally, we demonstrate the significant performance gain after combining CLSD and ALD. Some representative visualizations are provided in Fig. 3, which also illustrate the effectiveness of our PPID-IM for trimap-free baselines, especially in the scenarios with foregrounds that are semantically ambiguous (Row 6), chromaless (Row 1 to 4), or irregular (Row 5).

5 Conclusion

In this paper, we propose a privileged prior information distillation framework (PPID), that aims to effectively transfer privileged prior information from the trimap-based teachers to their poor environment-aware trimap-free student models. We also introduce a Cross-Level Semantic Distillation (CLSD) module that complements the student networks with both environmental awareness and higher-level semantic feature representations, for facilitating the cross-modality information distillation. Further, an Attention-Guided Local Distillation (ALD) is proposed to guide local region optimization for the students by efficiently transferring privileged local attributes and crucial information distribution from the teachers. Extensive experiments demonstrate the effectiveness and superiority of our PPID on image matting.

References

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12): 2481–2495.

Cai, S.; Zhang, X.; Fan, H.; Huang, H.; Liu, J.; Liu, J.; Liu, J.; Wang, J.; and Sun, J. 2019. Disentangled image matting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 8819–8828.

Cao, Y.; Xu, J.; Lin, S.; Wei, F.; and Hu, H. 2019. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 0–0.

Chen, Q.; Ge, T.; Xu, Y.; Zhang, Z.; Yang, X.; and Gai, K. 2018. Semantic human matting. In *Proceedings of the 26th ACM international conference on Multimedia*, 618–626.

Chen, Q.; Li, D.; and Tang, C.-K. 2013. KNN matting. *Proceedings of the IEEE transactions on pattern analysis and machine intelligence*, 35(9): 2175–2188.

Chen, X.; Zou, D.; Zhiying Zhou, S.; Zhao, Q.; and Tan, P. 2013. Image matting with local and nonlocal smooth priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1902–1907.

Cho, D.; Tai, Y.-W.; and Kweon, I. 2016. Natural image matting using deep convolutional neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, 626–643. Springer.

Ehsan, S.; Deepu, R.; Brian, P.; and Scott, C. 2013. Improving image matting using comprehensive sampling sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 636–643.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

He, K.; Sun, J.; and Tang, X. 2010. Fast matting using large kernel matting laplacian matrices. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2165–2172. IEEE.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Hou, Q.; and Liu, F. 2019. Context-aware image matting for simultaneous foreground and alpha estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4130–4139.

Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; and Wei, Y. 2018. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3588–3597.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Jue, W.; and Michael F, C. 2007. Optimized color sampling for robust matting. In *2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–8. IEEE.

Kaiming, H.; Christoph, R.; Carsten, R.; Xiaoou, T.; and Jian, S. 2011. A global sampling method for alpha matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2049–2056. IEEE.

Ke, Z.; Sun, J.; Li, K.; Yan, Q.; and Lau, R. W. 2022. MODNet: Real-Time Trimap-Free Portrait Matting via Objective Decomposition. In *AAAI*.

Lee, W.; Lee, J.; Kim, D.; and Ham, B. 2020. Learning with privileged information for efficient image super-resolution. In *European Conference on Computer Vision*, 465–482. Springer.

Levin, A.; Lischinski, D.; and Weiss, Y. 2007. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2): 228–242.

Li, B.; Wang, S.; Ye, H.; Gong, X.; and Xiang, Z. 2022a. Cross-Modal Knowledge Distillation for Depth Privileged Monocular Visual Odometry. *IEEE Robotics and Automation Letters*, 7(3): 6171–6178.

Li, J.; Zhang, J.; Maybank, S. J.; and Tao, D. 2022b. Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision*, 1–21.

Li, J.; Zhang, J.; and Tao, D. 2021. Deep Automatic Natural Image Matting. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 800–806. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Li, Y.; and Lu, H. 2020. Natural image matting via guided contextual attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11450–11457.

Lin, S.; Yang, L.; Saleemi, I.; and Sengupta, S. 2022. Robust High-Resolution Video Matting with Temporal Guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 238–247.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European conference on computer vision (ECCV)*, 740–755. Springer.

Liu, Y.; Xie, J.; Shi, X.; Qiao, Y.; Huang, Y.; Tang, Y.; and Yang, X. 2021. Tripartite information mining and integration for image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7555–7564.

Lopez-Paz, D.; Bottou, L.; Schölkopf, B.; and Vapnik, V. 2015. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*.

Lu, H.; Dai, Y.; Shen, C.; and Xu, S. 2019. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3266–3275.

Lutz, S.; Amplianitis, K.; and Smolic, A. 2018. AlphaGAN: Generative adversarial networks for natural image matting. In *British Machine Vision Conference (BMVC)*, 259. BMVA Press.

Park, G.; Son, S.; Yoo, J.; Kim, S.; and Kwak, N. 2022. Matteformer: Transformer-based image matting via prior-tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11696–11706.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Qiao, Y.; Liu, Y.; Yang, X.; Zhou, D.; Xu, M.; Zhang, Q.; and Wei, X. 2020. Attention-Guided Hierarchical Structure Aggregation for Image Matting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shen, X.; Tao, X.; Gao, H.; Zhou, C.; and Jia, J. 2016. Deep automatic portrait matting. In *Proceedings of the European conference on computer vision (ECCV)*, 92–107. Springer.

Vapnik, V.; Izmailov, R.; et al. 2015. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16(1): 2023–2049.

Wang, K.; and Chen, X. 2020. PMD-Net: Privileged Modality Distillation Network for 3D Hand Pose Estimation from a Single RGB Image. In *BMVC*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xu, N.; Price, B.; Cohen, S.; and Huang, T. 2017. Deep image matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2970–2979.

Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; and Yuan, C. 2022. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4643–4652.

Yu, Q.; Zhang, J.; Zhang, H.; Wang, Y.; Lin, Z.; Xu, N.; Bai, Y.; and Yuille, A. 2021. Mask guided matting via progressive refinement network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1154–1163.

Yung-Yu, C.; Brian, C.; David H, S.; and Richard, S. 2001. A bayesian approach to digital matting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 264–271. IEEE.

Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; and Agrawal, A. 2018. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 7151–7160.

Zhang, Y.; Gong, L.; Fan, L.; Ren, P.; Huang, Q.; Bao, H.; and Xu, W. 2019. A late fusion CNN for digital matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7469–7478.

Zhao, P.; Xie, L.; Zhang, Y.; Wang, Y.; and Tian, Q. 2020. Privileged knowledge distillation for online action detection. *arXiv preprint arXiv:2011.09158*.