ZIM: Zero-Shot Image Matting for Anything

Beomyoung Kim Chanyong Shin Joonhyun Jeong Hyungsik Jung
Se-Yun Lee Sewhan Chun Dong-Hyun Hwang Joonsang Yu
NAVER Cloud, ImageVision

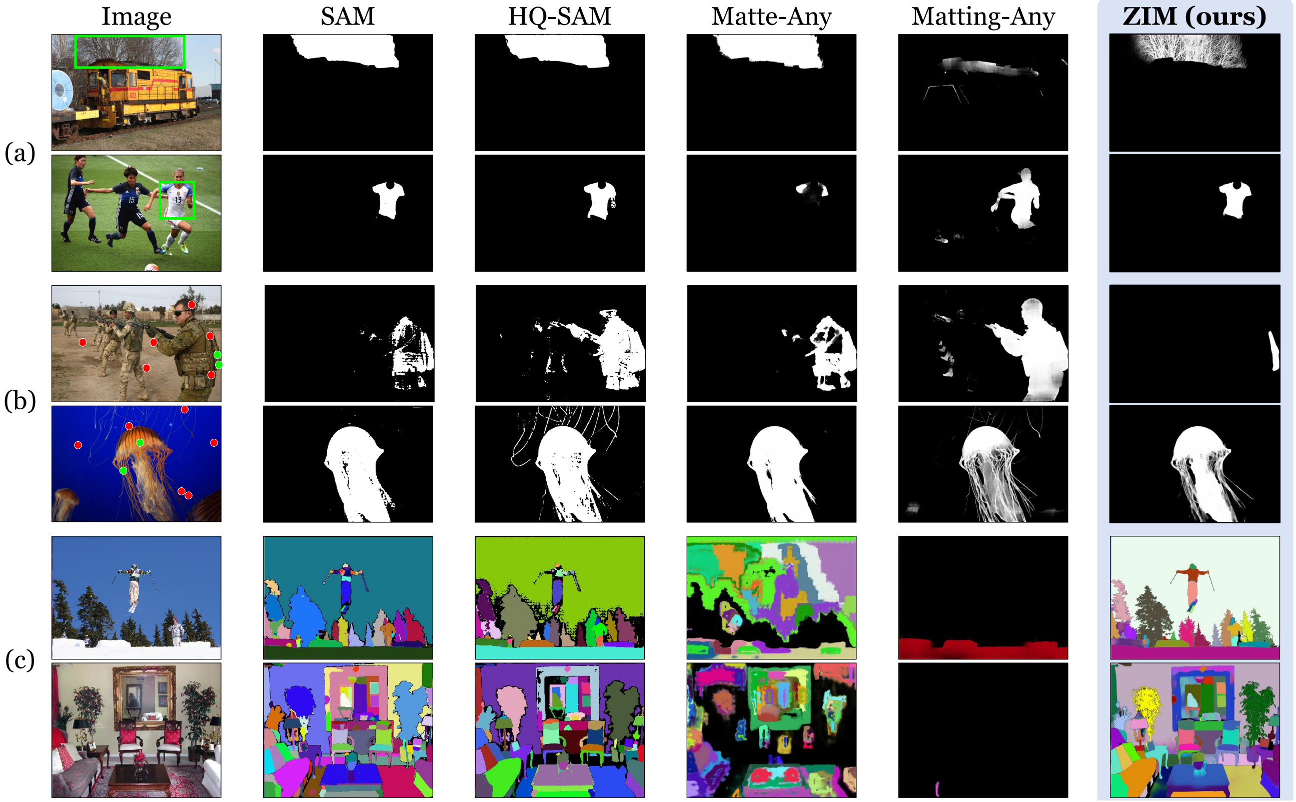


Figure 1. **Qualitative comparison** of ours with five existing zero-shot models (SAM [18], HQ-SAM [13], Matte-Any [50], and Matting-Any [25]). It showcases (a) box prompting results, (b) point prompting results, and (c) automatic mask generation results.

Abstract

The recent segmentation foundation model, Segment Anything Model (SAM), exhibits strong zero-shot segmentation capabilities, but it falls short in generating fine-grained precise masks. To address this limitation, we propose a novel zero-shot image matting model, called ZIM, with two key contributions: First, we develop a label converter that transforms segmentation labels into detailed matte labels, constructing the new SA1B-Matte dataset without costly manual annotations. Training SAM with this dataset enables it to generate precise matte masks while maintaining its zero-shot capability. Second, we design the zero-shot matting model equipped with a hierarchical pixel decoder to enhance mask representation, along with a prompt-aware masked attention mechanism to improve performance by enabling the model to focus on regions specified by visual prompts. We evaluate ZIM using the newly introduced MicroMat-3K test set, which contains high-quality micro-level matte labels. Experimental results show that ZIM outperforms existing methods in fine-grained mask generation and zero-shot generalization. Furthermore, we demonstrate the versatility of ZIM in various downstream tasks requiring precise masks, such as image inpainting and 3D segmentation. Our contributions provide a robust foundation for advancing zero-shot matting and its downstream applications across a wide range of computer vision tasks. The code is available at https://naver-ai.github.io/ZIM.

1. Introduction

Image segmentation, which divides an image into distinct regions to facilitate subsequent analysis, is a fundamental task in computer vision. Recent breakthroughs in segmentation models have made significant strides in this area, particularly with the emergence of the segmentation foundation model, Segment Anything Model (SAM) [18]. SAM is trained on the SA1B dataset [18] containing 1 billion micro-level segmentation labels, where its extensiveness enables SAM to generalize effectively across a broad range of tasks. Its strong zero-shot capabilities, powered by visual prompts, have redefined the state of the art in zero-shot interactive segmentation and opened new avenues for tackling more complex tasks within the zero-shot paradigm.

Despite these achievements, SAM often struggles to generate masks with fine-grained precision (see Figure 1). To address this limitation, recent studies [25, 50, 51] have extended SAM to the image matting task, which focuses on capturing highly detailed boundaries and intricate details such as individual hair strands. These approaches achieve enhanced mask precision by fine-tuning SAM on publicly available matting datasets [21, 36, 47]. However, this fine-tuning process can undermine the zero-shot potential of SAM, since most public matting datasets contain only macro-level labels (*e.g.*, entire human portrait) rather than the more detailed micro-level labels (*e.g.*, individual body parts), as illustrated in Figure 2. Fine-tuning with macro-level labels can cause SAM to overfit to this macro-level granularity, resulting in catastrophic forgetting of its ability to generalize at the micro-level granularity, as shown in Figure 1. Moreover, the scarcity of large-scale matting datasets with micro-level matte labels poses a significant obstacle in developing effective zero-shot matting solutions.

In this paper, we introduce a pioneering **Z**ero-shot **I**mage **M**atting model, dubbed **ZIM**, that retains strong zero-shot capabilities while generating high-quality micro-level matting masks. A key challenge in this domain is the need for a matting dataset with extensive micro-level matte labels, which are costly and labor-intensive to annotate. To address this challenge, we propose a novel label conversion method that transforms any segmentation label into a detailed matte label. For more reliable label transformation, we design two effective strategies to reduce noise and yield high-fidelity matte labels (Section 3.1). Subsequently, we construct a new dataset, called **SA1B-Matte**, which contains an extensive set of micro-level matte labels generated by transforming segmentation labels from the SA1B dataset via the proposed converter (see Figure 2). By training SAM on the SA1B-Matte dataset, we introduce an effective foundational matting model with micro-level granularity while preserving the zero-shot ability of SAM (see Figure 1).

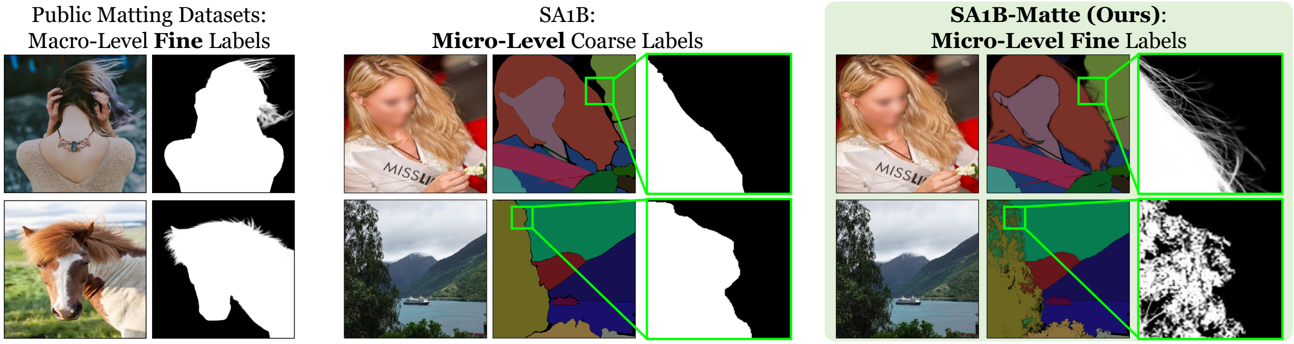
To further ensure effective image matting, we enhance the major bottleneck in the network architecture of SAM that impedes capturing robust and detailed feature maps. Specifically, SAM employs a simple pixel decoder to generate mask feature maps with a stride of 4, which is susceptible to checkerboard artifacts and often falls short in capturing fine details. To mitigate this, we design a more elaborated pixel decoder, enabling more robust and richer mask representations (Section 3.2). Furthermore, we introduce a prompt-aware masked attention mechanism that leads to the improvement of interactive matting performance.

To validate our zero-shot matting model, we present a new test set, called **MicroMat-3K**, consisting of 3,000 high-quality micro-level matte labels. Our experiments on this dataset demonstrate that while SAM exhibits strong zero-shot capabilities, it struggles to deliver precise mask outputs. In contrast, existing matting models show limited zero-shot performance. ZIM, however, not only maintains robust zero-shot functionality but also provides superior precision in mask generation. Additionally, we highlight the foundational applicability of ZIM in several downstream tasks requiring precise masks, such as image inpainting [54] and 3D segmentation [3]. We hope this work provides valuable insights to the research community, encouraging further development of zero-shot matting models.

2. Related Work

Image Segmentation. Image segmentation is a fundamental task in computer vision, enabling the division of an image into distinct regions. Recent advancements in segmentation models [6, 12, 19] have significantly improved the accuracy of segmentation tasks, including semantic, instance, and panoptic segmentation. The emergence of Segment Anything Model (SAM) [18] introduced a new paradigm in segmentation by leveraging visual prompts (*e.g.*, points or boxes). SAM is designed as a foundational segmentation model capable of handling diverse tasks due to its robust zero-shot capabilities, showing remarkable versatility across a wide range of tasks and domains. However, despite its strengths, SAM struggles to produce high-precision masks. In this paper, we address this limitation by developing a novel zero-shot model that enhances mask precision while maintaining SAM's generalization capabilities.

Image Matting. Image matting is a more complex task than image segmentation, as it focuses on estimating the soft transparency of object boundaries to capture fine details, which is critical in tasks like image compositing and background removal. Unlike segmentation, which assigns hard labels to each pixel, matting requires precise edge detection and soft labeling for smooth blending between objects and their background. Recent developments in zero-shot matting have aimed to build upon the foundational segmentation capabilities of SAM. Most approaches [25, 50, 51] fine-tune SAM on public matting datasets [21, 22, 36, 47].



Public Matting Datasets: Macro-Level **Fine** Labels

SA1B: **Micro-Level** Coarse Labels

SA1B-Matte (Ours): **Micro-Level Fine** Labels

Figure 2. **Qualitative samples from each dataset:** Public matting datasets [21, 22, 36, 47] with macro-level fine labels, the SA1B datset [18] with micro-level coarse labels, and our proposed SA1B-Matte dataset incorporating the micro-level labels with fine details.

However, these datasets predominantly contain macro-level labels, degrading SAM's ability to generalize on micro-level structures, such as individual body parts of a human. The reliance on these datasets can deteriorate the zero-shot generalization of the model. Furthermore, the lack of large-scale matting datasets with micro-level labels restricts progress in developing matting models with truly effective zero-shot ability. In this paper, we correspondingly construct a large-scale micro-level labeled matting dataset via our proposed label converter without laborious annotation procedures, enabling effective zero-shot matting modeling.

3. Methodology

Our contributions can be divided into two components: matting dataset construction (Section 3.1) and network architecture enhancements (Section 3.2).

3.1. Constructing the Zero-Shot Matting Dataset

Motivation. For effective zero-shot matting, a dataset with micro-level matte labels is essential. However, manually annotating matte labels at the micro-level requires extensive human labor and cost. To this end, we present an innovative **Label Converter** that transforms any segment label into a matte label, motivated by mask-guided matting works [35, 53]. We first collect public image matting datasets [20–22, 24, 44, 53] to train the converter. We derive coarse segmentation labels from matte labels by applying image processing techniques such as thresholding, resolution down-scaling, Gaussian blurring, dilation, erosion, and convex hull transformations. The converter takes an image and segmentation label as input source and is trained to produce a corresponding matte label, as illustrated in Figure 3a.

Challenges. (1) Generalization to unseen patterns: Public matting datasets predominantly contain macro-level labels (*e.g.*, entire portraits), as shown in Figure 2. Consequently, the converter trained on these datasets often struggles to generalize to unseen micro-level objects (*e.g.*, individual body parts). This limitation leads to the generation of noisy matte labels when applied to micro-level segmentation (see the 4th column in Figure 6a). **(2)** Unnecessary fine-grained representation: Some objects, such as cars or boxes, commonly do not require fine-grained representation. However, since the converter is trained to always transform segmentation labels into fine-grained matte labels, it often generates unnecessary noise into the output matte, particularly for objects that do not benefit from fine-grained representation (see the 4th column in Figure 6b).

Spatial Generalization Augmentation. To improve the converter's ability to generalize to diverse segmentation labels, we design Spatial Generalization Augmentation. This approach introduces variability into the training data by applying a random cut-out technique, as shown in Figure 3a. During training, both the segmentation label and the corresponding matte label are randomly cropped in the same regions. By exposing the converter to irregular and incomplete input patterns, this augmentation forces the converter to adapt to diverse spatial structures and unseen patterns, thus enhancing its generalization capability. This method ensures that the converter can better handle a variety of input segmentation labels, even those that deviate from training patterns (see the 3rd column in Figure 6a).

Selective Transformation Learning. To prevent the unnecessary transformation of objects that do not require fine-grained details (*e.g.*, cars or desks), we introduce Selective Transformation Learning. This technique enables the converter to selectively focus on objects requiring detailed matte conversion (*e.g.*, hair, trees) while skipping finer transformations for coarse-grained objects. We incorporate these non-transformable samples into the training process by collecting coarse-grained object masks from public segmentation datasets [56] (see Figure 3b). During training, the ground-truth matte label for the non-transformable samples is set to identical to the original segmentation label, allowing the converter to learn that no transformation is required. This selective approach reduces noise in the output

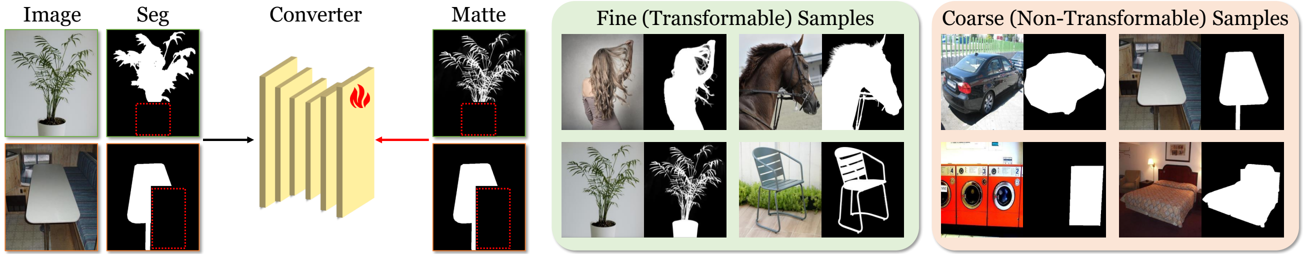


Image
Seg
Converter
Matte
Fine (Transformable) Samples
Coarse (Non-Transformable) Samples

(a) Spatial Generalization Augmentation

(b) Selective Transformation Learning

Figure 3. **Illustration of the key components of the Label Converter**. (a) Overview of the training procedure of the converter using Spatial Generalization Augmentation (indicated by red dotted boxes). (b) Examples of transformable (fine) and non-transformable (coarse) samples used in Selective Transformation Learning for the converter.

and ensures that fine-grained transformations are applied only when needed (see the 3rd column in Figure 6b).

Training. We employ standard loss functions commonly used in matting tasks, namely using a linear combination of L1 and Gradient losses [14, 21, 22] to minimize pixel-wise differences between the ground-truth and predicted matte:

$$L = L_{l1} + \lambda L_{grad} \tag{1}$$

$$L_{l1} = |M - M'| \tag{2}$$

$$L_{grad} = |\nabla_x(M) - \nabla_x(M')| + |\nabla_y(M) - \nabla_y(M')| \tag{3}$$

where M and M' represent the ground-truth and predicted matte label, respectively, and λ is a loss weighting factor. In addition, ∇_x and ∇_y represent the gradients along the horizontal and vertical axes, respectively. Moreover, we set a probability parameter p to control the random application of Spatial Generalization Augmentation during training.

SA1B-Matte Dataset. After training the label converter, we transform segmentation labels in the SA1B dataset [18] to matte labels using the converter, constructing a new SA1B-Matte dataset. As shown in Figure 2, the coarse labels in the SA1B dataset are successfully transformed into high-quality precise matte labels. Compared to existing public matting datasets consisting of macro-level fine labels, the SA1B-Matte dataset is a large-scale image matting dataset with micro-level fine labels, providing an ideal foundation for developing zero-shot matting models.

3.2. ZIM: Zero-Shot Image Matting Model

Overview of ZIM. Our proposed model, ZIM, builds upon SAM [18] and consists of four components, as illustrated in Figure 4: (1) Image Encoder: extracts image features from the input image, producing an image embedding with a stride of 16. (2) Prompt Encoder: encodes point or box inputs into prompt embeddings concatenated with learnable token embeddings, serving a role similar to the `[cls]` token in ViT [8]. (3) Transformer Decoder: takes the image and token embeddings to generate output token embeddings. It performs four operations: self-attention on the tokens, token-to-image cross-attention, an MLP layer, and image-to-token cross-attention that updates the image embedding. (4) Pixel Decoder: upsamples the output image embedding with a stride of 2. Lastly, the model produces matte masks by computing a dot product between the upsampled image embedding and output token embeddings.

Motivation. While SAM has shown success in segmentation tasks, its pixel decoder, which comprises two straightforward transposed convolutional layers, is prone to generating checkerboard artifacts, especially when handling challenging visual prompts, such as multiple positive and negative points placed near object boundaries or box prompts with imprecise object region delineation, as shown in Figure 1. Furthermore, their upsampled embeddings with a stride of 4 are often insufficient for image matting, which benefits from finer mask feature representations.

Hierarchical Pixel Decoder. To address these shortcomings, we introduce a hierarchical pixel decoder with a multi-level feature pyramid design, motivated by [49], as illustrated in Figure 4. The pixel decoder takes an input image and generates multi-resolution feature maps at strides 2, 4, and 8 using a series of simple convolutional layers. The image embedding is sequentially upsampled and concatenated with the corresponding feature maps at each resolution. The decoder is designed to be highly lightweight, namely adding only 10 ms of computational overhead compared to the original pixel decoder of SAM on a V100 GPU. Our hierarchical design serves two key purposes: First, it preserves high-level semantics while refining spatial details, reducing checkerboard artifacts and enhancing robustness to challenging prompts. Second, it generates high-resolution feature maps with a stride of 2, essential for capturing fine-grained structures in matting.

Prompt-Aware Masked Attention. To further boost the interactive matting performance, we propose a Prompt-Aware Masked Attention mechanism, inspired by Mask2Former [6] (See Figure 4). This mechanism allows the model to dynamically focus on the relevant regions within the image based on visual prompts (*e.g.*, points or

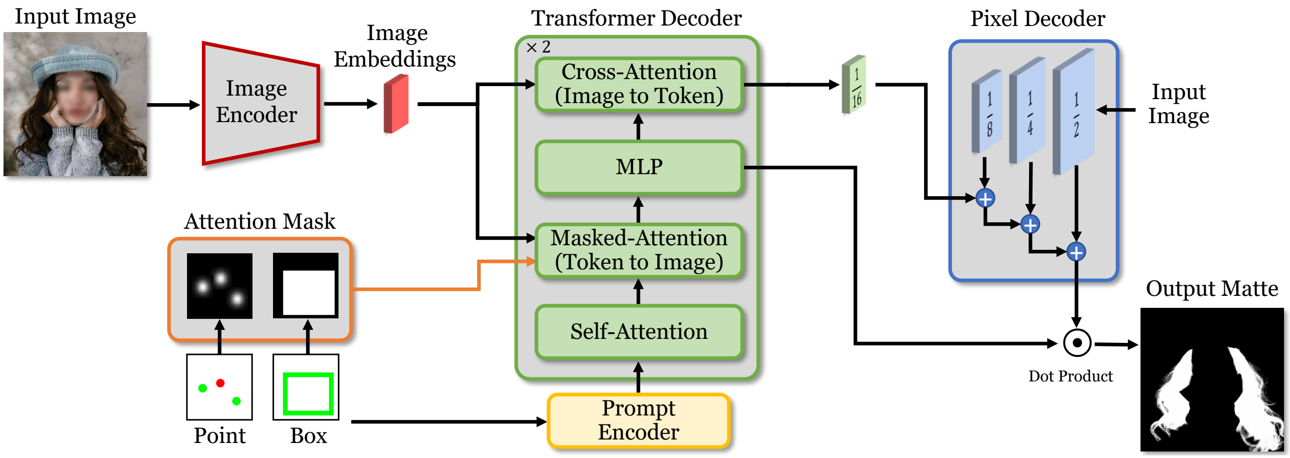


Figure 4. **Overview of the ZIM architecture.** Based on the SAM network architecture [18], we introduce two key improvements: (1) Hierarchical Pixel Decoder for more robust and higher-resolution mask feature map generation, and (2) Prompt-Aware Masked Attention mechanism to enhance interactive matting performance.

boxes), enabling more attention to the areas of interest.

For box prompts, we generate a binary attention mask $\mathcal{M}^b$ that indicates the specific bounding box region. The binary attention mask $\mathcal{M}^b \in \{0, -\infty\}$ is defined as:

$$\mathcal{M}^b(x,y) = \begin{cases} 0 & \text{if } (x,y) \in \text{box region} \\ -\infty & \text{otherwise} \end{cases} \tag{4}$$

where (x, y) represents the pixel coordinates. This forces the model to prioritize the region within the box prompt.

For point prompts, we generate a soft attention mask using a 2D Gaussian map distribution with standard deviation σ. The soft attention mask, $\mathcal{M}^p \in [0, 1]$, smoothly weighs the region around the point of interest, ensuring a graded focus that transitions smoothly to the surrounding regions.

The attention mask is incorporated into the cross-attention blocks of the transformer decoder. Specifically, the attention mask modulates the attention map as follows:

$$X_l = \begin{cases} \text{softmax}(\mathcal{M}^b + Q_l K_l^\intercal) V_l + X_{l-1} & \text{(box prompt)} \\ \text{softmax}(\mathcal{M}^p \odot Q_l K_l^\intercal) V_l + X_{l-1} & \text{(point prompt)} \end{cases} \tag{5}$$

where $\odot$ denotes element-wise multiplication, X_l represents the query feature maps at the l^{th} layer of the decoder, and Q_l, K_l, and V_l implies the query, key, and value matrices, respectively, at the l^{th} layer. This mechanism dynamically adjusts the model's attention according to the visual prompt, leading to performance improvement in prompt-driven interactive scenarios (see Table 3a).

Training. We train ZIM using the SA1B-Matte dataset. From the ground-truth matte label, we extract a box prompt from the given min-max coordinates and randomly sample positive and negative point prompts following [41]. The model is optimized using the same matte loss functions defined in Eq. (1).

4. MicroMat-3K: Zero-Shot Matting Test Set

We introduce a new test set, named MicroMat-3K, to evaluate zero-shot interactive matting models. It consists of 3,000 high-resolution images paired with micro-level matte labels. It includes two types of matte labels: (1) Fine-grained labels (*e.g.*, hair, tree branches) to primarily evaluate zero-shot matting performance, where capturing intricate details is critical. (2) coarse-grained labels (*e.g.*, cars, desks) to allow comparison with zero-shot segmentation models, which is still essential in zero-shot matting tasks. Moreover, It provides pre-defined point prompt sets for positive and negative points and box prompt sets for evaluating interactive scenarios. More detailed information about the MicroMat-3K is described in the supplementary material.

5. Experiments

5.1. Experimental Setting.

Training Dataset for Label Converter. To train the label converter, we collect six publicly available matting datasets (*i.e.*, AIM-500 [21], AM-2K [22], P3M-10K [20], RWP-636 [53], HIM-2K [44], and RefMatte [24]), consisting of 20,591 natural images and 118,749 synthetic images in total. For non-transformable samples, we extract coarse object categories (*e.g.*, car and desk) from the ADE20K dataset [56], sampling 187,063 masks from 17,768 images.

Evaluation Metrics. We use widely adopted evaluation metrics for the image matting task, including Sum of Absolute Difference (SAD), Mean Squared Error (MSE), Gradient Error (Grad), and Connectivity Error (Conn).

Implementation Details for Label Converter. The label converter model is based on MGMatting [53] with Hiera-base-plus [40] backbone network. For training the con-

Method	Prompt	MicroMat3K Fine-grained					MicroMat3K Coarse-grained				
		SAD↓	MSE↓	MAE↓	Grad↓	Conn↓	SAD↓	MSE↓	MAE↓	Grad↓	Conn↓
SAM [18]	point	68.076	21.651	23.307	16.496	67.730	17.093	5.569	5.756	4.800	17.035
	box	36.086	11.057	12.714	14.867	35.834	3.516	1.044	1.231	2.551	3.450
HQ-SAM [13]	point	110.681	36.674	38.331	16.855	110.421	18.842	6.457	6.645	4.599	18.792
	box	124.262	42.457	44.144	13.673	124.113	8.458	2.733	2.920	2.472	8.400
Matte-Any [50]	point	68.797	20.844	23.564	8.118	68.939	19.717	6.053	6.675	2.633	19.506
	box	34.661	9.746	12.182	7.021	34.856	6.950	1.983	2.445	2.142	6.905
Matting-Any [25]	point	275.398	77.335	97.141	20.019	270.722	164.145	36.187	55.943	23.244	155.780
	box	246.214	68.372	87.617	19.185	241.597	109.639	23.780	38.662	15.841	102.439
ZIM (ours)	point	**31.286**	**8.213**	**10.740**	**5.324**	**31.009**	**6.645**	**1.788**	**2.320**	**1.469**	**6.472**
	box	**9.961**	**1.893**	**3.426**	**4.813**	**9.655**	**1.860**	**0.448**	**0.659**	**1.281**	**1.807**

Table 1. **Quantitative comparison** of ZIM and six existing methods on the MicroMat-3K test set, evaluated separately on fine-grained and coarse-grained categories using point and box prompts across five evaluation metrics.

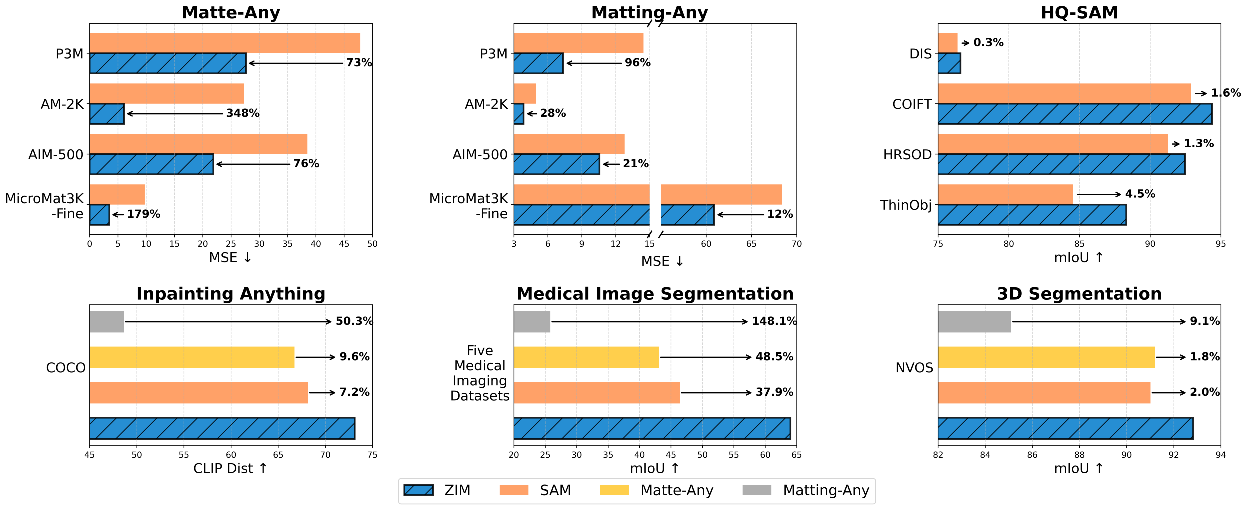


Figure 5. **Downstream Transferability Evaluation** by replacing SAM with ZIM as the segmentation foundation model in various downstream tasks: Matte-Any [50] and Matting-Any [25] are evaluated on MicroMat3K-Fine, AIM-500 [21], AM-2K [22], and P3M-500-NP [20] using the MSE metric. HQ-SAM [13] is assessed on DIS [37], COIFT [27], HRSOD [55], and ThinObject [27] using mIoU. Inpainting Anything [54] is tested on COCO [28] using CLIP distance. Following Medical Image Segmentation evaluation protocol [32] with five different prompting modes (mIoU). 3D segmentation follows the SA3D [3] framework and is evaluated on NVOS [39].

verter, we set the input size to 1024×1024, a batch size of 16, and a learning rate of 0.001 with cosine decay scheduling using the AdamW optimizer [30]. The training process runs for 500K iterations with the probability parameter p of 0.5 and the loss weight λ of 10.

Implementation Details for ZIM. For the ZIM model, we use the same image encoder (*i.e.*, ViT-B [8]) and prompt encoder as SAM. Leveraging the pre-trained weights from SAM, we fine-tune the ZIM model on 1% of the SA1B-Matte dataset, which amounts to approximately 2.2M matte labels. We set the input size to 1024×1024, batch size to 16, a learning rate to 0.00001 with cosine decay scheduling using the AdamW optimizer [30], and training iterations to 500K. The loss weight λ is set to 10 and the σ for the point-based attention mask is set to 21 by default.

5.2. Experimental Results

We evaluate ZIM against four related methods on MicroMat3K: SAM [18], HQ-SAM [13], Matte-Any [50], and Matting-Any [25]. All methods use the ViT-B [8] backbone network. Table 1 presents the evaluation scores across five metrics for both point and box prompts on fine-grained and coarse-grained masks. For coarse-grained results, SAM achieves reasonable zero-shot performance, while other methods (*e.g.*, Matting-Any and HQ-SAM) struggle to generalize to unseen objects. This is likely due to their fine-tuning on macro-level labeled datasets, which degrades their zero-shot capabilities. The qualitative results in Figure 1 show that these methods often produce macro-level outputs, even provided with micro-level prompts. Moreover,

SGA	STL	Fine-grained			Coarse-grained		
		SAD↓	MSE↓	Grad↓	SAD↓	MSE↓	Grad↓
		3.324	0.276	2.664	0.716	0.117	0.684
✓		2.440	0.122	2.139	0.697	0.092	0.634
	✓	3.153	0.239	2.457	0.635	0.089	0.653
✓	✓	**1.999**	**0.080**	**1.771**	**0.281**	**0.021**	**0.399**

Table 2. **Quantitative analysis of key components of the label converter**: Spatial Generalization Augmentation (SGA) and Selective Transformation Learning (STL).

SAM tends to suffer from checkerboard artifacts when challenging prompts are introduced. In contrast, ZIM generates a more robust quality of masks, due to our hierarchical feature pyramid decoder. The fine-grained results in Table 1 highlight ZIM's superiority in producing high-quality matting outputs while maintaining strong zero-shot capabilities.

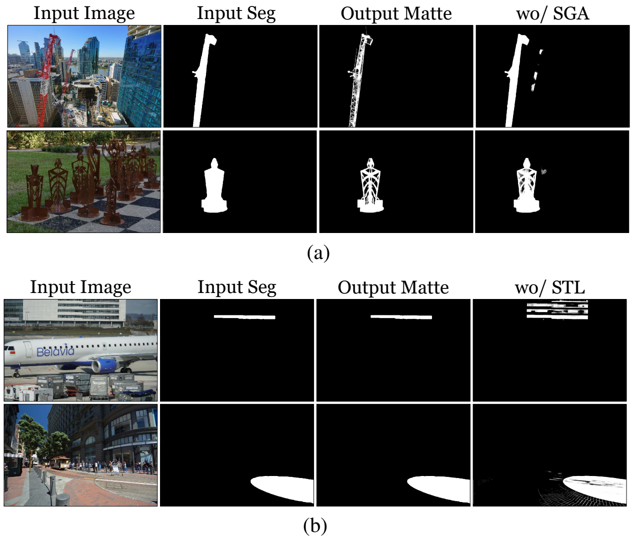
5.3. Downstream Transferability

We assess the transferability of ZIM compared to SAM across various downstream tasks. Specifically, we first integrate ZIM into existing matting methods, Matte-Any [50] and Matting-Any [25], and observe significantly improved box prompting MSE results across diverse matting benchmarks, including MicroMat3K-Fine, AIM-500 [21], AM-2K [22], and P3M-500-NP [20]. Likewise, replacing SAM with ZIM in HQ-SAM [13] notably enhances performance in fine-grained segmentation datasets, including DIS [37], COIFT [27], HRSOD [55], and ThinObject [27].

Moreover, we extend our analysis to broader vision tasks, including image inpainting, medical image segmentation, and 3D segmentation. Specifically, we evaluate image inpainting using the Inpainting Anything framework [54] on the COCO dataset [28] via the CLIP Distance metric [9, 52]. For medical image segmentation, we utilize the zero-shot evaluation protocol from [32] on five diverse medical imaging datasets [1, 11, 33, 42], measured by mean IoU (mIoU). Additionally, we assess 3D segmentation performance within the SA3D framework [3] on the NVOS dataset [39]. These tasks inherently require highly precise segmentation masks for optimal outcomes. While replacing SAM with existing matting models (*e.g.*, Matte-Any and Matting-Any) results in significant performance drops due to limited generalization capabilities, employing ZIM consistently boosts zero-shot performance across these diverse scenarios, highlighting its robust generalization and precise mask representation capabilities. Comprehensive qualitative and quantitative results for these tasks are provided in the supplementary materials.

6. Ablation Study

In this section, we analyze the impact of the key components of our method using the MicroMat-3K test set.



(a)

(b)

Figure 6. **Qualitative analysis of key components of the label converter**: (a) without Spatial Generalization Augmentation (SGA) and (b) without Selective Transformation Learning (STL).

Analysis of Label Converter. To analyze the effect of Spatial Generalization Augmentation (SGA) and Selective Transformation Learning (STL) strategies, we conduct an ablation study by removing each component individually. The SGA is designed to enhance the generalization ability of the converter by simulating diverse input patterns, particularly beneficial given that the converter is trained on macro-level labeled datasets. Without the SGA, the converter struggles to produce clear matte labels for unseen objects, as shown in Figure 6a. In addition, the STL is designed to help the converter avoid unnecessary label conversion for coarse objects. Without the STL, the converter attempts to transform every segmentation label into a matte label, resulting in noisy outputs for unseen coarse objects, as shown in Figure 6b. The quantitative results in Table 2 confirm that using both strategies yields the best label conversion performance on the MicroMat-3K test set.

Analysis of ZIM Model. We conduct experiments to analyze the effect of the prompt-aware masked attention and hierarchical mask decoder. The prompt-aware masked attention is designed to direct the model's focus on the regions of interest to improve the promptable matting performance. Table 3a shows that leveraging the masked attention yields a substantial improvement to our ZIM model. In addition, the hierarchical mask decoder is designed to produce more robust and higher-resolution mask feature maps to alleviate checkerboard artifacts and capture finer representation, simultaneously. Its effectiveness is particularly evident in reducing the gradient error for fine-grained objects in Table 3a, since the enhanced pixel decoder generates more solid and detailed mask outputs. Notably, the decoder remains lightweight, adding only 10 ms of additional inference time.

Attn	Dec	Fine-grained SAD↓	Fine-grained MSE↓	Fine-grained Grad↓	Coarse-grained SAD↓	Coarse-grained MSE↓	Coarse-grained Grad↓
		13.623	2.718	6.516	2.071	0.474	1.526
✓		13.198	2.504	6.445	2.049	0.471	1.486
	✓	11.074	2.094	5.401	2.069	0.487	1.355
✓	✓	**9.961**	**1.893**	**4.813**	**1.860**	**0.448**	**1.281**

(a)

Attn Mask T2I	Attn Mask I2T	Fine-grained SAD↓	Fine-grained MSE↓	Fine-grained Grad↓	Coarse-grained SAD↓	Coarse-grained MSE↓	Coarse-grained Grad↓
		11.074	2.094	5.401	2.069	0.487	1.355
✓		**9.961**	**1.893**	**4.813**	**1.860**	**0.448**	**1.281**
	✓	12.526	2.658	6.032	2.353	0.554	1.481
✓	✓	10.437	1.997	5.066	1.999	0.470	1.306

(b)

Model	Trainset	Fine-grained SAD↓	Fine-grained MSE↓	Fine-grained Grad↓	Coarse-grained SAD↓	Coarse-grained MSE↓	Coarse-grained Grad↓
ZIM	SA1B-Matte	**9.961**	**1.893**	**4.813**	**1.860**	**0.448**	**1.281**
	Public-Matte	120.571	38.332	6.730	9.506	2.760	1.688
Matting-Any [25]	SA1B-Matte	41.242	12.267	7.707	4.626	1.284	1.552
	Public-Matte	246.214	68.372	19.185	109.639	23.780	15.841

(c)

Table 3. **Analysis of ZIM** using box prompt evaluations: (a) Effect of `Attn` (prompt-aware masked attention) and `Dec` (hierarchical pixel decoder). (b) Effect of the masked attention in `T2I` (token to image) and `I2T` (image to token) cross-attention layers. (c) Effect of trainset: our SA1B-Matte and public matting datasets.

Moreover, we delve into the effect of prompt-aware masked attention in our transformer decoder, which comprises two kinds of cross-attention layers (see Figure 4): token-to-image (t2i) updating token embeddings (as queries) and image-to-token (i2t) updating the image embedding (as queries). As a result in Table 3b, applying masked attention to only the t2i layer leads to a meaningful improvement. This suggests that focusing attention on tokens based on visual prompts in the t2i layer enhances their ability to capture relevant features. In contrast, applying attention to specific regions within the image embedding in the i2t layer may disturb the capture of global features.

Analysis of Training Dataset. Table 3c investigates the influence of the training dataset on ZIM's zero-shot matting performance. When ZIM is trained on publicly available matting datasets, which predominantly contain macro-level masks, the performance on the micro-level MicroMat3K dataset significantly degrades. This result underscores the necessity of training with micro-level mask annotations for effective generalization to multi-granularity details. Despite Matting-Any [25] also being trained on SA1B-Matte, ZIM still surpasses its performance by a considerable margin, highlighting the advancements of our network architectural improvements for the zero-shot interactive matting task.

Discussion on Domain Shift. There is a distinct domain shift between traditional matting test sets and ZIM trained with SA1B-Matte. Traditional matting test sets [20, 21] typically regard entire salient objects as foreground, whereas ZIM mainly focuses on object- and part-level matting (Figure 7). This mismatch leads to substantial performance penalties for ZIM under box prompting on these test sets (Figure 7), which is inherent from SAM's prompt ambiguity issue. However, by leveraging dense multiple-point prompting, ZIM effectively mitigates this discrepancy, surpassing existing methods [21, 23, 25, 50, 51], even some of which are explicitly trained on datasets similar to the evaluation domain (Figure 8). This highlights the adaptability of our zero-shot interactive matting modeling.

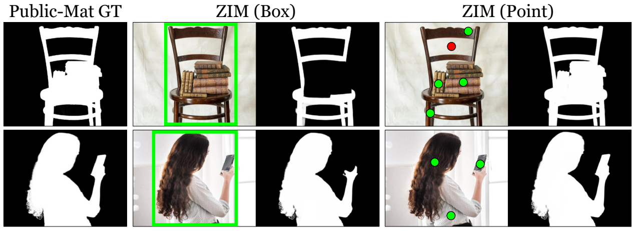


Figure 7. **Domain shift** between traditional public matting testsets and ZIM trained with SA1B-Matte (object-level).

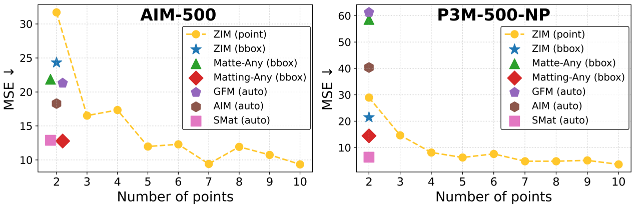


Figure 8. **Quantitative comparison** of ZIM (with varying numbers of point prompts) and existing matting methods (Matte-Any [50], Matting-Any [25], GFM [23], AIM [21], and SMat [51]) on the traditional matting test sets (AIM-500 [21] and P3M-500-NP [20]). The "auto" implies prompt-free mode.

7. Conclusion, Limitation, and Future Work

In this paper, we presented a pioneering zero-shot image matting model that advances the field by generating precise, fine-grained matte masks. We addressed the limitations of SAM, which struggles with high-detail segmentation tasks, by introducing a novel label conversion method and enhancing the network architecture with a hierarchical pixel decoder and prompt-aware masked attention mechanism. However, ZIM inherits inherent shortcomings from SAM, including ambiguous visual prompt handling and robustness in uncertain predictions. Future research should aim at resolving these distinct challenges through innovative approaches to prompt design and uncertainty-based modeling. We hope that the research community will continue to build on this work, exploring new applications in computer vision and enhancing its zero-shot matting performance.

References

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Ildoo Kim, Klaus Maier-Hein, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. The medical segmentation decathlon. *Nature Communications*, 13(1), 2022. 7

[2] Xue Bai and Guillermo Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[3] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, et al. Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems*, 36:25971–25990, 2023. 2, 6, 7

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

[5] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2, 4

[7] Henghui Ding, Hui Zhang, Chang Liu, and Xudong Jiang. Deep interactive image matting with feature propagation. *IEEE Transactions on Image Processing*, 31:2421–2432, 2022.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 6

[9] Yigit Ekin, Ahmet Burak Yildirim, Erdem Eren Caglar, Aykut Erdem, Erkut Erdem, and Aysegul Dundar. Clipaway: Harmonizing focused embeddings for removing objects via diffusion models, 2024. 7

[10] Leo Grady, Thomas Schiwietz, Shmuel Aharon, and Rüdiger Westermann. Random walks for interactive alpha-matting. In *Proceedings of VIIP*, pages 423–429. Citeseer, 2005.

[11] Daniel Gut. X-ray images of the hip joints, 2021. , Mendeley Data, V1, doi: 10.17632/zm6bxzhmfz.1. 7

[12] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023. 2

[13] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 6, 7

[14] 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 the AAAI Conference on Artificial Intelligence*, pages 1140–1147, 2022. 4

[15] Beomyoung Kim, Joonhyun Jeong, Dongyoon Han, and Sung Ju Hwang. The devil is in the points: Weakly semi-supervised instance segmentation via point-guided mask representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11360–11370, 2023.

[16] Beomyoung Kim, Myeong Yeon Yi, Joonsang Yu, Young Joon Yoo, and Sung Ju Hwang. Towards label-efficient human matting: A simple baseline for weakly semi-supervised trimap-free human matting. *arXiv preprint arXiv:2404.00921*, 2024.

[17] Beomyoung Kim, Joonsang Yu, and Sung Ju Hwang. Eclipse: Efficient continual learning in panoptic segmentation with visual prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3346–3356, 2024.

[18] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 2, 3, 4, 5, 6

[19] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023. 2

[20] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3501–3509, 2021. 3, 5, 6, 7, 8

[21] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235*, 2021. 2, 3, 4, 5, 6, 7, 8

[22] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision*, 130(2):246–266, 2022. 2, 3, 4, 5, 6, 7

[23] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision*, 130(2):246–266, 2022. 8

[24] Jizhizi Li, Jing Zhang, and Dacheng Tao. Referring image matting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22448–22457, 2023. 3, 5

[25] Jiachen Li, Jitesh Jain, and Humphrey Shi. Matting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1785, 2024. 1, 2, 6, 7, 8

[26] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11450–11457, 2020.

[27] Jun Hao Liew, Scott Cohen, Brian Price, Long Mai, and Jiashi Feng. Deep interactive thin object selection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 305–314, 2021. 6, 7

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6, 7

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

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 6

[31] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3266–3275, 2019.

[32] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. 6, 7

[33] Anna Montoya, Hasnin, kaggle446, shirzad, Will Cukierski, and yffud. Ultrasound nerve segmentation. `https://kaggle.com/competitions/ultrasound-nerve-segmentation`, 2016. Kaggle. 7

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

[35] Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Mask-guided matting in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1992–2001, 2023. 3

[36] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13676–13685, 2020. 2, 3

[37] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision*, pages 38–56. Springer, 2022. 6, 7

[38] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[39] Zhongzheng Ren, Aseem Agarwala†, Bryan Russell†, Alexander G. Schwing†, and Oliver Wang†. Neural volumetric object selection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. († alphabetic ordering). 6, 7

[40] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, pages 29441–29454. PMLR, 2023. 5

[41] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 5

[42] Yuxin Song, Jing Zheng, Long Lei, Zhipeng Ni, Baoliang Zhao, and Ying Hu. Ct2us: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. *Ultrasonics*, 122:106706, 2022. 7

[43] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Semantic image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11120–11129, 2021.

[44] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Human instance matting via mutual guidance and multi-instance refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2647–2656, 2022. 3, 5

[45] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020.

[46] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15374–15383, 2021.

[47] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2970–2979, 2017. 2, 3

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

[49] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pretrained plain vision transformers. *Information Fusion*, 103: 102091, 2024. 4

[50] Jingfeng Yao, Xinggang Wang, Lang Ye, and Wenyu Liu. Matte anything: Interactive natural image matting with segment anything model. *Image and Vision Computing*, 147: 105067, 2024. 1, 2, 6, 7, 8

[51] Zixuan Ye, Wenze Liu, He Guo, Yujia Liang, Chaoyi Hong, Hao Lu, and Zhiguo Cao. Unifying automatic and interactive matting with pretrained vits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25585–25594, 2024. 2, 8

[52] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models, 2023. 7

[53] 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 the IEEE/CVF conference on computer vision and pattern recognition*, pages 1154–1163, 2021. 3, 5

[54] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 2, 6, 7

[55] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7234–7243, 2019. 6, 7

[56] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 3, 5