Alpha Local Difference Loss Function for Deep Image Matting

Jiehong Li[1], Peijie Huang[1(✉)], Wensheng Li[2], and Yihui Liang[2]

[1] College of Mathematics and Informatics, South China Agricultural University, Guangzhou, China
pjhuang@scau.edu.cn

[2] School of Computer, University of Electronic Science and Technology of China, Zhongshan Institute, Zhongshan, China

Abstract. In recent years, deep learning-based matting methods have received increasing attention due to their superior performance. The design of the loss function plays a important role in the performance of matting models. Existing loss functions train the network by supervising it to learn specific value, gradient, and detailed information of the ground-truth alpha matte. However, these loss functions only supervise network learning based on the value of alpha matte, and the matting network may not fully understand the uniqueness of the matting task. We introduce a loss function which supervises image features. On one hand, it effectively extracts useful information from the ground-truth alpha. On the other hand, this loss function combines the mathematical model of matting, which constrains the image features to satisfy local differences. Multiple experiments have shown that our loss function enhances the generalization ability of matting networks.

Keywords: Natural Image Matting · Loss Function · Deep Learning

1 Introduction

Image matting is a challenging tasks in computer vision that aims to separate the foreground from a natural image by predicting the transparency of each pixel. It has been applied in the field of biometric recognition, such as finger-vein [1], gait recognition [2,3], and face verification [4], as it can finely delineate the target contours, thus facilitating biometric recognition tasks.

The image $\mathbf{I}$ can be represented as a convex combination of the foreground $\mathbf{F}$ and the background $\mathbf{B}$.

$$\mathbf{I}_i = \alpha_i \mathbf{F}_i + (1 - \alpha_i)\mathbf{B}_i \qquad \alpha_i \in [0, 1] \tag{1}$$

where α_i, $\mathbf{F}_i$, and $\mathbf{B}_i$ respectively represent the transparency, foreground color, and background color at position i in the image. This problem is a highly underdetermined mathematical problem. There are three unknowns and only one known in the equation. The trimap is introduced to provide additional constrains. It consists of three parts: the known foreground region where the alpha

W. Jia et al. (Eds.): CCBR 2023, LNCS 14463, pp. 407–416, 2023.
https://doi.org/10.1007/978-981-99-8565-4_38

value is known to be 1, the known background region where the alpha value is 0, and an unknown region where the alpha value needs to be determined. Existing deep learning-based matting methods have greatly surpassed traditional methods in terms of the quality of alpha mattes, attracting a rapid increase in attention to deep learning-based matting methods.

The loss function is a fundamental component of deep learning, as it measures the difference between the predicted output of a model and the true labels. It provides guidance for model training and optimization objectives, allowing the model to gradually improve its prediction accuracy. The alpha prediction loss is computed as the average absolute difference between the predicted alpha matte and the ground-truth alpha matte. The composition loss, introduced by [5], utilizes the ground-truth foreground and background colors to supervise the network at the pixel level. Gradient loss [6] has been proposed to improve the sharpness of the predicted alpha matte and reduce excessive smoothness. The Laplacian Pyramid loss [7], a multi-scale technique, is employed to measure the disparities between the predicted alpha matte and the ground-truth alpha matte in local and global regions. Indeed, the loss functions used for image matting encompass supervision at the pixel level as well as supervision of the gradient and detail changes in the alpha channel, which improves the accuracy and quality of the matting results. But these loss functions only focus on the differences between the alpha matte predicted by the network and the ground-truth alpha matte. Consequently, the network may not effectively learn valuable information inherent in the ground-truth across different feature layers. In general, increasing the depth of a neural network can improve its representation ability to some extent. To better train the network, it is common to add auxiliary supervision to certain layers of the neural network. Some methods [8,9] supervise the multi-scale features obtained by the decoder at different scales. However, directly supervising neural networks with ground-truth alpha mattes causes the decoder at a small scale to strictly approximate the ground-truth alpha mattes, which may result in overfitting. Figure 1 provides an example. When the image matting method is applied to scenarios different from the training images, the prediction of the decoder at a small scale may not be accurate. Any prediction error of the decoder would degrade the quality of alpha mattes.

We introduce a loss function called Alpha Local Difference Loss(ALDL), which leverages the local differences within the ground-truth to supervise features at various resolution scales. Unlike gradient loss, ALDL captures the differences between the pixel and its surrounding pixels in the ground-truth, and utilizes these differences as constraints to supervise the features of the image. Gradient loss only describes the gradient of the central pixel in the x and y directions, without explicitly capturing the specific variations between the central pixel and local surrounding pixels. Furthermore, instead of applying strict supervision on early decoders [8,9], ALDL is a loose supervision that leads the matting network to learn the relationships between features, rather than strictly adhering to specific numerical values.

This work's main contributions can be summarized as follows:

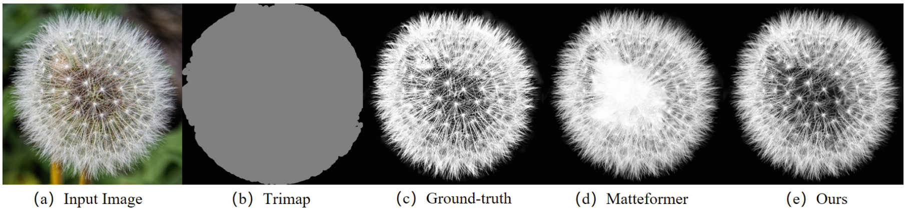
(a) Input Image (b) Trimap (c) Ground-truth (d) Matteformer (e) Ours

Fig. 1. From left to right, the images are the input, trimap, ground-truth, the predicted results by the MatteFormer and ours. We can see that there are serious errors in the prediction of the intermediate details of alpha. These errors are the result of inaccurate alpha prediction caused by low-resolution feature estimation.

1. We propose a loss function called Alpha Local Difference Loss specifically designed for matting networks, which utilizes the supervision of local feature relationships. This loss function can be easily integrated into existing networks with hardly any need to add extra parameters.
2. Through experiments conducted on multiple networks and datasets, our Alpha Local Difference Loss demonstrates the ability to improve the generalization capability of matting networks, resulting in enhanced object details in the matting process.

2 Methodology

In this section, we illustrate how to define the difference between each point and its local neighboring points based on the local information of the ground-truth alpha. The local difference is embedded into the image features, and the Alpha Local Difference Loss is proposed to constrain the network in learning this difference. Furthermore, an analysis is conducted to determine which features in the neural network should be supervised.

2.1 Local Similarity of Alpha Labels and Features

Consistent with the assumption of closed-form matting [10], we assume that pixels within a local region have the same foreground color F and background color B. According to Eq. (1), we can obtain the pixel value difference ΔI between two points $\mathbf{x}$ and $\mathbf{y}$ within a local region. Similarly, by using the ground-truth alpha, we can also obtain the alpha value difference $\Delta\alpha$ between point $\mathbf{x}$ and $\mathbf{y}$.

$$I_x - I_y = \alpha_x F + (1-\alpha_x) - \alpha_y F - (1-\alpha_y)B = (\alpha_x - \alpha_y)(F - B) \tag{2}$$

$$\Delta I = \Delta\alpha(F - B) \tag{3}$$

$$\Delta f = \Delta\alpha(f_F - f_B) \tag{4}$$

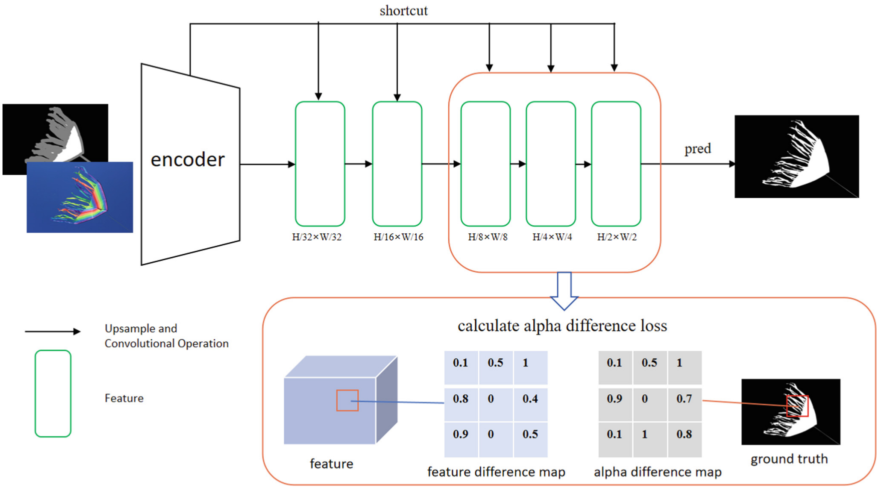

Fig. 2. The process of calculating ALDL

It can be observed that there is a linear relationship between the color value difference ΔI and the $F - B$ within a local region on the image. Because F and B are invariant within the local region, $F - B$ is a fixed vector. By analogy, we can consider feature difference Δf as a linear combination of features f_F and f_B. In spatial terms, for two features f_F and f_B within a local region, Eq. (4) is obtained. The features should also be constrained to satisfy this relationship as much as possible. This relationship embodies the intrinsic meaning of matting, and it is believed that it will help the network learn to synthesize Eq. (1).

2.2 The Design of Loss Function

For a position i, let $\partial\{i\}$ denote the set of points within the M1 × M2 region R, where M1 and M2 respectively denote the height and width of the R, and pixel i is located at the center position of the R. The set of values for the ground-truth alpha at position i is: $\partial\{\alpha_i\} = \{\alpha_{i1}, \alpha_{i2}, \alpha_{i3}, ..., \alpha_{iM1\times M2}\}$. It is worth noting that α_{ij} is a scalar. We can compute the differences between α_i and each element in its set $\partial\{\alpha_i\}$.

$$dif\left(\alpha_i, \alpha_{ij}\right) = \alpha_i - \alpha_{ij} \tag{5}$$

$$sim_\alpha\left(\alpha_i, \alpha_{ij}\right) = 1 - \left|dif\left(\alpha_i, \alpha_{ij}\right)\right| \tag{6}$$

$$sim_f\left(f_i, f_{ij}\right) = \varphi(\cos(norm(f_i), norm(f_{ij}))) \tag{7}$$

$$loss = \sum_i \sum_j sim_\alpha\left(\alpha_i, \alpha_{ij}\right) - sim_f\left(f_i, f_{ij}\right) \tag{8}$$

$dif\left(\alpha_i, \alpha_{ij}\right)$ represents the difference between the alpha value of the central pixel i and the alpha values of other positions within the region R. To facilitate computation, we normalize the values between 0 and 1 using the sim_α function. The smaller the difference between α_i and α_{ij}, the closer the value of sim_α tends to approach 1. Given the feature $X \in R^{H/r \times W/r \times C}$, for any point at the location i in X, $\partial\left\{f_i\right\} = \left\{f_{i1}, f_{i2}, f_{i3}, ..., f_{iM1 \times M2}\right\}$, where $f_{iM1 \times M2} \in R^{1 \times 1 \times C}$, r is the downsampling factor. In order to align the resolution of alpha with the feature, the ground-truth alpha is downsampled to obtain $\partial\left\{\alpha_i^r\right\}$. Each element in the set $\partial\left\{\alpha_i^r\right\}$ and $\partial\left\{f_i\right\}$ corresponds to each other based on their spatial positions. It is worth noting that our goal is to correspond the vector Δf with the scalar $\Delta\alpha$, so the similarity between the two features is calculated to convert the vector into a scalar. The definitions of distance between features is (7), $norm(f_i)$ denotes the calculation of the norm of vector f_i, φ represents a mapping function, cos refers to the calculation of the cosine similarity. The aim is to maintain consistency in terms of both the differneces of alpha values and the differneces of features between each point and its neighboring adjacent points. Hence, the definition of Alpha Local Difference Loss is Eq. (8).

2.3 The Supervisory Position of ALDL

[11] indicates that different layers in a convolutional neural network tend to learn features at different levels. Shallow layers learn low-level features such as color and edges and the last few layers learn task-relevant semantic features. If the features at shallow layers are supervised to capture task-related knowledge, the original feature extraction process in the neural network would be overlooked. Therefore, we only supervise the features outputted by the decoder. Additionally, our supervision relationship is derived from the ground-truth alpha in local regions, which can be considered as extracting features at a lower-level semantic level. Alpha Local Difference Loss should not be used for supervising features representing higher-level semantic features with very low resolution. As shown in the Fig. 2, taking MatteFormer [9] as an example, its decoder outputs features with resolutions of 1/32, 1/16, 1/8, 1/4, and 1/2. Supervision is only applied to the features with resolutions of 1/8, 1/4, and 1/2 in the decoder, while the feature with a resolution of 1 is not supervised in order to reduce computational cost.

3 Experiments

To validate the effectiveness of the suggested Alpha Local Difference Loss function, we extensively perform experiments on various matting baselines using multiple benchmark datasets. The performance is assessed in real-world scenarios to verify its generalization capability.

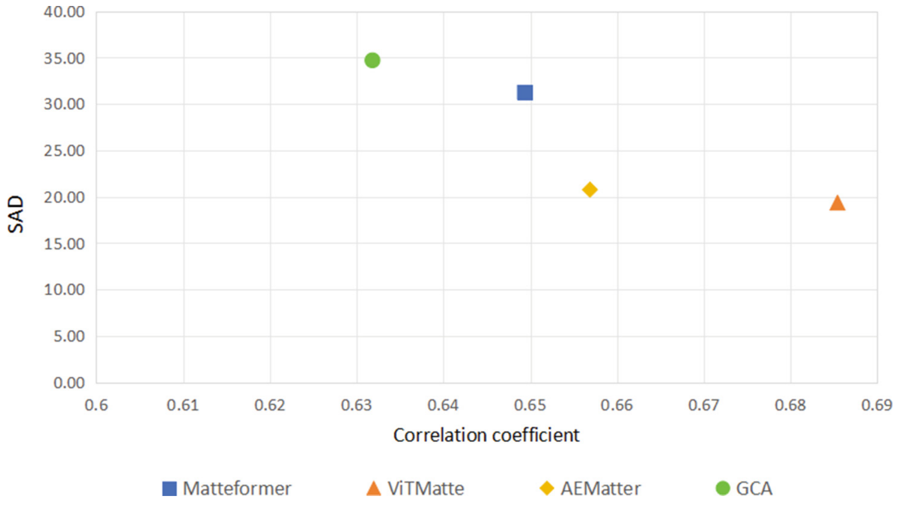

Fig. 3. Y-axis: the SAD error on AIM-500. X-axis: the correlation coefficient between the difference of alpha and the difference of feature.

Table 1. The effectiveness of implementing ALDL

	AIM-500				AM-2K				P3M			
	MSE	SAD	Grad	Conn	MSE	SAD	Grad	Conn	MSE	SAD	Grad	Conn
GCA	40.00	**35.25**	27.86	35.89	**14.49**	**9.23**	8.92	**8.17**	17.49	8.50	13.41	7.90
GCA+ALDL	**38.33**	35.65	**27.82**	**36.13**	15.28	9.45	**8.75**	8.50	**16.98**	**8.27**	**12.92**	**7.74**
MatteFormer	34.32	31.25	23.53	31.25	17.00	9.93	9.81	9.16	22.62	10.07	**14.39**	9.63
MatteFormer+ALDL	**31.11**	**28.23**	**21.79**	**28.21**	**15.68**	**9.68**	**9.73**	**8.93**	**21.42**	**9.93**	14.48	**9.51**
VitMatte	15.69	19.35	12.99	18.74	7.65	6.50	6.07	5.46	11.34	6.40	10.33	6.79
VitMatte+ALDL	**15.45**	**17.76**	**12.85**	**17.09**	**7.87**	**6.44**	**5.83**	**5.41**	**10.84**	**6.16**	**9.82**	**5.56**

- MSE values are scaled by 10^{-3}
- The best results are in bold

3.1 Datasets and Implementation Details

We train models on the Adobe Image Matting [5] dataset and report performance on the real-world AIM-500 [8], AM-2K [12], P3M [13]. AIM-500 contains 100 portrait images, 200 animal images, 34 images with transparent objects, 75 plant images, 45 furniture images, 36 toy images, and 10 fruit images. The AM-2k test set comprises 200 images of animals, classified into 20 distinct categories. P3M-500-NP contains 500 diverse portrait images that showcase diversity in foreground, hair, body contour, posture, and other aspects. These datasets comprise a plethora of human portrait outlines and exhibit numerous similarities to datasets employed for tasks like gait recognition and other biometric recognition tasks. Our implementation is based on PyTorch. No architectural changes are required. We only modify the loss function. The height M1 and width M2 of the local region R are both set to 3, and the center position of R belongs to an unknown region in the trimap. Various matting models utilize distinct loss functions. In order to effectively illustrate the efficacy of ALDL, we directly incorporate ALDL into the existing loss function. In line with the approach outlined in [14], four widely adopted metrics are employed to assess the quality of the predicted alpha matte. These metrics include the sum of absolute differences (SAD), mean squared errors (MSE), gradient errors (Grad), and connectivity errors (Conn). Four matting baselines, namely: GCA Matting [15], MatteFormer [9], VitMatte [16], AEMatter [17] are evaluated. GCA implements a guided contextual attention module to propagate opacity information based on low-level features. MatteFormer introduces prior-token for the propagation of global information. VitMatte proposes a robust matting method based on Vit [18].

3.2 Proof of the Local Similarity Hypothesis Between Alpha and Feature

In order to validate the effectiveness of the local similarity hypothesis in improving image matting, during the inference stage, we extracted the feature outputs from the intermediate layer. Based on (6) and (7), the correlation coefficient of sim_α and sim_f for each point in the unknown region of the trimap have been calculated. It can be observed that the higher the correlation coefficient, the better the matting performance of the method from Fig. 3. This indicates that if the features satisfy the local differences defined by ground-truth alpha, it can improve the quality of the matting.

3.3 Generalization

ALDL was applied to three different baselines and compared with their counterparts without ALDL, as shown in the Table 1. It can be observed that for MatteFormer and VitMatte, ALDL improves their generalization ability on three datasets. This suggests that constraining the relationships between local features can help the network better understand the matting task. The combination of GCA with ALDL demonstrates its generalization ability, particularly on the P3M dataset. GCA incorporates a shallow guidance module to learn feature relationships, but evaluating the quality of these relationships poses a challenge. In contrast, ALDL explicitly constrains local feature relationships using ground-truth alpha, aligning with the objective of GCA's shallow guidance module. Consequently, the addition of ALDL to GCA results in moderate performance improvements on the AIM-500 and AM-2K datasets. GCA consistently performs well according to the Grad metric, indicating that ALDL excels at capturing intricate details, accurately defining contours, and proves advantageous for downstream tasks involving matting.

Table 2. Ablation experiment of ALDL

			AIM-500				AM-2K				P3M			
MatteFormer	R1	R2	MSE	SAD	Grad	Conn	MSE	SAD	Grad	Conn	MSE	SAD	Grad	Conn
			34.32	31.25	23.53	31.25	17.00	9.93	9.81	9.16	22.62	10.07	14.39	9.63
	✓		**26.99**	**23.95**	**20.59**	**23.41**	**15.42**	**9.10**	**9.14**	**8.23**	**18.11**	**8.69**	**13.32**	**8.16**
		✓	33.16	29.10	23.23	28.92	15.63	9.39	9.94	8.52	22.18	9.74	14.65	9.28
GCA			40.00	**35.25**	27.86	35.89	**14.49**	**9.23**	8.92	**8.17**	17.49	8.50	13.41	7.90
	✓		**38.33**	35.65	**27.82**	**36.13**	15.28	9.45	**8.75**	8.50	**16.98**	**8.27**	**12.92**	**7.74**
		✓	39.04	35.23	28.34	35.82	15.39	9.81	8.96	8.80	17.31	8.53	13.26	7.80

- MSE values are scaled by 10^{-3}
- R1: ALDL supervises features with resolutions of 1/2, 1/4, 1/8. R2: ALDL supervises features with all resolutions of decoder output.

3.4 Ablation Study of Deep Supervision

An ablation experiment was conducted using MatteFormer, as its decoder's output features are supervised with ground-truth. The difference is that ALDL supervises the local differential relationships between features, while MatteFormer directly supervises the alpha values at the feature level. As shown in the Table 2, MatteFormer marked with R1 or R2 denotes removing the structure that originally outputs alpha values from the decoder and instead directly supervising the feature level with ALDL. GCA marked with R1 or R2 represents the application of the ALDL to the intermediate layer features of the decoder. Experimental results demonstrate that applying ALDL to features, which is a relatively weak constraint, yields better performance than directly supervising with alpha values. Additionally, since ALDL explores local information from ground-truth, which essentially belongs to low-level features, it is more suitable for shallow features rather than deep features.

4 Conclusion

This study focuses on the loss function of deep image matting methods. We analyzed the shortcomings in the loss functions of existing matting models, and proposed the alpha local difference loss function, which takes the ground-truth alpha matte and the composition formula of image matting as the starting point, to supervise the image features. Extensive experiments are performed on several test datasets using state-of-the-art deep image matting methods. Experimental results verify the effectiveness of the proposed ALDL and demonstrate that ALDL can improve the generalization ability of deep image matting methods.

Acknowledgements. This work was supported in part by National Natural Science Foundation of China (62002053,62276103), Guangdong Basic and Applied Basic Research Foundation (2020A1515110504), Natural Science Foundation of Guangdong Province (2021A1515011866, 2020A1515010696, 2022A1515011491), and of Sichuan Province (2022YFG0314), the Guangdong University Key Platforms and Research

Projects (2018KZDXM066), Key Research and Development Program of Zhongshan (2019A4018), the Major Science and Technology Foundation of Zhongshan City (2019B2009, 2019A40027, 2021A1003), and Zhongshan Science and Technology Research Project of Social welfare (210714094038458, 2020B2017), Natural Science Foundation of Guangdong Province (2021A1515011864).

References

1. Yang, J., Shi, Y.: Finger-vein network enhancement and segmentation. Pattern Anal. Appl. **17**, 783–797 (2014)
2. Hofmann, M., Schmidt, S.M., Rajagopalan, A.N., Rigoll, G.: Combined face and gait recognition using alpha matte preprocessing. In: 2012 5th IAPR International Conference on Biometrics (ICB), pp. 390–395. IEEE (2012)
3. Han, Y., Wang, Z., Han, X., Fan, X.: Gaitpretreatment: robust pretreatment strategy for gait recognition. In: 2022 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI), pp. 1–6. IEEE (2022)
4. Wasi, A., Gupta, M.: Dfvnet: real-time disguised face verification. In: 2022 IEEE India Council International Subsections Conference (INDISCON), pp. 1–5. IEEE (2022)
5. Xu, N., Price, B., Cohen, S., Huang, T.: Deep image matting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2970–2979 (2017)
6. Tang, J., Aksoy, Y., Oztireli, C., Gross, M., Aydin, T.O.: Learning-based sampling for natural image matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3055–3063 (2019)
7. Hou, Q., Liu, F.: Context-aware image matting for simultaneous foreground and alpha estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4130–4139 (2019)
8. Yu, Q., et al.: Mask guided matting via progressive refinement network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, PP. 1154–1163 (2021)
9. Park, G., Son, S., Yoo, J., Kim, S., Kwak, N.: Matteformer: transformer-based image matting via prior-tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11696–11706 (2022)
10. Levin, A., Lischinski, D., Weiss, Y.: A closed-form solution to natural image matting. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 228–242 (2007)
11. Zhang, L., Chen, X., Zhang, J., Dong, R., Ma, K.: Contrastive deep supervision. In: Avidan, S., Brostow, G., Cisse, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13866, pp. 1–19. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-19809-0_1
12. Li, J., Zhang, J., Maybank, S.J., Tao, D.: Bridging composite and real: towards end-to-end deep image matting. Int. J. Comput. Vision **130**(2), 246–266 (2022)
13. Li, J., Ma, S., Zhang, J., Tao, D.: Privacy-preserving portrait matting. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 3501–3509 (2021)
14. Rhemann, C., Rother, C., Wang, J., Gelautz, M., Kohli, P., Rott, P.: A perceptually motivated online benchmark for image matting. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1826–1833. IEEE (2009)

15. Li, Y., Lu, H.: Natural image matting via guided contextual attention. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11450–11457 (2020)
16. Yao, J., Wang, X., Yang, S., Wang, B.: Vitmatte: boosting image matting with pretrained plain vision transformers. arXiv preprint arXiv:2305.15272 (2023)
17. Liu, Q., Zhang, S., Meng, Q., Li, R., Zhong, B., Nie, L.: Rethinking context aggregation in natural image matting. arXiv preprint arXiv:2304.01171 (2023)
18. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)