BACKGROUND MATTING VIA RECURSIVE EXCITATION

Junjie Deng†, Yangyang Xu†, Zeyang Zhou and Shengfeng He§

School of Computer Science and Engineering, South China University of Technology, China

ABSTRACT

We propose a simple yet effective technique that significantly improves the performance of the current state-of-the-art background matting model without compromising its original speed. We achieve this by carefully exciting the proper neural activations using an excitation map in the training phase and performing recursive inference in the testing phase. To avoid being over-reliant on perfect excitations, we follow the idea of curriculum learning to divide the training phase into three easy-to-hard stages and gradually shift the excitation map from GT alpha matte to pseudo GT alpha matte. In the testing phase, we propose a recursive inference mechanism that uses the output alpha matte as the excitation map to further refine the output alpha matte. Our method is a simple plug-in for arbitrary matting models. Compared with the original ones, the enhanced models alleviate the problem of performance degradation with complex background and thus boosts the matting accuracy.

Index Terms—image matting, recursive excitation, curriculum learning

1. INTRODUCTION

Image matting refers to accurately extracting the foreground object of an image. The estimated opacity (alpha) value can be further used to absorb the foreground object and composed with a new background image to render a new one, which has been well-studied in both research and industrial communities. Mathematically, given an input image $I$, we assume it is composed by the foreground image $F$, background image $B$, and the alpha matte $\alpha$ by the following equation:

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

where $i = (x, y)$ denotes the pixel position in the image.

It can be seen that image matting is an ill-posed problem that given 3 known values but 7 unknown values need to be solved for each pixel. Therefore, auxiliary inputs are necessary to solve this problem. One solution is to turn the unknown background image to a blue/green screen, known as Blue Screen Matting [1].

†The first two authors contributed equally to this work.
§Corresponding author (hesfe@scut.edu.cn).

(a) Input Image (b) BGM (c) BGM* (d) GT

Fig. 1: Alpha matte comparison between the background matting (BGM) and the enhanced model (BGM*) with our proposed excitation mechanism.

However, a green screen studio is not so common in real-world scenes. As a result, a user-designed trimap which indicates foreground, background and unknown regions is further introduced as extra prior information to constrain the solution space. Traditional matting approaches [2–5] require a trimap as input and rely on the prior assumptions on image statistics to predict the alpha values of unknown regions based on the foreground/background areas. Deep learning based approaches [6–9] take the advantages of large scale datasets and show the superior performance compared to traditional methods in the presence of trimap. Although some approaches [10–12] focus on matting without any external input, these approaches often end up with inferior results. More recently, background matting (BGM) [13, 14] is proposed to replace the trimap with an additional background image for providing the auxiliary information, and achieves state-of-the-art performance. Though much progress has been gained, those learning based works still suffer from ambiguous semantic and discrepant alpha values, especially for the images with complex background. We argue that they do not take full use of the guidance information, which leads the inaccurate feature extraction.

In this paper, inspired by curriculum learning [15, 16], we propose a simple yet effective way for further absorbing the guidance from the ground truth (GT). In particular, we excite the output feature maps of the certain layer of the background matting model carefully as an additional guideline, to help the network learn the more accurate features and converge faster during training. During testing, we utilize the recursive
excitation to further refine the output for obtaining a better result. Our curriculum learning-based block shows a strong ability for natural image matting with complex background. As shown in Fig. 1, we present two samples at first column that share the same foreground images and ground truth alpha matte (shown in the last column). The first sample is composed with an easy background and the second one has a much complex background. For the first input image, both models are capable of producing a precise alpha matte. But for the second input image that has a much-complex background, the original BGM model fails to generate reliable results while the BGM plugged with our proposed matting block shows strong robust performance. Besides, our method is super lightweight and can plug into many existing matting models. Our contributions are summarized as follows:

- We propose a feature activation excitation mechanism for background matting with a curriculum learning strategy that takes full use of the auxiliary guidance information, producing more accurate features and faster convergence. We further demonstrate the generalization ability to other models.
- We present an online refinement that allows recursively inferencing better matting alpha mattes.

2. RELATED WORK

Traditional Image Matting. Traditional matting approaches can roughly be divided into affinity-based approaches [2,3,10,17–20] and sampling-based approaches [4,5,21–24]. The former solves alpha values iteratively from the foreground and background regions to unknown region, while the latter holds the assumption that the observed color in the unknown region can always be sampled from the foreground and background regions.

Deep Image Matting. Recent works have achieved great improvement by training the neural networks on the large scale dataset [6–9]. Xu et al. [6] proposes a large image matting dataset and an encoder-decoder based network that estimates alpha matte directly. Tang et al. [25] combines learning-based approaches with traditional sampling-based techniques to estimate much accurate alpha mattes. Lu et al. [8] propose a learning-based upsampling and pooling operators for matting problem. Li et al. [26] propose a contextual attention mechanism to propagate high-level opacity information from low-level affinity. However, they need the pre-annotated trimap as the input, which is a practical problem in real-world scenarios. Hence, there are many works aim at getting rid of additional annotated inputs. In particular, Zhang et al. [11] proposes a segmentation network with two decoder branches for foreground and background classification followed by a fusion network for alpha matte estimation. Qiao et al. [12] employs spatial and channel-wise attention to construct an end-to-end network which can predict alpha mattes from single RGB image.

More recently, Sengupta et al. [13] proposes a background matting model that replaces the trimap with an additional background image, which reliefs the demand of annotated knowledge and achieves state-of-the-art performance. Following the spirit of background matting, Lin and Ryabtsev et al. [14] employ two neural networks to predict results in a coarse-to-fine manner, which yields higher quality results while achieves a dramatic boost in speed and resolution. However, these methods usually predict an inferior results in complex background since they cannot take full use of the auxiliary guidance. In this paper, we follow the idea of curriculum learning, and propose a feature activation excitation mechanism, which takes full use of the auxiliary knowledge and boosts the existing performance.

3. METHOD

3.1. Excitation Formulation

Current matting models degrade performances in the scenarios of complex background. Existing works solve this problem by introducing additional training data or designing a complicated network architecture. In contrast, we propose an excitation mechanism based on curriculum learning [15], that takes full advantages of the guidance from the auxiliary guidance information. The pipeline of our excitation mechanism is illustrated in Fig. 2. Given an input feature map \( F_{in} \) produced by a matting model, an excitation map \( E \), we can obtain the excited feature map \( F_e \) by applying element-wise multiplication, which can be presented as:

\[
F_e = F_{in} \ast E. \tag{2}
\]

Then, a learnable parameter \( \gamma \) is further introduced to control the intensity of the excitation map:

\[
F_{out} = \gamma \ast F_e + (1 - \gamma) \ast F_{in}, \tag{3}
\]

where \( F_{out} \) is the final output feature map from the excitation layer. We delegate this parameter \( \gamma \) to the network to find an optimum value that balances the contribution between the excited feature map and the original feature map. We initialize \( \gamma \) with a value of 0.5 at the beginning of training phase.
Combining Eq. 2 and Eq. 3, our excitation mechanism can strengthen the activations of the feature map at foreground area which is indicated by the excitation map \( E \) and suppress the responses at the background area which is not effective on the matting tasks. As shown in Fig. 3, after our excitation, the excited feature map reduces the influence of background and more dedicated to the foreground region. As a result, the boosted matting model can handle the complex background scenes more effectively.

### 3.2. Enhanced Background Matting

Due to the excellent performance of Background Matting [13], we adopt the basic network structure of BGM which can jointly predict the alpha matte and the foreground. Besides, we also introduce an excitation map to guide the network to predict more accurate results. As shown in Fig. 4, we add an excitation layer to excite the certain activations of feature map derived from the first residual block. Similar to BGM [13], the input of Enhanced Background Matting network consists of an image \( I \) with foreground, a perturbed version of background \( B' \), a soft segmentation \( S \), and an additional excitation map \( E \). The network \( G \) can be denoted as:

\[
F, \alpha = G(I, B', S, E),
\]

where \( F \) and \( \alpha \) denote the predicted foreground image and alpha matte respectively. In following sections, we will present the details of excitation during training and testing phrases. Meanwhile, we will plug it into many existing models in Sec. 4.5.

### 3.3. Excitation During Training

Since the neural activation on the feature map reflects the final output result, the excitation map has a strong connection with the alpha matte generated by the network. We utilize the \( GT_{\text{pseudo}} \) alpha mattes produced by a pre-trained deep matting network called IndexNet [8], along with \( GT \) alpha mattes provided from the dataset, to composite excitation maps. Inspired by curriculum learning, we divide the training phase into three stages from easy to hard. It can be formulated as follows:

\[
E = \begin{cases} 
GT & \text{epoch} \leq N_1, \\
GT \ast \beta + GT_{\text{pseudo}} \ast (1 - \beta) & N_1 < \text{epoch} \leq N_2, \\
GT_{\text{pseudo}} & \text{epoch} > N_2.
\end{cases}
\]

In the first stage, we utilize the \( GT \) alpha matte to be the excitation map for exciting the neural activations correctly. Besides, it accelerates the convergence speed during the beginning of the training phase significantly. In the second stage, we gradually turn the excitation map from \( GT \) to \( GT_{\text{pseudo}} \) by linearly decreasing the value of \( \beta \) from 1 to 0 with the following equation:

\[
\beta = 1 - \frac{\text{epoch} - N_1}{N_2 - N_1}.
\]

We excite the certain neural activation with an imperfect excitation map during this stage, forcing the network to be tolerance to real-world cases, and relief the demand of \( GT \). At the third stage, we use the \( GT_{\text{pseudo}} \) matte as the excitation map for improving the generalization performance of the network.

### 3.4. Recursive Excitation During Testing

In our design, the quality of the excitation map plays an important role in final alpha matte estimation. In training phrase, we slowly shift the input excitation map from \( GT \) to \( GT_{\text{pseudo}} \) and obtain a robust matting model. During the testing, we can use this model directly. However, our model still
benefits from the accurate excitation map that properly excites the certain activations. Hence, we propose a recursive excitation testing strategy. In particular, as shown in the dotted lines in Fig. 4, we reuse the output alpha mattes as the excitation map to refine the matting model, that is, we apply the third stage in Eq. 5 on the testing data to obtain the better performance. The recursive excitation procedure can be formulated as:

\[(F_{i+1}, \alpha_{i+1}) = G(F_i, B', S, \alpha_i),\]  

(7)

where \(F_i\) and \(\alpha_i\) are foreground image and alpha matte of network output in iteration \(i\).

3.5. Loss Function

We only train our network with the Adobe Image Matting Dataset [6] which provides ground truth foreground image \(F^*\) and alpha matte \(\alpha^*\). Hence, we adopt the same supervised loss as BGM to optimize our network:

\[L = \|\alpha - \alpha^*\|_1 + \|\nabla(\alpha) - \nabla(\alpha^*)\|_1 + 2\|F - F^*\|_1 + \|I - \alpha F - (1 - \alpha)B\|_1,\]  

(8)

where \(B\) is ground truth backgrounds of the input images \(I\), \(\|\cdot\|_1\) is the \(l_1\) distance between two inputs and \(\nabla(\cdot)\) denotes the gradient map of input. The gradient term loss of \(\alpha\) is beneficial for the network to predict sharper results [11].

4. EXPERIMENT

4.1. Implementation Details and Evaluation Metrics

Our model is implemented by following exactly the same settings as BGM. We re-train both BGM and the enhanced version of BGM with the same number of epoch, same learning rate, and same batch-size to avoid unfair comparison. We divide the training phase of the enhanced BGM into three easy-to-hard stages and set the value of \(N_1, N_2\) to be 2, 8, respectively. We also compare with a natural image matting model IndexNet [8] from reference. The parameter number of our model is noted in Table 1.

### Table 1

<table>
<thead>
<tr>
<th>Models</th>
<th>SAD↓</th>
<th>MSE↓</th>
<th>CONN↓</th>
<th>GRAD↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>IndexNet</td>
<td>3.9872</td>
<td>0.0168</td>
<td>3.7657</td>
<td>5.7170</td>
</tr>
<tr>
<td>Original BGM</td>
<td>2.0657</td>
<td>0.0054</td>
<td>1.9623</td>
<td>1.8602</td>
</tr>
<tr>
<td>BGM(_s=1)</td>
<td>1.7206</td>
<td>0.0037</td>
<td>1.1708</td>
<td>1.4910</td>
</tr>
<tr>
<td>BGM(_s=2)</td>
<td>1.7123</td>
<td>0.0035</td>
<td>1.0798</td>
<td>1.4733</td>
</tr>
<tr>
<td>BGM(_s=3)</td>
<td>1.7394</td>
<td>0.0038</td>
<td>1.2491</td>
<td>1.5103</td>
</tr>
</tbody>
</table>

4.2. Quantitative Results

Table 1 shows the quantitative evaluation on the IndexNet, original BGM, and enhanced BGM (BGM\(_s=2\)) with different steps in the recursive testing. BGM\(_s=1\) means we use the output alpha matte as the excitation map. We can see that the BGM\(_s=2\) has gained a better result than the original BGM and IndexNet, and BGM\(_s=3\) achieves the best performance over four evaluation metrics by a large margin, demonstrating the effectiveness of our proposed excitation mechanism. However, BGM\(_s=3\) decreases its performance which we believe due to overfitting. Hence, we choose \(s = 2\) as the recursive testing setting on the enhanced networks in the following experiments.

4.3. Qualitative Results

Fig. 5 shows qualitative evaluation on the origin BGM and enhanced BGM (BGM\(_s=2\)). The enhanced BGM can obtain much better result when compare to original BGM. Even if the background is very complex (e.g. Fig. 5c, 5d, 5e), which indicates that our excitation mechanism can avoid the influence of the complex background meanwhile boosting the matting performance.

4.4. Robust Comparison

For evaluating the robustness of our excitation mechanism, we conduct experiment across different background images. In particular, the testing dataset of Adobe Image Matting Dataset [6] has 11 foreground images, and each foreground image is composited with 20 random background images. We treat the composited images with the same foreground image as a category, then calculate the variance of these four metrics over 11 categories, finally obtain four average values on four evaluation metrics. Table 2 shows the comparison results between the original BGM and our enhanced BGM. We can see that our enhanced BGM has achieved lower variance in
Fig. 5: Qualitative evaluation between the original BGM and the enhanced BGM. We can see that the enhanced BGM improves the matting results significantly. (Zoom in for a better view.)

Table 3: Quantitative evaluation by applying our method to other matting models.

<table>
<thead>
<tr>
<th>Models</th>
<th>SAD↓</th>
<th>MSE↓</th>
<th>CONN↓</th>
<th>GRAD↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original IndexNet</td>
<td>3.9872</td>
<td>0.0168</td>
<td>3.7657</td>
<td>5.7170</td>
</tr>
<tr>
<td>Original DIM</td>
<td>4.2620</td>
<td>0.0146</td>
<td>4.0114</td>
<td>5.7514</td>
</tr>
<tr>
<td>Enhanced IndexNet</td>
<td>3.8851</td>
<td>0.0164</td>
<td>3.6812</td>
<td>5.5117</td>
</tr>
<tr>
<td>Enhanced DIM</td>
<td>3.5715</td>
<td>0.0141</td>
<td>3.6458</td>
<td>5.8304</td>
</tr>
</tbody>
</table>

all four metrics, which demonstrate that the enhanced BGM is more robust than the original one. That should contributed to our excitation mechanism, the enhanced BGM can avoid the influence of complex background and more concentrating on the foreground.

4.5. Generalization

Our proposed excitation mechanism is very simple and light-weight, it can be plugged into any matting networks. To demonstrate the generalization of our proposed mechanism, we plug the excitation mechanism into two recent deep matting models, IndexNet [8] and Deep Image Matting (DIM) [6]. Quantitative results on two matting models and their enhanced one are shown in Table 3. The enhanced versions of two models achieve better results compared to the original ones, demonstrating the effectiveness and the generalization ability of our excitation mechanism.

5. CONCLUSION

In this paper, we propose a novel excitation mechanism for enhancing the performance of background matting as well as other matting models. By carefully exciting certain activations guided by real GT and pseudo GT, our enhanced models absorb more knowledge from the guidance information, and produce a more precise alpha matte than the original ones, especially for the samples with complex backgrounds. Furthermore, we can refine the alpha matte with recursive inference strategy. Besides, our excitation mechanism is light-weighted, and can be plugged into any existing models meanwhile boosting their performances.

Acknowledgement. This project is supported by the National Natural Science Foundation of China (No. 61972162); Guangdong International Science and Technology Cooperation Project (No. 2021A0505030009); Guangdong Natural Science Foundation (No. 2021A1515012625); Guangzhou Basic and Applied Research Project (No. 202102021074); and CCF-Tencent Open Research fund (RAGR20210114).
6. REFERENCES


