Unifying Automatic and Interactive Matting with Pretrained ViTs

Zixuan Ye Wenze Liu He Guo Yujia Liang Chaoyi Hong* Hao Lu Zhiguo Cao
School of AIA, Huazhong University of Science and Technology
{yezixuan,wzliu,hguo01,yjl,cyhong,hlu,zgcao}@hust.edu.cn

Abstract

Automatic and interactive matting largely improve image matting by respectively alleviating the need for auxiliary input and enabling object selection. Due to different settings on whether prompts exist, they either suffer from weakness in instance completeness or region details. Also, when dealing with different scenarios, directly switching between the two matting models introduces inconvenience and higher workload. Therefore, we wonder whether we can alleviate the limitations of both settings while achieving unification to facilitate more convenient use. Our key idea is to offer saliency guidance for automatic mode to enable its attention to detailed regions, and also refine the instance completeness in interactive mode by replacing the binary mask guidance with a more probabilistic form. With different guidance for each mode, we can achieve unification through adaptable guidance, defined as saliency information in automatic mode and user cue for interactive one. It is instantiated as candidate feature in our method, an automatic switch for class token in pretrained ViTs and average feature of user prompts, controlled by the existence of user prompts. Then we use the candidate feature to generate a probabilistic similarity map as the guidance to alleviate the over-reliance on binary mask. Extensive experiments show that our method can adapt well to both automatic and interactive scenarios with more light-weight framework. Code available at github.com/coconut/SMat.

1. Introduction

Image matting, as a fundamental task in computer vision with wide application scenarios, has developed rapidly in deep-learning era. As a milestone work for deep image matting, DIM [30] first introduces an end-to-end network to solve α from an ill-posed equation. Following the paradigm of DIM, subsequent work [7, 16, 19–21, 28] also takes trimap as additional input towards high-quality alpha mattes. While there is unanimity that trimap can significantly enhance precision, this labour-consuming guidance is not desired. Therefore, some researchers replace trimap with more easy interactions like click [29] and scribble [32]. Based on the simplified interaction, they further explore the possibility of object selection, then interactive matting emerges. The success of SAM [9] lifts interactive matting to a new level by providing an intermediate mask result given interactions. Instead of simplifying the interactions, another stream of work tries to entirely discard the requirement of guidance, delving into automatic matting [3, 12, 13, 25, 36]. The two types of methods have their own focuses which leads to different strengths. As shown in Fig. 1, due to the effort made to find salient objects, automatic methods have better global awareness, and are easier to depict complete instances than interactive methods do (the feet of bird), but fail in detail compensation (tail feather); while interactive methods user prompts thus easily focus on details, but loses completeness caused by over-reliance on binary intermediate mask guidance (Fig. 2). Also, when facing different applications, we have to turn to different models, where the model switch will additionally increase heavy workload and inconvenience. Therefore, to combine the advantages of two modes and facilitate more convenient use, we desire *a unified solution for interactive matting and automatic matting and overcome their certain limitations.*

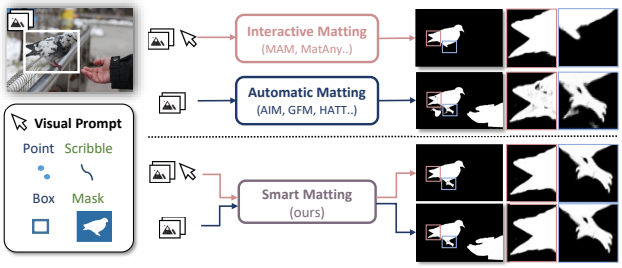

Figure 1. **A path to unify automatic and interactive matting while addressing the weaknesses.** Previous methods address the two types of image matting in separate structures, and show sub-optimal results on either object completeness or region details. We propose a unified structure for more convenient use, and address their certain weakness for better performance.

*Corresponding author.

The unification, however, is difficult to achieve, due to the challenge in designing a framework compatible with both automatic and interactive matting. Automatic matting cannot accommodate interactions, while interactive matting fails to give up its reliance on interactions. Therefore, we desire a hub to receive both empty prompt and user prompt, then output a uniform intermediate representation. In our method, we achieve the hub with a **candidate feature mechanism**, where the input is the prompts (empty or user prompts) and the output is the average feature of the target instance, termed candidate feature. If given user prompt, we use mask average pooling function to extract the average feature in prompt area as candidate feature. For automatic mode, we employ the class token in ViT structure, which implicitly carries the saliency information and can be regarded as the average feature for salient objects. With saliency guidance, automatic mode can easily locate the object and then concentrate on details. With the candidate feature automatically adapting between prompt feature and class token, we achieve the unification.

However, current candidate feature somehow only reflects "sporadic feature" rather than "average feature", e.g., given a single point user prompt, the point feature is hard to represent the complete instance. Therefore, we interact the candidate feature with image features to enable global awareness through a cross-attention module. By computing similarity between updated candidate feature and image features, we can obtain similarity map indicating the target instance. In this step, features with strong semantics are required to ensure feature consistency inside an instance, therefore, we employ pretrained-ViT to function as feature extractor. With the similarity map, through a simple similarity-guided decoder, we can implement the transition from similarity map to alpha matte. It is worth noting that previous interactive methods utilize SAM to turn user prompts into a binary mask, then use the mask-guided decoder to generate alpha matte. However, it causes over-reliance (Fig. 2) on mask and also brings heavy complexity and computation. Instead, our similarity map guidance is more probabilistic to alleviate the over-reliance and more easy-to-get without requiring large models.

For the first time, we delve into the unification of automatic and interactive matting and present a Smart Matting pipeline (SMat) as the solution, which also addresses limitations in each mode. Extensive evaluation on various benchmarks, including AIM-500 [12], RefMatte [15], AM-2k [13] and P3M [11] show that SMat can achieve superior performance on two modes without the requirement of extra models. To sum up, our contributions are as follows:

- We address the weaknesses in two matting types by redesigning the guidance.
- We achieve the unification by a candidate feature mechanism, with a simple and light-weight framework.

2. Related Work

2.1. Interactive Matting

As a fundamental task in computer vision, image matting has attracted much research interest over the past years [2, 5, 10]. DIM first addresses the alpha matte prediction in an end-to-end network with trimap as auxiliary input. Following DIM [30], subsequent methods [7, 16, 20, 23, 33] modify the network design for better global and detail information capture and fusion, while the requirement for a trimap still remains. However, drawing a trimap is time-consuming, therefore prior arts replace it with other guidance, e.g., background [18, 27], coarse binary mask [24, 35] and click [29], which are all easier to obtain than a trimap. With diverse guidance, researchers move a step further, not only using the guidance to focus on the unknown region, but also to specify instances for interactive matting. UIM [32] unifies the different visual prompts like point, scribble and box, and shows promising results on composition dataset. The emergence of SAM [9] inspires MatAny [34] and MAM [14], both of which use SAM to turn user prompts into intermediate binary masks, then achieve mask-to-matte transition with a designed module or existing models. However, the direct use of the binary mask can cause over-reliance on the results of SAM, which limits the modifications of the transition. Therefore, we tend to change the guidance into a more probabilistic form to alleviate the strong restriction caused by binary form. Additionally, the employment of SAM increases the complexity and heavy computation, which deserves a re-consideration.

2.2. Automatic Matting

Compared with simplifying the guidance, another stream of matting approaches tries to entirely remove the interaction from network, aiming for automatic matting. Due to the difficulty of finding the correct object, it starts from category-specific matting, e.g., SHM [3], PPM [11] and MODNet [8] for human matting and GFM [13] for animal matting. To remove the category restrictions, subsequent works extend category-specific to natural. Although there is no explicit definition for target object for natural images, AIM [12] and Deora et al. [4] implicitly extract the salient object by pretraining with Salient Object Detection (SOD) dataset or directly using the SOD model as priority respectively. Even with these designs, the generalization ability of these methods is still limited. Therefore, we desire more stable guidance that can be self-contained in the model instead of using other extra SOD models.

2.3. Pretrained ViTs

The emergence of ViT inspires great changes in computer vision, fueling increased interest in how to pre-train ViT to extract ideal feature representations for downstream tasks,

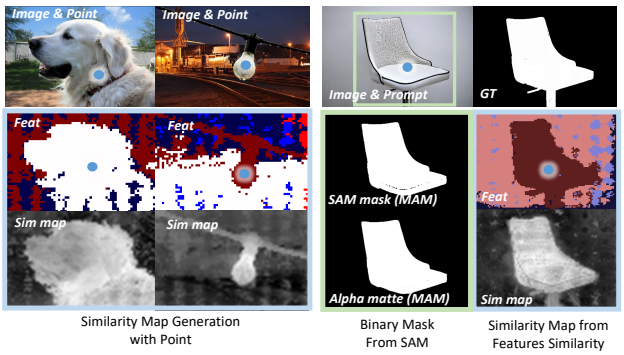

Figure 2. **Comparison of different ways for guidance generation.** For Matting Anything Model (MAM), we use box mode (best performance) to predict mask and alpha matte. For our similarity map generation, we use a point feature to perform similarity computation with image features. The probabilistic manner carries more instance information without the requirement of heavy models to get a binary mask.

such as MAE [6], DINO [1] and DINOv2 [22], and how to better utilize the ViT feature in downstream tasks. For instance, ViTDet [17] showcases how to benefit object detection with a plain ViT structure. In image matting, ViTMatte [33] demonstrates how to adapt plain ViT to matting and how to design the decoder based on ViT feature, MatAny and MAM utilize ViT by employing SAM [9] as priory model. In our method, we not only employ the good semantics in ViT features, but also think outside of the box, using the class token as the saliency cue to guide the salient object location for automatic mode.

3. Smart Matting: Unified Automatic and Interactive Matting

In this section, we illustrate how to establish a unified structure while addressing the respective weaknesses in a step-by-step manner. We begin by analyzing the rationale behind adopting a probabilistic similarity map as guidance instead of a binary mask, and introduce the candidate feature mechanism for unification (Sec. 3.1). Then, we introduce how to model candidate feature in different cases, i.e., automatic (Sec. 3.2) and interactive (Sec. 3.3). With candidate feature, we show how to generate and update the similarity map (Sec. 3.4), and how to obtain alpha matte with the similarity map (Sec. 3.5). To enhance instance perception, we introduce a foreground duplication strategy during the generation of training samples (Sec. 3.6).

3.1. Generate Guidance with Candidate Feature

It is common sense that interactive matting always gains higher performance than automatic matting, which owes much to the guidance it receives, either indicating unknown regions or prompting the target instance. For a matting model, the guidance can help alleviate the pressure for the target location, such that the model could focus on transparency prediction.

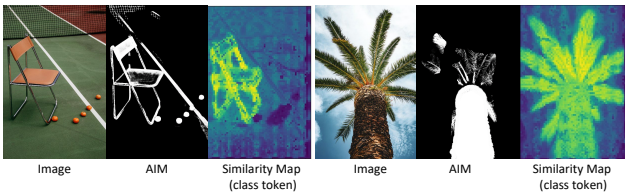

Figure 3. **The results from AIM and similarity map with class token as candidate feature.** The class token in ViT structure implicitly carries the salient information within an image.

Considering that mask-level can provide stronger guidance than sparse ones, after the emergence of SAM, interactive methods choose to convert the sparse interactions into binary masks with SAM. However, implementing such a sparse-to-dense transformation with SAM is complex and unnecessary. First, using a binary mask as guidance will cause over-reliance, which limits the modification of subsequent mask-to-matte modules. As shown in Fig. 2, due to the high confidence paid to SAM results, MAM can only make minor modifications instead of checking leaks and filling the vacancy, therefore the binary mask is somehow the restriction. Second, the utilization of SAM always leads to model tandem (MatAny uses SAM, DINO, and ViTMatte in sequence) and introduces higher computation.

Motivated by this, we seek more probabilistic guidance that carries more information, reflected in telling the probability of each point belonging to the instance rather than an absolute decision. Therefore, we present the notion of **candidate feature**, which represents the average feature of the target instance and is adaptable in different scenarios. By computing the similarity between candidate feature and image features, we can obtain a similarity map highlighting the instance region with possibility. As shown in Fig. 2, by setting the point feature as candidate feature, the obtained similarity map shows probabilistic instance region. Compared with the binary mask, the probabilistic map carries more information for subsequent modules for refinements.

Therefore, we formulate the guidance as the similarity between candidate feature and image features, i.e.,

$$F_{\text{sim}} = \text{SIM\_FUNC}(F_{\text{can}}, F_{\text{image}}), \tag{1}$$

where $F_{\text{can}} \in \mathcal{R}^{B\times C\times 1\times 1}$ denotes the candidate feature to be computed similarity with $F_{\text{image}} \in \mathcal{R}^{B\times C\times (H/S)\times (W/S)}$, and S stands for the strides when dividing image into patches.

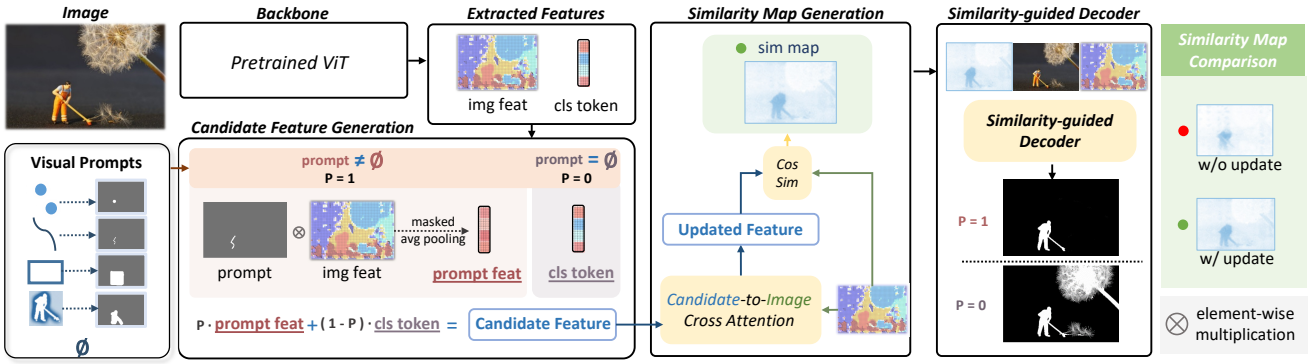

Figure 4. **The pipeline of our Smart Matting (SMat).** We obtain image features and class token from pretrained-ViT, and generating prompt feature by mask average pooling. Then, we merge the prompt feature and class token into candidate feature, the weights is controlled by the existance of visual prompts. By updating candidate feature with global interaction, we generate similarity map by computing the similarity between candidate feature and image features. Through a simple similiarity-guided decoder, we implement the transition from similarity map to alpha matte.

3.2. Employ Class Token as Saliency Guidance

As aforementioned, automatic matting lacks extra guidance, which hinders its concentration on transparency prediction. Therefore, we require a stable candidate feature for automatic mode, which is capable of understanding the meaning of the ideal object in automatic matting. In automatic matting, although the properties of ideal objects are not specified clearly, much effort has been made to recognize the salient objects inside an image. AIM [12] adopts Salient Object Detection (SOD) dataset to pretrain the matting model, which empowers the network to quickly locate the salient objects, while Deora et al. [4] directly use the results of SOD model as the prior information. However, as shown in Fig. 3, such approaches also suffer from poor generalization ability. This naturally raises a question: how can we capture saliency in a more robust and sound way?

Fortunately, we find that class token, a special design in ViT structure which is used for classification, can help locate the salient object in images. Since the class assignment to images always relies on the most salient object, therefore as the important feature for classification, the class token can be viewed as average feature of the salient object. As shown in Fig. 3, the salient object information included in class token can largely help with target searching. More importantly, it is self-contained in the ViT structure without the need for extra models and is easy to access. Therefore, we select the class token as the candidate feature to guide automatic matting by setting $F_{\text{can}} = F_{\text{cls\_token}}$.

3.3. Merge Prompt Feature as Instance Guidance

After confirming the candidate feature in automatic mode, we should consider how to design the candidate feature in interactive mode, to transform prompts of multiple forms to similarity map identifying the target objects. Given prompt of a single point, we can easily use the point feature as candidate feature, as shown in Fig. 2. For more complex prompts like scribble and mask, we select to adopt the average notion to the prompts. Specifically, as shown in Fig. 4, we treat all prompts as masks of the same resolution as image, and mark the regions covered by visual prompts by assigning a value of 1 then downsample it to $M_{\text{prompt}} \in 0$ or 1 with the resolution of F_{image}. Then we take out all the point features with mask value of 1 and calculate the average feature as the candidate feature, and compare it with image features to obtain the similarity map, i.e., $F_{\text{can}} = \text{MEAN}(M_{\text{prompt}} \odot F_{\text{image}})$. The denominator in the MEAN function is the number of pixels belonging to the highlighted area instead of all pixels within an image.

3.4. Update Candidates with Global Awareness

With the adaptable candidate feature for two modes, we can easily obtain the similarity map F_{sim} through:

$$F_{\text{sim}} = \text{SIM\_FUNC}(F_{\text{can}}, F_{\text{image}}),$$
$$\begin{cases} F_{\text{can}} = F_{\text{cls\_token}}, & \text{prompt} = \varnothing \\ F_{\text{can}} = \text{MEAN}(M_{\text{prompt}} \odot F_{\text{image}}), & \text{prompt} \neq \varnothing \end{cases} \quad (2)$$

However, the candidate feature in interactive mode only contains local information, which is more like "sporadic feature" rather than "average feature". Moreover, considering the noise and inconsistency it may include, the similarity map will be sensitive to the positions of given prompts and thereby hinders its robustness. Consequently, the direct use of current candidate feature for similarity comparison poses challenge on generating an ideal similarity map identifying a complete instance.

Motivated by this, we suggest an updation to candidate feature, to turn "sporadic" into "average", the core of which lies in the interaction between candidate feature with global

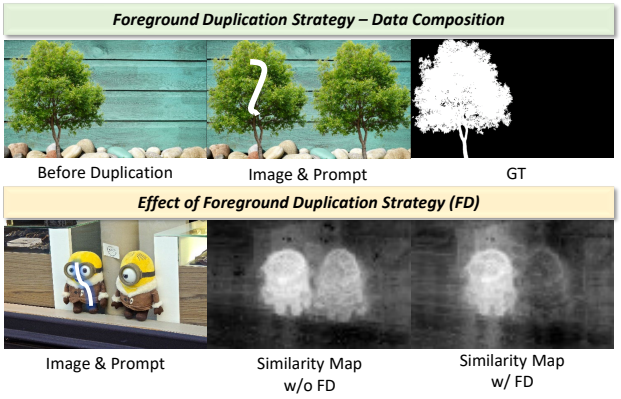

Figure 5. **Foreground Duplication Strategy and its effect.** The strategy can effectively help the difficult distinction for instances with the same semantics and similar appearance by enforcing the model to learn position difference.

features. Therefore, we introduce a cross-attention module to enable the interaction between candidate and image features. As shown in Fig. 4, the updated candidate feature owns a more stable performance and instance instruction capability. Moreover, the module can help bring the feature representation of class token and prompt features closer, which further benefits the unification of two modes.

3.5. Obtain Alpha Matte with Similarity Map

Through the candidate feature mechanism, we achieve Smart Matting with a unified structure, and the main pipeline is illustrated in Fig. 4. We first adopt a pretrained-ViT as backbone to extract image features and class token, then generate candidate feature by merging class token and prompt feature with different weights controlled by whether the prompt is empty, then it can adapt to saliency guidance and prompt guidance in different cases. With the updated candidate feature, we select COSINE_SIMILARITY as the similarity function to generate the similarity map. To realize a similarity-to-matte transition, we employ a simple and lightweight decoder in ViTMatte, with the concatenation of image features and similarity map as input, decode the features to alpha matte in a pyramid manner.

3.6. Enhance Instance Perception with Foreground Duplication

Since we distinguish the instance mainly by the semantics, then instances with same semantics and similar appearance will become hard cases, as shown in Fig. 5. Drawing lessons from the instance segmentation, we learn that the positional information is also a key-point for the distinction. In ViT structure, it naturally contains positional embedding when dividing images to patches, then what we should consider is how to amplify the role of position information in similarity comparisons. We take an extreme approach termed Foreground Duplication Strategy (FD), which duplicates the foreground in an image but sets only one target and enforces the network to learn the position difference. With the same appearance, the only way to distinguish the instance is to strengthen the positional awareness ability. Still in Fig. 5, one can see that the proposed FD strategy encourages a more precise and ideal similarity map.

4. Experiments

4.1. Implementation Details

Datasets We employ the same training set as AIM [12], which consists of Distinction-646 [25], AM-2k [13], Composition-1k [30], and DUTS [31], denoted as set 1. MAM additionally adopts RefMatte [15]; we define the entire set as set 2. However, we find the samples in RefMatte override the test set in other datasets, therefore, we only use set 1. Without multi-instance dataset, we need to generate multi-instance samples to enable the training of interactive matting. By randomly merging several unique foregrounds in set 1 into one image like RefMatte, we obtain 10000 samples for interactive matting, which re-utilizes the existing foregrounds rather than inviting new foreground patterns. We use the modified l_1 loss presented by ViTMatte [33] and l_{lap} to supervise the prediction.

Benchmarks and Metrics We test our method on a variety of benchmarks for image matting, including AIM-500 [12] for natural image matting, AM-2k [13] for animal matting, P3M [11] and RWP-636 [35] for human matting. We also evaluate on RefMatte-RW100 [15] to verify the ability for object switch in interactive matting. The quality of alpha matte is assessed with four standard metrics [26]: SAD, MSE, Grad and Conn; the lower value on four metrics denotes higher performance.

Training Details We use a mixed training strategy in our method, i.e., samples with prompt and without prompt randomly appear during training, where the only difference is the way to generate candidate feature. Only real-world image will be set as automatic samples, otherwise the composition image will interrupt the saliency capture ability. For the composition image, we take the common data augmentations following most of matting methods [16, 23], and randomly generate different prompts for the foregrounds. We perform foreground duplication with a possibility of 0.6, and only use box prompt for the duplicated foreground. We take a batch size of 32 with 50k iterations on two GTX3090 GPUs, AdamW is selected as the optimizer with a learning rate of 5e-4. We initialize the backbone with DINOv2, and set a decay rate of 0.01.

method	params (M)	mode automatic	AIM-500 SAD	MSE	MAD	Grad	Conn	mode interactive	AIM-500 SAD	MSE	MAD	Grad	Conn	RefMatte-RW100 SAD	MSE	MAD	Grad	Conn
LF[1] [36]	37.9	✔	191.74	0.0667	0.1130	64.51	181.26	✘										
HATT[1] [25]	107.0	✔	479.17	0.2700	0.2806	473.98	238.63	✘			/					/		
GFM[1] [13]	55.3	✔	52.66	0.0213	0.0313	46.11	52.69	✘										
AIM[1] [12]	55.3	✔	48.09	0.0183	0.0285	47.58	21.74	✘										
MatAny[0] [34]	363.2	✘			/			✔	124.36	0.0639	0.0753	37.12	21.84	52.91	0.0270	0.0293	25.17	5.13
MAM[2] [14]	96.4	✘						✔	42.62	0.0144	0.0258	**22.14**	22.00	29.23	0.0151	0.0166	25.85	**2.83**
SMat[1](ours)	26.9	✔	**34.30**	**0.0129**	**0.0203**	**31.49**	**13.98**	✔	**26.63**	**0.0083**	**0.0209**	33.03	**15.77**	**25.60**	**0.0120**	**0.0146**	**22.62**	5.31

Table 1. **Quantitative Results on AIM-500 and RefMatte (multi-instance) compared with other natural image matting methods.** [0] stands for no training process, [1] denotes training on AIM training set, and [2] represents additionally adding RefMatte to training set like MAM. Our method can provide automatic and interactive results while other methods can only address one scenario.

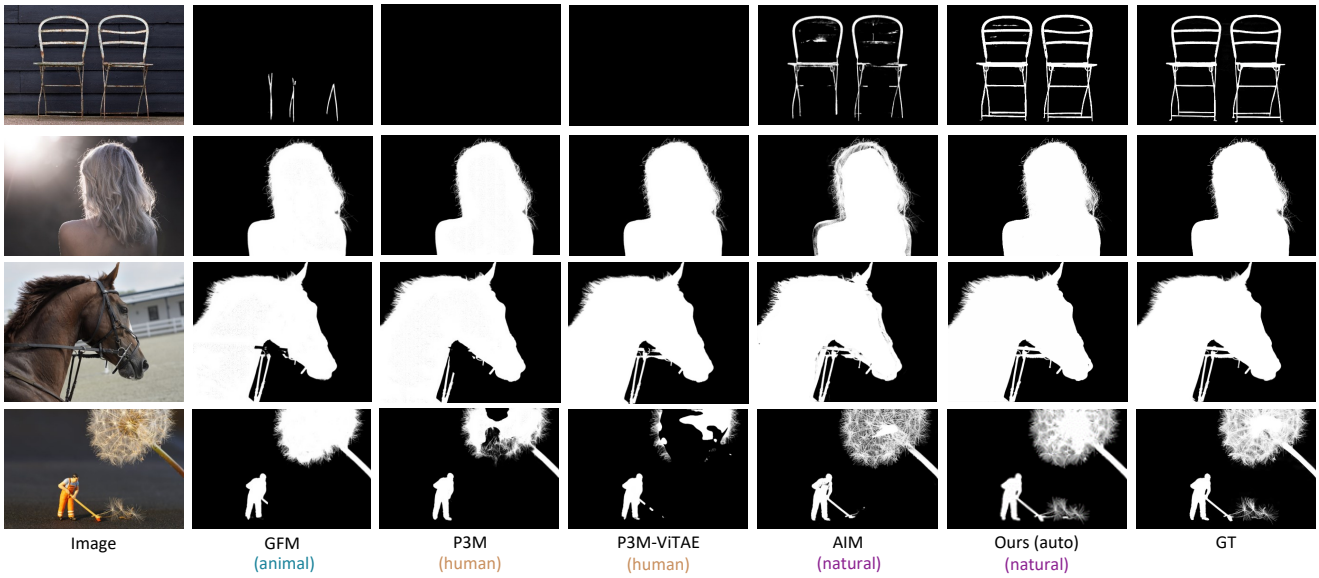

Figure 6. **Qualitative comparison with automatic methods.** Category-specific methods can only perform well on its target domain, while existing natural automatic matting method fail to provide ideal results. In contrast, our automatic mode experts in both saliency and details.

4.2. Main Results

We compare our methods with previous methods from three aspects: compatibility of different matting types, generalization ability of automatic mode on various categories and robustness to different interactions in interactive mode.

Compatibility of Automatic and Interactive Matting We first compare our Smart Matting (SMat) with other approaches in terms of compatibility. All natural automatic methods are trained with training set 1 in order to provide a fair comparison. Although GFM [13] is an animal matting approach, we re-train it with natural matting data set 1 to act as a natural method. Since MatAny [34] consists of three fixed models with no training process, we denote it as 0 which means no training set, while MAM uses training set 2. We report the results on natural matting dataset AIM-500 [12] and multi-object dataset RefMatte-RW-100 [15] in Table 1, note that the interactive performance here is the best behavior among different prompts, the complete form can refer to Table 3. As shown in Table 1, automatic matting methods cannot perform interactive object selection, while interactive methods fail to give automatic results. In contrast, our method can achieve ideal results in both automatic and interactive scenarios with a lightweight structure, the compatibility suggests that using the candidate feature mechanism for the unification is effective and reasonable.

Compared with other natural automatic methods, under the same training dataset, our method achieves the *state-of-the-art* performance with the fewest parameters. Our automatic mode surpasses the previous method AIM with an obvious 13.79 improvement on SAD and 29.5% relative improvement on MSE metric.

In interactive scenarios, since MatAny concatenates three models, the incorrect predictions of transparency from GroundingDINO can lead to substantial performance degradation which requires an extra user prompt to refine, therefore it fails to provide ideal predictions without manual re-

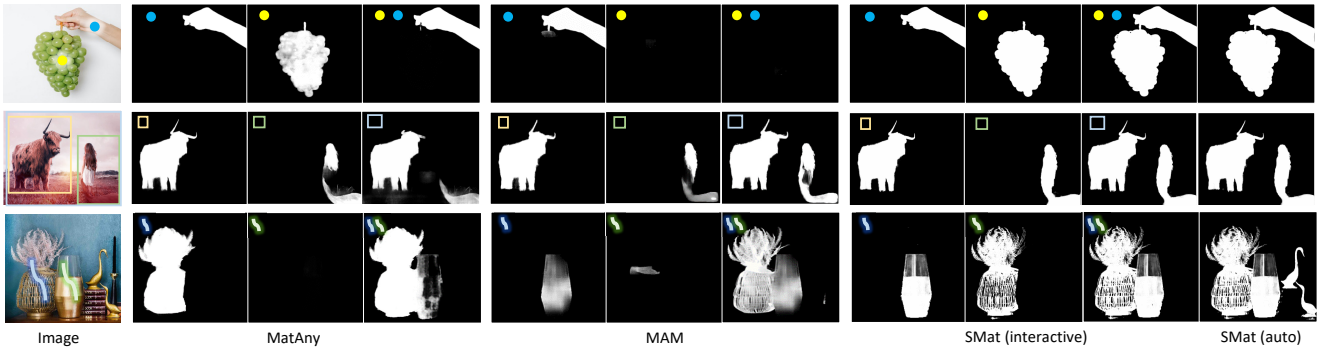

Figure 7. **Qualitative comparison with interactive methods.** Our method is more robust to different interactions and generates more appealing results.

automatic methods	category	AM2K (animal)		P3M-500-NP (human)		AIM-500 (natural)	
		SAD	MSE	SAD	MSE	SAD	MSE
GFM	animal	11.11	0.0031	111.98	0.0613	95.84	0.0505
PPM	human	23.06	0.0096	13.38	0.0042	97.36	0.0512
PPM-ViTAE	human	37.84	0.0189	7.80	0.0017	109.69	0.0584
AIM	natural	32.03	0.0124	65.57	0.0404	48.09	0.0183
SMat-auto	natural	16.84	0.0047	18.93	0.0064	34.30	0.0129

Table 2. **Quantitative comparison with different automatic matting methods.** The results are generated with official models provided by the authors. Category-specific models can only address limited scenarios, while natural image matting can generalize to all foreground types.

finement. Unlike MatAny, MAM only uses a mask-to-matte module after SAM for refinement, leading to large improvement. However, much of its strong performance can be attributed to SAM, which shows 33.51 SAD even without MAM module. Instead, our interactive mode achieve 25.60 SAD without leveraging the power of SAM, which alleviates the complexity and invites higher performance.

Generalization on class-specific automatic matting Since we target our automatic mode in natural rather than category-specific, it is important to know the generalization ability on different foreground types. As shown in Table 2, animal- and human-specific matting methods are like specialists, which can only excel at their domain, but are poor in other scenarios. In contrast, our method is more like an all-rounder; although it slightly falls behind the performance of specialists in their domain, it maintains stable performance on all scenarios and outperforms other methods on natural scenarios by a large margin where category-specific approaches cannot provide ideal results. Qualitative results are demonstrated in Fig. 6, where GFM and P3M generate poor predictions on natural images, and AIM brings discontinuity within the prediction.

interactive methods	prompt	P3M-500-NP			RefMatte-RW100		
		SAD	MSE	MAD	SAD	MSE	MAD
MatAny	point	139.09	0.0750	0.0813	63.99	0.0340	0.0363
	box	132.60	0.0730	0.0793	52.91	0.0270	0.0293
	box-u	-	-	-	85.51	0.0456	0.0482
MAM	point	211.54	0.0877	0.1177	614.34	0.3450	0.3489
	box	25.81	0.0920	0.0153	29.23	0.0151	0.0166
	box-u	-	-	-	32.74	0.0139	0.0188
SMat(ours)	point	52.48	0.0250	0.0304	25.60	0.0120	0.0146
	box	14.10	0.0037	0.0081	34.86	0.0172	0.0199
	box-u	-	-	-	25.96	0.0123	0.0148

Table 3. **Quantitative comparison with different interactive matting methods.** Box-u denotes user-provided boxes. Our method is more robust to different prompts, and invites higher performance without the requirement of SAM.

Robustness to different interactions We also compare the robustness to different prompts with other interactive methods, the results are reported in Table 3. For the point prompt, we provide ten points for each method to cover as many instance regions as possible. One can see that MAM is very sensitive to prompt types, while MatAny cannot provide an ideal prediction even with box prompt. Our SMat achieves more detailed results on easy cases, reflected in the SAD on P3M-500 using box prompt (14.10 vs 25.81 vs 132.60). It can also locate the target instance more precisely during object selection, reflected in MSE metric on RefMatte-RW100 (0.0120 with point prompt). We also manually label the box and conduct the evaluation. The manual label randomly zooms in and out the box to simulate the inaccurate box provided by users, the results demonstrate that our method can overcome the inaccurate prompt well. Since the GT box may completely cover multiple instances and cause confusion, while the manual label avoids this situation, our method generates better performance with user-provided box prompts than GT box.

method	params (M)	Speed(s) 512*512	FPS
LF	37.9	0.1103	9.07
HATT	107.0	0.0862	11.60
AIM	55.3	0.0165	60.61
MatAny	363.2	0.5408	1.85
MAM	96.4	0.2940	3.40
SMat-Auto.	26.9	0.0246	40.65
SMat-Inter.		0.0302	33.11

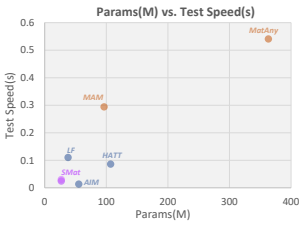

Table 4. **Efficiency comparison with other methods.** We compare the efficiency from parameters and inference speed. The results show that our model has both faster inference and a more lightweight structure.

	AIM-500			RWP-636		
	SAD	MSE	MAD	SAD	MSE	MAD
random embedding	413.86	0.2328	0.2431	634.61	0.3856	0.3971
learnable embedding	36.62	0.0133	0.0214	37.93	0.0145	0.0252
class token	34.30	0.0129	0.0203	35.34	0.0136	0.0245

Table 5. **The effect of class token for saliency guidance.** We conduct the experiments on natural dataset AIM-500 and human dataset RWP-636, the results show that using class token as saliency guidance can help automatic mode.

Efficiency Comparison with Other Methods Here we compare the efficiency following the setting which tests the average inference speed of processing a 512 $\times$ 512 image over 100 runs. The results are illustrated in Table 4. One can see that our method outperforms other interactive methods like MAM and MatAny in efficiency by a large margin, because our model does not rely on SAM and uses feature similarity to locate the target object. Also, our model has a more lightweight structure which only adopts the most straightforward design of the modules.

4.3. Ablation Study

Saliency guidance of class token To verify whether the class token can indeed help to find salient object, we conduct an ablation study on the candidate feature for automatic mode. We replace the class token with random embedding and learnable embedding, the results are shown in Table 5. With random embedding, the model fails to locate the object, leading to 413 SAD. Learnable embedding can owns the ability for saliency capture, however, it still falls behind the performance of class token with 2.32 SAD.

Generate, update and accurate the guidance We also explore the effect of our specific designs on guidance. First, we compare the binary and probabilistic form of guidance. We binary the similarity map with a threshold of 0.5. The results shown in Table 6 show that the probabilistic similarity map carries more information than a direct decision, with an obvious performance boost on MSE metric. Based on the initial similarity map, we further probe the effect of the updation process. Comparing B2 with B3, an updation to the candidate feature can introduce a 7.81 improvement on SAD. The advantage of FD strategy is reflected in the 4.6% relative error reduction on MSE, which indicates an accurate location of instance.

	Generate		Update	Accurate	RefMatte-RW100		
	binary	prob.	cross_attn	FD	SAD	MSE	MAD
B1	✔				231.28	0.1221	0.1312
B2		✔			44.18	0.0223	0.0252
B3		✔	✔		36.37	0.0181	0.0216
B4		✔		✔	42.47	0.0210	0.0242
B5		✔	✔	✔	34.86	0.0172	0.0199

Table 6. **Ablation study on generating, updating and accurating the guidance.**

	AIM-500			RefMatte-RW100		
	prompt	SAD	MSE	prompt	SAD	MSE
automatic	none	35.32	0.0130	box	354.05	0.1987
interactive	none	363.45	0.2042	box	35.68	0.0177
mix	none	34.30	0.0129	box	34.86	0.0172

Table 7. **Synergy between interactive and automatic mode.**

Synergy between automatic and interactive modes To verify whether there exists a synergy between automatic and interactive matting, we train the model under only automatic setting and interactive setting separately; the results are illustrated in Table 7. By comparing Row 1 and Row 3, one can see that the unification can help boost the transparency prediction (0.98 SAD improvement), while maintaining a good location on salient object (0.0130 vs 0.0129 MSE). Also, the unification can strengthen the instance completeness during object selection, reflected in 2.8% relative improvement on MSE (Row 2 vs Row 3).

5. Conclusion

In this paper, we observe the contrary behavior of automatic and interactive matting and the inconvenience of model switching in different cases. For the limitations, we add saliency guidance for automatic mode and change binary mask guidance into probabilistic similarity maps. Then we achieve unification with a novel candidate feature mechanism, which adaptively switches between class token and prompt feature, decided by whether the prompt is empty, to enable saliency capture and instance selection for automatic and interactive mode. We also invite specific designs to accurate the similarity map like the foreground duplication strategy. Beyond the unification, we achieve superior performance on both automatic and interactive scenarios, and the slim network opens up more possibilities for real-world applications.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[2] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. KNN Matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2175–2188, 2013. 2

[3] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *Proceedings of ACM International Conference on Multimedia*, pages 618–626, 2018. 1, 2

[4] Rahul Deora, Rishab Sharma, and Dinesh Samuel Sathia Raj. Salient image matting. *arXiv preprint arXiv:2103.12337*, 2021. 2, 4

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

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

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

[8] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. Modnet: real-time trimap-free portrait matting via objective decomposition. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 1140–1147, 2022. 2

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 2, 3

[10] A. Levin, D. Lischinski, and Y. Weiss. A Closed-Form Solution to Natural Image Matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 228–242, 2008. 2

[11] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *Proceedings of ACM International Conference on Multimedia*, pages 3501–3509, 2021. 2, 5

[12] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep Automatic Natural Image Matting. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 800–806, 2021. 1, 2, 4, 5, 6

[13] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. *Int. J. Comput. Vis.*, 130(2):246–266, 2022. 1, 2, 5, 6

[14] Jiachen Li, Jitesh Jain, and Humphrey Shi. Matting anything. *arXiv preprint arXiv:2306.05399*, 2023. 2, 6

[15] Jizhizi Li, Jing Zhang, and Dacheng Tao. Referring image matting. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pages 22448–22457, 2023. 2, 5, 6

[16] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 11450–11457, 2020. 1, 2, 5

[17] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 280–296. Springer, 2022. 3

[18] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pages 8762–8771, 2021. 2

[19] Qinglin Liu, Haozhe Xie, Shengping Zhang, Bineng Zhong, and Rongrong Ji. Long-range feature propagating for natural image matting. In *Proceedings of ACM International Conference on Multimedia*, pages 526–534, 2021. 1

[20] Yuhao Liu, Jiake Xie, Xiao Shi, Yu Qiao, Yujie Huang, Yong Tang, and Xin Yang. Tripartite information mining and integration for image matting. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 7555–7564, 2021. 2

[21] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 3265–3274, 2019. 1

[22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[23] GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, and Nojun Kwak. Matteformer: Transformer-based image matting via prior-tokens. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pages 11696–11706, 2022. 2, 5

[24] Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Mask-guided matting in the wild. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pages 1992–2001, 2023. 2

[25] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pages 13676–13685, 2020. 1, 5, 6

[26] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pages 1826–1833, 2009. 5

[27] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pages 2291–2300, 2020. 2

[28] Jingwei Tang, Yagiz Aksoy, Cengiz Oztireli, Markus Gross, and Tunc Ozan Aydin. Learning-Based Sampling for Natural Image Matting. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pages 3050–3058, 2019. 1

[29] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Hanqing Zhao, Weiming Zhang, and Nenghai Yu. Improved image matting via real-time user clicks and uncertainty estimation. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pages 15374–15383, 2021. 1, 2

[30] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep Image Matting. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pages 311–320, 2017. 1, 2, 5

[31] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013. 5

[32] Stephen DH Yang, Bin Wang, Weijia Li, YiQi Lin, and Conghui He. Unified interactive image matting. *arXiv preprint arXiv:2205.08324*, 2022. 1, 2

[33] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pretrained plain vision transformers. *Information Fusion*, page 102091, 2023. 2, 3, 5, 1

[34] Jingfeng Yao, Xinggang Wang, Lang Ye, and Wenyu Liu. Matte anything: Interactive natural image matting with segment anything models. *arXiv preprint arXiv:2306.04121*, 2023. 2, 6

[35] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pages 1154–1163, 2021. 2, 5

[36] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pages 7469–7478, 2019. 1, 6