EFormer: Enhanced Transformer towards Semantic-Contour Features of Foreground for Portraits Matting

Zitao Wang Qiguang Miao* Yue Xi* Peipei Zhao
Xidian University, China
{21181214174@stu., qgmiao@, xiyue@, zhpp2023@}xidian.edu.cn

Abstract

The portrait matting task aims to extract an alpha matte with complete semantics and finely detailed contours. In comparison to CNN-based approaches, transformers with self-attention module have a better capacity to capture long-range dependencies and low-frequency semantic information of a portrait. However, recent research shows that the self-attention mechanism struggles with modeling high-frequency contour information and capturing fine contour details, which can lead to bias while predicting the portrait's contours. To deal with this issue, we propose EFormer to enhance the model's attention towards both the low-frequency semantic and high-frequency contour features. For the high-frequency contours, our research demonstrates that cross-attention module between different resolutions can guide our model to allocate attention appropriately to these contour regions. Supported by this, we can successfully extract the high-frequency detail information around the portrait's contours, which were previously ignored by self-attention. Based on the cross-attention module, we further build a semantic and contour detector (SCD) to accurately capture both the low-frequency semantic and high-frequency contour features. And we design a contour-edge extraction branch and semantic extraction branch to extract refined high-frequency contour features and complete low-frequency semantic information, respectively. Finally, we fuse the two kinds of features and leverage the segmentation head to generate a predicted portrait matte. Experiments on VideoMatte240K (JPEG SD Format) and Adobe Image Matting (AIM) datasets demonstrate that EFormer outperforms previous portrait matte methods.

1. Introduction

Portrait matting aims to extract a precise alpha matte of persons from natural images. It has gathered great momentum in vision applications, including but not limited to photo editing and background replacement.

Existing methods[20, 30, 33, 35] attempt to estimate a portrait matte with Convolutional Neural Networks (CNNs) under the guidance of a pre-determined trimap, which is a three-class map that indicates the foreground, background, and unknown region. However, it is time-consuming and labor-intensive to obtain abundant trimaps by manually an-

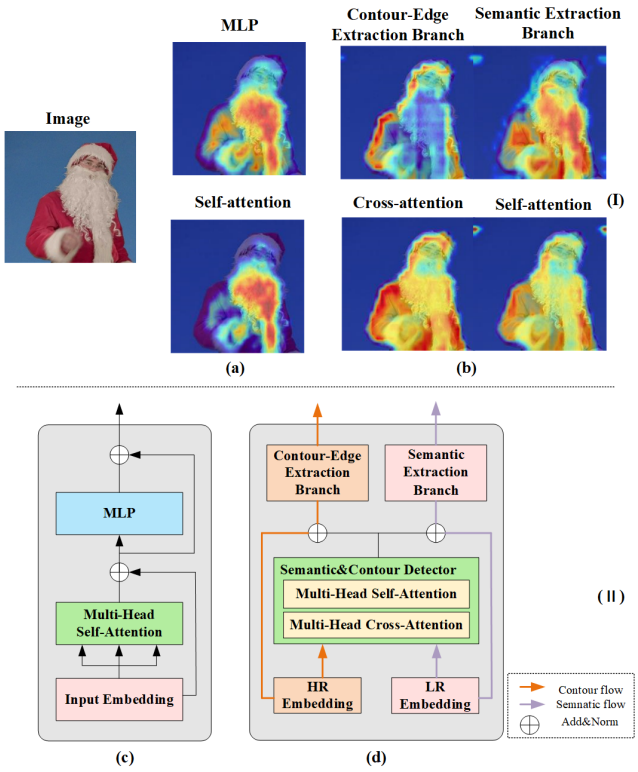

Figure 1. (I) Visualization with Grad-CAM of output from each layer within different transformer blocks. (a) The features captured from Self-attention and MLP layers in the original transformer block. (b) In comparison, the features captured from each layer in the proposed transformer block. (II) Comparisons of different transformer blocks. (c) The original transformer block in ViT. (d) The proposed transformer block for portrait matte.

*Corresponding authors.

notating images. Some methods eliminate the trimap, such as background matting series[21, 28] uses a separate background image instead of the trimap as an auxiliary reference for model prediction. However, in practical cases, it is necessary to provide two aligned images for the model, which include one with only the background and the other one including people. An alternative solution[17, 22, 23] is to rely solely on RGB images as input for predicting portrait mattes. Although the previous models have made strides in portrait matte, there is a constraint on the receptive field in CNN-based approaches.

In contrast, Vision Transformer (ViT)[7] and Swin Transformer (Swin) [24] for semantic segmentation employ a self-attention mechanism, which conducts interactions among all pixels in an image to obtain a global receptive field. Previous research [25] shows that ViT and its variants can capture the low-frequency components in images effectively, such as global shapes and structures of a scene or object. However, they usually neglect high-frequency ones in images, such as edges and textures. They work similarly to low-pass filters.

As shown in Fig. 1(a), we can observe the transformer block with self-attention clearly pays little attention to the portrait's high-frequency contour regions. We believe that this phenomenon is attributed to the excessive propagation of global information because global attention is performed in each transformer block during the calculation of multiple blocks. So, the global information is propagated throughout the model in this iterative calculation, ultimately leading it to only focus on low-frequency semantic features with stronger clustering in portraits. Over time, the model gradually ignores the discrete high-frequency detail features around the contours. Therefore, it is desired to capture low- and high-frequency components simultaneously to improve the segmentation performance.

To achieve this goal, Inception Transformer (IFormer)[29] splits all channels into a convolutional path and self-attention path, respectively, to capture high and low frequencies. However, it is required for a channel ratio of each block at different levels in the IFormer. These channel ratios typically require manual adjustment, resulting in significant uncertainty when aiming for optimal model performance. To enhance the model's stability and adaptive capability, we propose a novel architecture EFormer without manually adjusted parameters. In EFormer, we use the cross-attention layer between different resolution features, as shown in Fig. 1(d), which is replaced by convolution to handle high-frequency detail features.

Our research further demonstrates that the cross-attention module between different resolutions can formulate a more reasonable attention allocation mechanism, which can accurately capture contour features, as shown in Fig. 1(b). Particularly, EFormer with a cross-attention layer not only can not lose the semantic information within the portrait but also can successfully capture the high-frequency details around the portrait's contours, which were previously ignored by self-attention. It is worth noting that high-frequency components only around the boundary between foreground and background are critical to the improvement of segmentation performance. Therefore, we follow a coarse-to-fine manner. Initially, EFormer needs to locate all high-frequency regions in the image. Then, it filters and extracts the high-frequency regions around the foreground contours.

Based on the cross-attention module, we design a semantic and contour detector (SCD) in EFormer, which cascades the cross-attention layer and self-attention layer to sequentially locate the contour and semantic features. Firstly, the model adjusts its attention to contour features with the guidance of a cross-attention module. Secondly, the self-attention module captures semantic information within a portrait's contours and reverses filtering and correcting the contour information output by the cross-attention layer. This allows the model to gradually capture and match the high-frequency contour information and the low-frequency semantic information. Then, we use multi-layer perceptron (MLP) to build the contour-edge extraction branch (CEEB) and the semantic extraction branch (SEB), respectively. With this support, we independently purify the contour flow and the semantic flow to obtain finer contour information and more comprehensive portrait semantic information, as shown in Fig. 1(b). Finally, we fuse the contour features and semantic features and send them to the segmentation head to estimate the portrait matte. We perform extensive experiments on VideoMatte240K (JPEG SD Format)[21], Adobe Image Matting (AIM)[35], and BG10K. Comparisons with other state-of-the-art models and ablation studies verify that our model performs better than previous works.

To summarize, our contributions are as follows:

- We propose EFormer: an approach that can enhance the transformer's attention towards both semantic and contour features of the foreground.
- We further build a semantic and contour detector (SCD) to accurately capture the semantic and contour features and design two separate extraction branches to purify the contour flow and the semantic flow.
- We leverage the cross-attention module between different resolution features in transformer block to autonomously capture and extract high-frequency detail features, which are sparsely distributed along the foreground contour.

2. Related Works

2.1. Image Matting

Image matting is an essential task in the field of computer vision that aims to accurately estimate the foreground in an

image. Mathematically speaking, an image I is a combination of an unknown foreground image F and a background image B with a probability coefficient alpha matte maps α.

$$I = \alpha F + (1 - \alpha) B \tag{1}$$

Previous image matting solutions have predominantly focused on low-level features, such as color cues[9, 10, 14, 15] or propagation[1, 2, 19], to distinguish the transition areas between foregrounds and backgrounds. However, such traditional matting algorithms commonly struggle to perform properly in complex scenes.

With the significant advancements making in deep learning, numerous methods based on convolutional neural networks (CNNs) are proposed, leading to notable successes. Some approaches[36, 37, 39] incorporate auxiliary trimap supervisions to enhance matting performance, while other methods[22, 23] leverage trimap-free solutions to estimate alpha mattes from image feature maps using an end-to-end segmentation network. In the background matting series[21, 28], an auxiliary input of the background image is used to compute alpha matte. Furthermore, MODNet[17] proposes an end-to-end manner without any auxiliary input. Although prior works based on CNN kernels have improved the accuracy of image matting, the prediction of models is in the limited receptive field.

2.2. Vision Transformer

Different from CNNs, vision transformers with self-attention mechanisms can capture long-term dependencies. Therefore, in the realm of computer vision, vision transformers have gained significant attention in recent times. For pixel-level prediction tasks, such as segmentation, SETR[38], SegFormer[34] and DPT[27] apply transformer as an encoder to attain feature maps of images. Their performances demonstrate the transformer with self-attention is capable of building complete contextual information through global interactions. However, recent studies[25, 29] indicate that transformer "yet is incompetent in capturing high frequencies that predominantly convey local information".

Compared to self-attention, cross-attention has more potential to induce customized features. Max-Deeplab[31], MaskFormer[5, 6] and SeMask[13] use query-based methods through cross-attention, which are inspired by DETR[3]. They view segmentation as a set prediction problem. U-Transformer[26] and EUT[8] modify the cross-attention in the original transformer to leverage the information from the encoder, allowing a fine spatial recovery in the decoder. The cross-resolution attention employed by RTFormer[32] enables the gathering of comprehensive contextual information for high-resolution features.

In this work, we propose a strategy that using of cross-attention between different resolutions guides the model to autonomously locate and capture high-frequency features, particularly those near the portrait's contours. Contrary to IFormer[29], this method eliminates the need for predefined parameters to enhance the model's adaptive capabilities. At the same time, this approach can significantly improve the model's performance in capturing contour details.

2.3. Refinement for Segmentation

In the field of portrait segmentation, refining the high-frequency details near the contours is crucial. Many prior studies on segmentation refinement depend on convolutional networks or MLPs that are designed specifically for this purpose.

PointRend[18], based on the coarse predictions of Mask R-CNN[12], calculates confidence scores as the point-wise uncertainty measure. And it focuses on the points with the higher uncertainty according to set hyperparameters. Finally, PointRend uses a shared MLP to refine the labels of the selected points. Although PointRend has implemented segmentation refinement to some extent, the model still requires manual adjustment of certain hyperparameters to assist in filtering uncertain points. Transfiner[16] processes detected error-prone and incoherent regions with transformer, which include high-frequency regions strewn along object boundaries. However, during this procedure, excessive focus is placed on the local details and incoherent regions, without effectively associating the local features of the object with the overall semantic information.

We believe that locating effective detail distribution areas requires information interaction between local details and global semantics. To achieve this, we design a semantic and contour detector that adopts a cascade structure of cross-attention and self-attention. This approach uses global semantic information garnered from self-attention to autonomously select out local details situated at the borders of the contours. On this base, we further utilize the MLP-based extraction branch to purify contour details, thereby significantly enhancing the effect of refinement.

3. Model

In this section, we introduce the process of establishing the complete model architecture of EFormer, which comprises the backbone for encoder, the transformer block for decoder, and the prediction stage, as shown in Fig. 2.

3.1. Backbone for Encoder

The CNN backbone takes RGB images $I \in \mathbb{R}^{B\times 3\times H\times W}$ as input, where B is the number of images, to obtain a pyramid of feature maps $F = \left\{F_{enc\frac{1}{4}}, F_{enc\frac{1}{8}}, F_{enc\frac{1}{16}}\right\}$ with dimension $F_{enc\frac{1}{i}} \in \mathbb{R}^{B\times C_i\times \frac{H}{i}\times \frac{W}{i}}$. As we all know, in the pyramid of feature maps, shallow features have high resolution and rich details, while deep features have low reso-

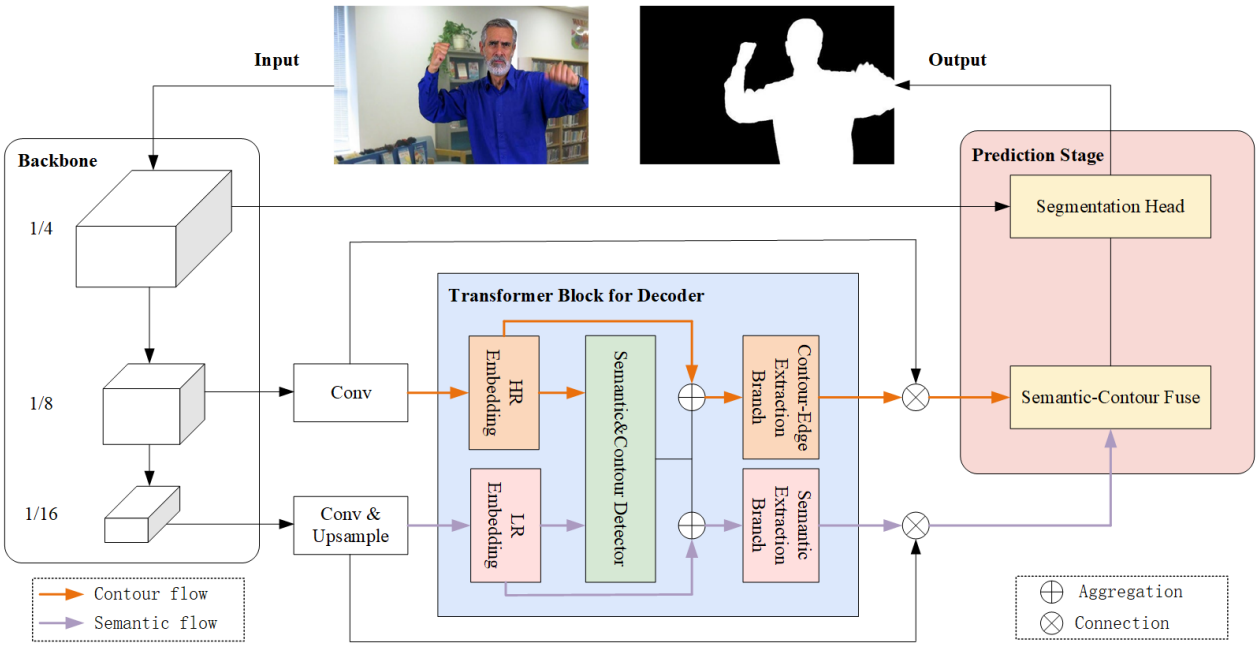

Figure 2. The Architecture of the EFormer. It contains a backbone for encoder, a transformer block for decoder, and the prediction stage, where the transformer block includes a semantic and contour detector, a semantic extraction branch, and a contour-edge extraction branch.

lution but more concentrated semantic information. Based on this, our research makes the cross-attention mechanism to filter the high-frequency contour features from high-resolution features, using low-frequency semantic information in low-resolution features. However, since the large resolution of feature maps would bring large computation cost, we use the $F_{enc\frac{1}{8}}$ as F_{HR} (high-resolution features) and use the $F_{enc\frac{1}{16}}$ as F_{LR} (low-resolution features) in our base model. The impact of other choices of F_{HR} and F_{LR} on model performance is demonstrated in ablation experiments Sec. 4.4. As shown in Fig. 2, the F_{HR} serves as the source of contour flow, providing high-frequency contour detail features for the following processes. Correspondingly, the F_{LR} serves as the source of semantic flow, providing low-frequency semantic information. Then, the F_{HR} and F_{LR} are projected to same dimension C and respectively flattened into high-resolution feature embedding as F_{HR}^{em} and low-resolution feature embedding as F_{LR}^{em}, where $F_{HR}^{em} \in R^{N\times B\times C}$, $F_{LR}^{em} \in R^{N\times B\times C}$, and $N = \frac{H}{8} \times \frac{W}{8}$. Both of them are utilized as inputs for the transformer block.

3.2. Transformer Block for Decoder

We propose a novel transformer block in the model decoding stage to enhance the ability to capture and extract visual features. It not only does not reduce the focus on low-frequency semantic features but also significantly increases the attention to high-frequency contour features. Its detailed architecture is depicted in Fig. 3. The novel transformer block first uses a Semantic and Contour Detector (SCD) to locate and capture high-frequency contour features and low-frequency semantic features. To achieve this, we cascade a cross-attention layer (CA) and a self-attention layer (SA) in the SCD. Then, as shown in Fig. 3, a Contour-Edge Extraction Branch (CEEB) and a Semantic Extraction Branch (SEB) are further built to handle high-frequency contour feature flow and low-frequency semantic feature flow respectively. These branches can separately purify more refined high-frequency contour features and extract more complete low-frequency semantic features. This ensures an outcome that is optimal for both types of features.

Semantic and Contour Detector (SCD). There is a fact that low-frequency semantic features far outperform high-frequency contour detail features in terms of quantity and distribution density. So, in the global information propagation process, the low-frequency semantic information easily dominates the feature representations of the original transformer block with self-attention module.

To prevent the loss of high-frequency contour detail information, we first use a cross-attention layer with different resolution features rather than a self-attention layer. The cross-attention layer can capture high-frequency detail features with high semantic correlation in high-resolution features, based on the low-frequency semantic information in low-resolution features. Technically, after getting F_{HR}^{em} and F_{LR}^{em}, we use Layer Norm (LN) processing them to maintain consistency. Due to the sensitivity of the attention mechanism towards positional information, we use absolute and

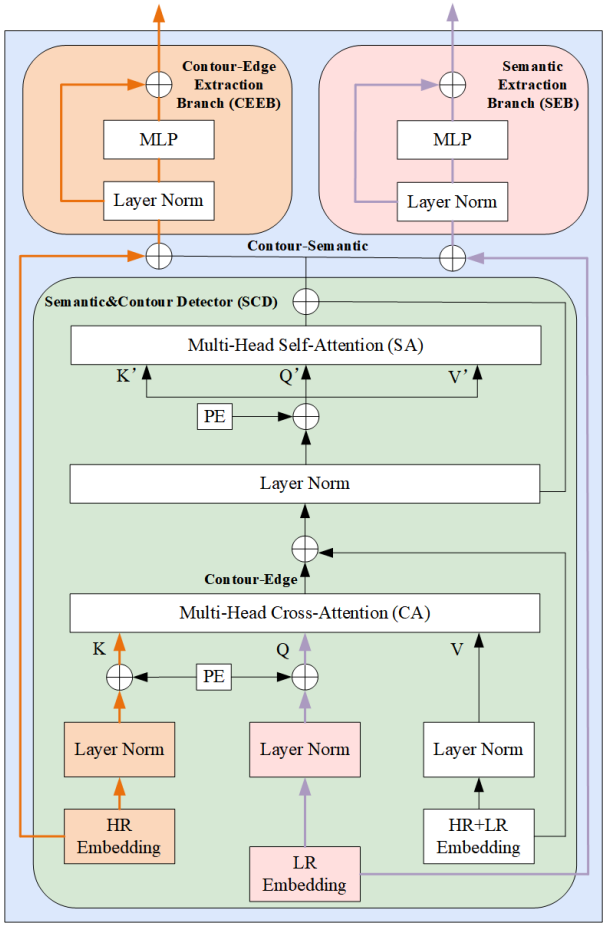

Figure 3. Our transformer block. It includes a Semantic and Contour Detector (SCD), a Semantic Extraction Branch (SEB), and a Contour-Edge Extraction Branch (CEEB). The detailed composition of the module is shown in the figure.

learnable position encoding (PE) in the SCD. The input K, Q, V of the cross-attention layer (CA) is calculated as

$$K = LN\left(F_{HR}^{em}\right) + PE \tag{2}$$

$$Q = LN\left(F_{LR}^{em}\right) + PE \tag{3}$$

$$V = LN\left(F_{LR}^{em} + F_{HR}^{em}\right) \tag{4}$$

By performing cross-attention calculations using features of different resolutions, it is possible to automatically locate and capture the discrete contours along the portrait and allocate more attention to corresponding positions. The contour-edge features can be derived from the output of the CA.

$$F_{contour-edge} = CA\left(K,\ Q,\ V\right) \tag{5}$$

Although CA pays extra attention to contour-edge features, it still captures the overall semantics. By the residual connection between CA and V, both the semantic and contour-edge features of the portrait are enhanced, especially the contour-edge features.

$$F_{enhance} = CA\left(K,\ Q,\ V\right) + V \tag{6}$$

Since semantic information is still present, we use the self-attention layer (SA) after CA. SA can locate the semantic feature within the portrait and contours, thereby emphasizing the semantic attributes at their respective positions. After using LN to normalize $F_{enhance}$, the input K', Q', V' of the SA is calculated as

$$K' = Q' = LN\left(F_{enhance}\right) + PE \tag{7}$$

$$V' = LN\left(F_{enhance}\right) \tag{8}$$

$$F_{contour-semantic} = SA\left(K',\ Q',\ V'\right) + V' \tag{9}$$

Through SA filtering the semantic and contour features, the residual connection between V' and the output of SA, we can produce $F_{contour-semantic}$ as the output of the semantic and contour detector.

Feature Extraction Branches. After SCD, we use multi-layer perceptron (MLP) to build a Contour-Edge Extraction Branch (CEEB) and a Semantic Extraction Branch (SEB). To prevent the loss of high-frequency contour detail information, we aggregate $F_{contour-semantic}$ getting from SCD and F_{HR}^{em} in the contour flow. After getting the merged feature normalized by LN, we employ CEEB to purify it, in order to further extract and refine high-frequency contour details individually. Similarly, to obtain more comprehensive and consistent portrait semantic features separately, we merge $F_{contour-semantic}$ and F_{LR}^{em} in the semantic flow. Then, we normalize the aggregated feature by applying LN and use SEB to filter it.

$$F_{contour} = MLP\left(LN\left(F_{HR}^{em} + F_{contour-semantic}\right)\right) \tag{10}$$

$$F_{semantic} = MLP\left(LN\left(F_{LR}^{em} + F_{contour-semantic}\right)\right) \tag{11}$$

3.3. Prediction Stage

After getting key characteristics $F_{semantic} \in R^{N\times B\times C}$ and $F_{contour} \in R^{N\times B\times C}$ from transformer block, where $N = \frac{H}{8}\times\frac{W}{8}$, we transform them into $F'_{semantic} \in R^{B\times C\times\frac{H}{8}\times\frac{W}{8}}$ and $F'_{contour} \in R^{B\times C\times\frac{H}{8}\times\frac{W}{8}}$ by converting the vector into the matrix. To achieve improved outcomes in portrait segmentation, we propose fusing the two kinds of features and ultimately acquiring the predicted portrait matte from the segmentation head module.

Fuse. We attain the semantic feature through the residual connection between F'_{LR} and $F'_{semantic}$ of semantic flow. Similarly, we get the contour feature from contour flow and subsequently merge the $F'_{semantic}$ and $F'_{contour}$ to generate the fused features $F'_{contour-semantic}$.

$$F'_{semantic} = Conv\left(F'_{semantic} + F'_{LR}\right) \tag{12}$$

$$F'_{contour} = Conv\left(F'_{contour} + F'_{HR}\right) \quad (13)$$

$$F'_{contour-semantic} = Fuse\left(F'_{semantic} + F'_{contour}\right) \quad (14)$$

Segmentation Head. Finally, We gather $F'_{contour-semantic}$ and $F_{enc\frac{1}{4}}$ with higher resolution to predict portrait matte in 1/4 scale of the input image and upscale it to the original scale through bi-linear interpolation as the output of our model.

$$matte = Head\left(F'_{contour-semantic} + F_{enc\frac{1}{4}}\right) \quad (15)$$

4. Experiments

4.1. Dataset and Evaluation

Dataset. We utilize composite training data from two sources: the foreground image dataset VideoMatte240K (JPEG SD Format)[21] and Adobe Image Matting (AIM)[35]. For the background images, we select BG10K following the approach of BGMv2[21], which is a collection of photographs depicting various life scenes without any human portraits. We select foregrounds from VideoMatte240K-JPEG-SD and AIM, and backgrounds from BG10K to composite image datasets. We split VideoMatte240K-JPEG-SD into 234,982/3,007/2720 image sets with image resolutions of 224×224 and 512×288 for training, validating, and testing our model. Similarly, BG10K is split into 9000/1000 image sets, and AIM is split into 214/10 image sets for training and testing, where AIM only includes the portrait images with resolutions at 512×512. Finally, we use the aforementioned testing datasets to compare and evaluate the performance of EFormer against other models.

Evaluation. We mainly consider portrait matting accuracy for evaluation. For portrait matting accuracy, we use Mean Absolute Difference (MAD), Mean Squared Error (MSE), Gradient (Grad), and Connectivity (Conn) as evaluation metrics. We also scale MAD, MSE, Grad, and Conn by 10^3, 10^3, 10^{-3}, and 10^{-3} respectively, for convenience of reference. For all these metrics, the lower number represents better performance.

4.2. Implementation Details

Training Settings. For training the network, we use a single RTX 3090 GPU with batch size at 24. The optimizer is AdamW and the initial learning rate is set to 10^{-4}. All models in ablation studies are trained for 25 epochs, with the learning rate decaying by a factor of 0.8 every 5 epochs. Additionally, to augment the data, each input image is subjected to random horizontal flipping.

Backbone. The backbone we use for the network is the pretrained ResNet50[11], since it is the most frequently used backbone in prior works. We use the implementation and weights from torchvision.

Model	MAD↓	MSE↓	Grad↓	Conn↓
DeepLabv3[4]	14.4700	9.6700	8.5500	1.6900
MODnet[17]	10.3900	5.6500	2.0200	1.0400
BGMv2[21]	4.1858	1.7934	1.3011	0.5326
RVM[22]	5.9900	1.1700	1.1000	0.3400
EFormer	**2.3097**	**0.6637**	**0.4580**	**0.2740**

Table 1. Comparisons on the test set of VideoMatte240K-JPEG-SD. **Bold** indicates the best performance among these models under the inputs with the same resolution at 512×288.

Model	MAD↓	MSE↓	Grad↓	Conn↓
DeepLabv3[4]	29.64	23.78	20.17	7.71
MODnet[17]	21.66	14.27	5.37	5.23
BGMv2[21]	44.61	39.08	5.54	11.60
RVM[22]	14.84	8.93	4.35	3.83
EFormer	**7.47**	**2.13**	**2.83**	**1.90**

Table 2. Comparisons on the test set of AIM. **Bold** indicates the best performance among these models under the inputs with the same resolution at 512×512.

Transformer Block for Decoder. The transformer block is based on the transformer decoder of DETR. We continuously use four transformer blocks in the decoder of the network. We set the channel $C = 256$ and the number of attention heads $M = 8$ for each multi-head attention module.

Prediction Stage. After obtaining the semantic and contour feature maps of the portrait from the stacked transformer blocks, we fuse both of them. Then, the alpha mattes predicted by the segmentation head are upsampled to the original size of the input image with bi-linear interpolation.

Loss function. To compute losses during the training stage, we use a Binary CrossEntropy Loss function.

$$L_{bce} = -\left(g \times \log\left(m\right) + \left(1 - g\right) \times \log\left(1 - m\right)\right) \quad (16)$$

Where, m is the prediction of the network, g is the ground truth.

4.3. Comparisons to State-of-the-art Methods

To demonstrate the progressiveness of EFormer, we compare it against the latest state-of-the-art trimap-free portrait matting solutions. These are including DeepLabV3[4] with ResNet101[11] backbone, BGMv2[21] with ResNet50[11] backbone, MODNet[17], and RVM[22]. Our assessment is based on the testing datasets of VideoMatte240K-JPEG-SD and AIM, respectively. As the training files for MODNet are not available, we use its official weights. Furthermore, since BGMv2, DeepLabV3, and RVM have already been trained on all datasets, our method can be compared with them in a fair manner.

Figure 4. Visualization of portrait matte predictions from MODnet, BGMv2, and EFormer under challenging image from the test set of VideoMatte240K (JPEG SD Format). Our model shows a better ability to distinguish ambiguous foreground contours, as indicated by the red box. Please zoom in to view the detailed information.

Model	MAD↓	MSE↓	Grad↓	Conn↓
DeepLabv3[4]	18.1586	14.7623	6.0015	2.8824
MODnet[17]	13.1400	8.9656	2.6028	2.1543
BGMv2[21]	7.0471	2.3543	1.5621	1.0651
RVM[22]	6.1357	2.2647	0.6523	0.2866
EFormer	**4.1357**	**1.4280**	**0.3760**	**0.1753**

Table 3. Comparisons on the test set of VideoMatte240K-JPEG-SD. **Bold** indicates the best performance among these models under the inputs with the same resolution at 224×224.

	MAD↓	MSE↓	Grad↓	Conn↓
5-epoch	4.3953	1.4824	0.4089	0.1865
10-epoch	4.3108	1.4766	0.3908	0.1834
15-epoch	4.2867	1.5238	0.4087	0.1833
20-epoch	**4.1357**	**1.4280**	**0.3760**	**0.1753**
25-epoch	4.1811	1.4797	0.3858	0.1779

Table 4. Ablation Study on the Training Epochs with the test set of VideoMatte240K-JPEG-SD (resolution at 224×224). **Bold** indicates the best performance among these models. We use the result of 20-epoch training as a strong baseline model.

As shown in the Tab. 1, Tab. 2 and Tab. 3 , our model exhibits superior performance in portrait matte tasks. It has a stronger capability to distinguish the contour of a portrait. And it can finely segment edge details (*e.g.* fingers), as exemplified in Fig. 4.

4.4. Ablation Studies

Training Epochs. In the ablation study with an image resolution of 224×224, we conduct 25 training epochs. Every 5 epochs of training, we evaluate the training effect of the model. At the same time, the learning rate is reduced by 0.8 times. The experimental results indicate that our model achieves optimal performance in the 20 epoch, as shown in the Tab. 4.

The Functional Role of Each Module. To demonstrate the effectiveness of the model in capturing and extracting contour details and portrait semantics, we analyze the outputs of CA, SA, CEEB, and SEB, and utilize Class Activation Map (CAM) to visualize the attention allocation of each layer's features, as shown in Fig. 5.

From the first line of Fig. 5, it can be evident that after the cross-attention layer, the model clearly strengthens its attention to the contours of the portrait. Following the contour prior, the self-attention layer captures the portrait's semantics thoroughly. CEEB and SEB models provide sharper contour and more comprehensive semantic information.

Comparing the second and first lines of Fig. 5, it is ob-

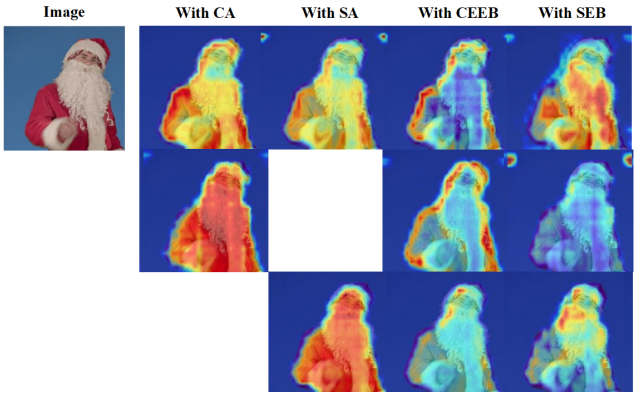

Figure 5. We use Grad-CAM to visualize the attention distribution of various features in CA, SA, CEEB, and SEB. The model in the first row contains four complete modules. For comparison, the SA layer and CA layer were removed from the models in the second and third rows, respectively.

CA	SA	MAD↓	MSE↓	Grad↓	Conn↓
✓		4.2762	1.5344	0.4092	0.1823
	✓	4.1710	1.4354	0.3773	0.1769
✓	✓	**4.1357**	**1.4280**	**0.3760**	**0.1753**

Table 5. Ablation Study on the Functional Role of CA and SA with the test set of VideoMatte240K-JPEG-SD (resolution at 224 × 224). **Bold** indicates the best performance among these models.

served that removing the self-attention layer considerably weakens the semantic information extracted by SEB. In the comparison between the third and first lines of Fig. 5, it can be indicated that removing the cross-attention layer significantly weakens the contour information extracted by CEEB.

After qualitative analysis, the prediction results of all the above models are evaluated. It is concluded that the model with CA, SA, CEEB, and SEB is both effective and rational, as demonstrated in the Tab. 5.

As is well known, when it comes to extracting features for portrait segmentation tasks, semantic information plays a significant role, in contrast, contour details account for a relatively small proportion. Owing to this, as shown in the Tab. 5 and Fig. 5, the model without SA leads to a significant increase in all metrics due to the lack of amount of semantic information, while the model without CA experiences a slight increase in all indicators due to the loss of contour details. In order to fully harness the potential of the vision transformer, we integrate both CA and SA to ensure semantic completeness and enhance attention towards contour details, thereby achieving more comprehensive and fine-grained segmentation results.

The Selection of F_{HR} and F_{LR}. In our base model, considering the large resolution of feature maps would bring large computation cost, we use the $F_{enc\frac{1}{8}}$ as F_{HR} (high-resolution features) and use the $F_{enc\frac{1}{16}}$ as F_{LR} (low-resolution features). Then the F_{HR} serves as the source of contour flow, providing high-frequency contour detail features. Correspondingly, the F_{LR} serves as the source of semantic flow, providing low-frequency semantic information. As shown in the Tab. 6, after using $F_{enc\frac{1}{4}}$ with higher resolution as F_{HR}, our model can capture more high-frequency contour detail features and its performance is better. And after using $F_{enc\frac{1}{16}}$ with lower resolution as F_{LR}, our model can use more concentrated low-frequency semantic information, which makes all indicators optimal.

F_{HR}	F_{LR}	MAD↓	MSE↓	Grad↓	Conn↓
$F_{enc\frac{1}{8}}$	$F_{enc\frac{1}{16}}$	4.1357	1.4280	0.3760	0.1753
$F_{enc\frac{1}{4}}$	$F_{enc\frac{1}{8}}$	3.9970	1.3343	0.3328	0.1679
$F_{enc\frac{1}{4}}$	$F_{enc\frac{1}{16}}$	**3.9730**	**1.2949**	**0.3215**	**0.1665**

Table 6. Ablation Study on the Selection of F_{HR} and F_{LR} with the test set of VideoMatte240K-JPEG-SD (resolution at 224 × 224). **Bold** indicates the best performance among these models.

5. Conclusion

In this paper, we propose a strategy that use of cross-attention module between different resolutions guides the EFormer to autonomously locate and capture high-frequency contour features. Without manual parameter tuning, the proposed method can adaptability capture high-frequency details neglected by previous transformer-based methods. Based on the cross-attention module, we further construct a semantic and contour detector (SCD), which can enhance the transformer's attention to both the high-frequency contour and low-frequency semantic features. Additionally, we build a contour-edge extraction branch (CEEB) and a semantic extraction branch (SEB) to extract finer high-frequency contour details and more comprehensive portrait's low-frequency semantic information. Extensive experiments show that our method outperforms previous portrait matting solutions and exhibits robust performance during testing and inference.

Acknowledgement

The work was jointly supported by the National Natural Science Foundations of China under grant No.62272364, the Teaching Reform Project of Shaanxi Higher Continuing Education under Grant No.21XJZ004.

References

[1] Yagiz Aksoy, Tunc Ozan Aydin, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 29–37, 2017. 3

[2] Yağiz Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. Semantic soft segmentation. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018. 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 6, 7

[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. 3

[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. 3

[7] 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. 2

[8] Dazhao Du, Bing Su, Yu Li, Zhongang Qi, Lingyu Si, and Ying Shan. Efficient u-transformer with boundary-aware loss for action segmentation. *arXiv preprint arXiv:2205.13425*, 2022. 3

[9] Xiaoxue Feng, Xiaohui Liang, and Zili Zhang. A cluster sampling method for image matting via sparse coding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 204–219. Springer, 2016. 3

[10] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In *CVPR 2011*, pages 2049–2056. Ieee, 2011. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[13] Jitesh Jain, Anukriti Singh, Nikita Orlov, Zilong Huang, Jiachen Li, Steven Walton, and Humphrey Shi. Semask: Semantically masked transformers for semantic segmentation. *arXiv preprint arXiv:2112.12782*, 2021. 3

[14] Jubin Johnson, Ehsan Shahrian Varnousfaderani, Hisham Cholakkal, and Deepu Rajan. Sparse coding for alpha matting. *IEEE Transactions on Image Processing*, 25(7):3032–3043, 2016. 3

[15] Levent Karacan, Aykut Erdem, and Erkut Erdem. Image matting with kl-divergence based sparse sampling. In *Proceedings of the IEEE international conference on computer vision*, pages 424–432, 2015. 3

[16] Lei Ke, Martin Danelljan, Xia Li, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Mask transfiner for high-quality instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4412–4421, 2022. 3

[17] 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. 2, 3, 6, 7

[18] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020. 3

[19] Anat Levin, Alex Rav-Acha, and Dani Lischinski. Spectral matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(10):1699–1712, 2008. 3

[20] 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. 1

[21] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 2, 3, 6, 7

[22] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 238–247, 2022. 2, 3, 6, 7

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2, 3

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[25] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 2, 3

[26] Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-net transformer: Self and cross attention for medical image segmentation. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 267–276. Springer, 2021. 3

[27] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of*

the IEEE/CVF international conference on computer vision, pages 12179–12188, 2021. 3

[28] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2291–2300, 2020. 2, 3

[29] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng Yan. Inception transformer. *Advances in Neural Information Processing Systems*, 35:23495–23509, 2022. 2, 3

[30] Yanan Sun, Guanzhi Wang, Qiao Gu, Chi-Keung Tang, and Yu-Wing Tai. Deep video matting via spatio-temporal alignment and aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6975–6984, 2021. 1

[31] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5463–5474, 2021. 3

[32] Jian Wang, Chenhui Gou, Qiman Wu, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. Rtformer: Efficient design for real-time semantic segmentation with transformer. *Advances in Neural Information Processing Systems*, 35:7423–7436, 2022. 3

[33] Yu Wang, Yi Niu, Peiyong Duan, Jianwei Lin, and Yuanjie Zheng. Deep propagation based image matting. In *IJCAI*, pages 999–1006, 2018. 1

[34] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3

[35] 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. 1, 2, 6

[36] Haichao Yu, Ning Xu, Zilong Huang, Yuqian Zhou, and Humphrey Shi. High-resolution deep image matting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3217–3224, 2021. 3

[37] Yunke Zhang, Chi Wang, Miaomiao Cui, Peiran Ren, Xuansong Xie, Xian-Sheng Hua, Hujun Bao, Qixing Huang, and Weiwei Xu. Attention-guided temporally coherent video object matting. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5128–5137, 2021. 3

[38] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 3

[39] Bingke Zhu, Yingying Chen, Jinqiao Wang, Si Liu, Bo Zhang, and Ming Tang. Fast deep matting for portrait animation on mobile phone. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 297–305, 2017. 3