electronics

MDPI

Article

Bust Portraits Matting Based on Improved U-Net

Honggang Xie *, Kaiyuan Hou *, Di Jiang and Wanjie Ma

School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan 430068, China
* Correspondence: xiehg@hbut.edu.cn (H.X.); 102110265@hbut.edu.cn (K.H.)

Abstract: Extracting complete portrait foregrounds from natural images is widely used in image editing and high-definition map generation. When making high-definition maps, it is often necessary to matte passers-by to guarantee their privacy. Current matting methods that do not require additional trimap inputs often suffer from inaccurate global predictions or blurred local details. Portrait matting, as a soft segmentation method, allows the creation of excess areas during segmentation, which inevitably leads to noise in the resulting alpha image as well as excess foreground information, so it is not necessary to keep all the excess areas. To overcome the above problems, this paper designed a contour sharpness refining network (CSRN) that modifies the weight of the alpha values of uncertain regions in the prediction map. It is combined with an end-to-end matting network for bust matting based on the U-Net target detection network containing Residual U-blocks. An end-to-end matting network for bust matting is designed. The network can effectively reduce the image noise without affecting the complete foreground information obtained by the deeper network, thus obtaining a more detailed foreground image with fine edge details. The network structure has been tested on the PPM-100, the RealWorldPortrait-636, and a self-built dataset, showing excellent performance in both edge refinement and global prediction for half-figure portraits.

Keywords: image matting; bust portraits; deep learning; driver monitoring

Citation: Xie, H.; Hou, K.; Jiang, D.; Ma, W. Bust Portraits Matting Based on Improved U-Net. *Electronics* **2023**, *12*, 1378. https://doi.org/10.3390/electronics12061378

Academic Editor: Savvas A. Chatzichristofis

Received: 24 February 2023
Revised: 9 March 2023
Accepted: 10 March 2023
Published: 14 March 2023

1. Introduction

Natural image matting is widely used in image and video processing. It cuts out the exact area in the foreground of an image by predicting the alpha value of the target. There are currently two methods for matting images: traditional matting methods and deep learning methods. The pixel value of the unknown area is predicted from the known area in the traditional methods. Among them, the more representative ones are the Ruzon method based on parameter sampling [1], the Poisson matting method [2], and the Bayesian matting method [3]. However, the traditional methods rely on trimaps and are limited by their working principle, which leads to the poor prediction of foregrounds while predicting discontinuous objects.

Unlike traditional matting methods, deep learning methods can automatically find the representational information for natural-image-matting tasks in large amounts of training data without human intervention [4]. Most of the current deep learning matting methods in the field of image matting rely on deep convolutional neural networks. It can cause image distortion during upsampling, result in rough local details in the portrait, and reduce the accuracy of the matting results. At the same time, existing matting methods face the problem of a lack of training data due to the difficulty of obtaining real transparency masks. Most of the masking methods require additional input trimaps, which undoubtedly increase the difficulty of the matting work.

To address these issues, this paper proposes an end-to-end neural network for bust matting, which focuses on the uncertain region at the edges of the bust and can obtain an alpha map with good edge detail without additional input. The work in this paper has three aspects:

1. Aiming at the problem of the automatic matting of half-body portraits, a new contour sharpness refinement network is proposed to improve the prediction of edge details;
2. Focusing on high-quality bust portrait matting improves the accuracy of predictions;
3. It produces a bust-matting dataset containing high-resolution busts and their corresponding alpha images.

2. Related Work

2.1. Matting

As shown in Equation (1), mathematical models of natural image matting were first proposed by T Porter and T Duff [5].

$$I^i = \alpha^i F^i + \left(1 - \alpha^i\right) B^i \tag{1}$$

where I is the composite image, α is the transparency of the image, F is the foreground of each image, B is the background of each image, and i is the pixel index. At this stage, the methods predict F, B, and α. In this equation, only I is known and predicting α is a pathological problem. Therefore, many image-matting methods use additional inputs to assist with the matting process, with trimaps being the most common method. The use of additional inputs leads to an increase in workload and raises the difficulty of the matting task. In recent decades, image-matting methods based on deep learning have become mainstream. Ruan et al. optimized the U-Net backbone network and proposed an edge segmentation network that improved the speed of segmenting target locations and smoothed the edges of the detected targets [6]. Traditional deep learning architectures are not convenient to use when the foreground color is close to the background color. Qiao et al. got rid of the limitation of the trimap and proposed an end-to-end hierarchical attentional matting network and successfully predicted alpha masks with a strong visual quality and matting accuracy [7]. Yihui et al. developed a multi-scale evolutionary pixel pair optimization framework that reduces computational effort while predicting high-quality alpha mattes [8]. Wang et al. separated the matting task into two stages, and by adopting a segmented refinement approach, the alpha value of the image was obtained more accurately, but the segmented approach resulted in more convolution and pooling layers, which increased the computing time [9]. Hou et al. devised a dual encoder and dual decoder model, allowing the model to output both foreground and alpha values [10]. Liu et al. proposed a novel cascaded segmented matting network that used a shared encoder and two separate decoders, achieving the end-to-end human image matting [11]. In recent years, many researchers have tried to apply a transformer to vision tasks. Park et al. designed the first transformer-based image-matting network and proposed a prior-token to be used as a global prior [12]. Cai et al. proposed a transformer-based image-matting network and used global features and a non-background mask to guide the propagation of multiscale features from the encoder to the decoder in order to maintain the context of transparent objects [13]. The transformer model can be used for matting, but is not the most suitable model. Although the transformer model can learn the relationship between different positions in a sequence through a self-attentive mechanism, for a pixel-level task such as image matting, the spatial relationship between pixels needs to be considered. The transformer model does not directly consider it, and the current transformer-based matting model suffers from the need for trimaps. In addition, the transformer takes a long time to train and is not computationally efficient. Therefore, it may not perform as well as the CNN and FCN models in the image matting task.

Qin et al. proposed a hybrid loss for boundary-aware saliency target detection by superimposing two U-Nets to enhance the sharpness of saliency target edges, in response to the problem of the blurred edges of traditional target detection [14]. For the complex instance segmentation problem, Xie et al. devised a framework to fuse target detection with instance segmentation, simplifying the computational complexity [15]. Xiao et al. proposed a multi-feature selection module to improve the accuracy of detecting foreground targets

in complex backgrounds [16]. Luo et al. proposed a multi-task collaborative network to achieve the joint learning of RES and REC, solving the problem of segmentation and detection needing to be performed separately [17]. Traditional deep learning architectures such as VGG [18], Res Net [19], and other networks are used to extract deep features using the existing backbone, leading to their extracting the global information of the image first and ignoring the edge parts of the image.

In a network for edge detail processing, Zhou et al. proposed an attention transfer network, effectively reducing the computational complexity and exactly extracting an object with fine structures from a complex background [20]. Fang et al. proposed a user-guided approach for practical human matting, enhancing the accuracy of edge prediction by an interactive matting strategy [21]. Islam MA et al. proposed a stage-wise refinement mechanism to make the edge detail more visible in matting images to a certain extent [4]. Wang et al. proposed an image-matting method based on graph theory and guided feathering, which improves the prediction of the local details of images [22].

Inspired by these, this paper designs a contour sharpness refinement network and a natural image-matting method based on the U-Net deep learning model framework to make the portrait-matting network improve the prediction of local details on the premise of complete global prediction.

2.2. Datasets

As shown in Figure 1, existing portrait-matting datasets are divided into two categories: synthetic-background datasets and natural-background datasets. A more representative synthetic-background dataset is the Adobe Image Matting dataset [23] published by Xu et al., which consists of 431 portrait images with fine annotation and 100 background images without portraits. Lin et al. released the PhotoMatte85 dataset [24] containing 85 portrait foregrounds with their corresponding hand-annotated alpha images and later released 193 background images without portraits for post-composition. The synthetic background can significantly increase the data volume of the dataset and reduce the training cost of deep learning.

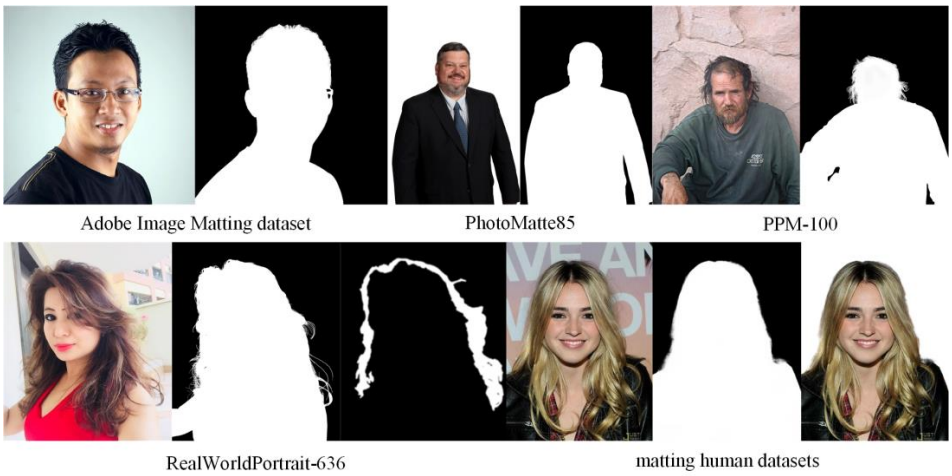

Figure 1. Comparison of the datasets.

J Li et al. showed that models trained in a natural context have a better generalization ability than those trained in a synthetic context [25], so a large number of researchers choose to train the models using natural-context datasets. The PPM-100 datasets [26] published by Ke et al. contain 100 portraits with a sharpness between 1080 p and 4 k, and the ground truth of each image is carefully hand-annotated. The RealWorldPortrait-636 dataset [27] published by Yu et al. contains 636 portraits with fine annotations in the range of 720 p to 1080 p, and additional annotated masks for hair areas and other soft tissue areas. The Matting Human Datasets published by www.aisegment.cn contain 34,427 images and

the corresponding matting results, which are 600 × 800 portraits generated after face detection and region cropping. This is the largest known natural-background portrait dataset. However, this dataset suffers from the problems of rough edge markings of characters and blurred ground truth details, and although the data volume is large, it does not generate a model that achieves the expected effect of clear visibility of hairs in the actual training process. However, the production cost of high-quality natural-background datasets is higher than that of synthetic-background datasets, so natural-background datasets generally have the problem of insufficient training data due to the small amount of data. To enrich the training data, 100 natural-background portraits are collected and each image is manually labeled meticulously and combined with the PPM-100 datasets, and then the images are randomly rotated for data enhancement to finally generate a bust dataset containing 2000 high-quality labeled portraits.

3. Approach

3.1. Overall Network Structure

The entire bust-matting network is shown in Figure 2. To make the global prediction of the portrait-matting network more accurate, this paper adopts the U2-Net network structure with Residual U-blocks (RSU) to improve the capture of multi-scale features. To make the predicted local details more precise, this paper adds the contour sharpness refinement network after the third and fourth decoder layers to realize the refinement of the uncertain regions at the edges of the portrait, and then fuses the output images of each level to generate the final feature maps.

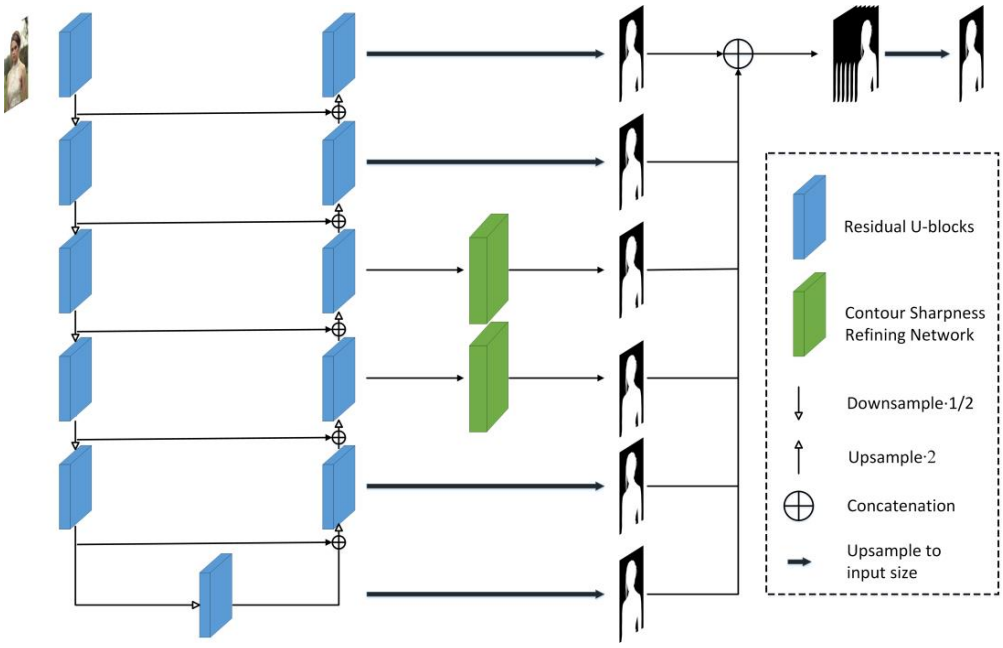

Figure 2. Network structure.

3.2. Contour Sharpness Refining Network

To address the problem of the inaccurate prediction of edges in deep learning matting models, this paper designs a contour sharpness refinement network that processes the uncertain regions that appear in the deep network during upsampling. The main function of the contour sharpness refinement network designed in this paper is to judge the uncertain regions in the input αl-1 whose weights are between (0, 1) and set their weights to 0, to generate a new mask gl. The specific effect is shown in Figure 3.

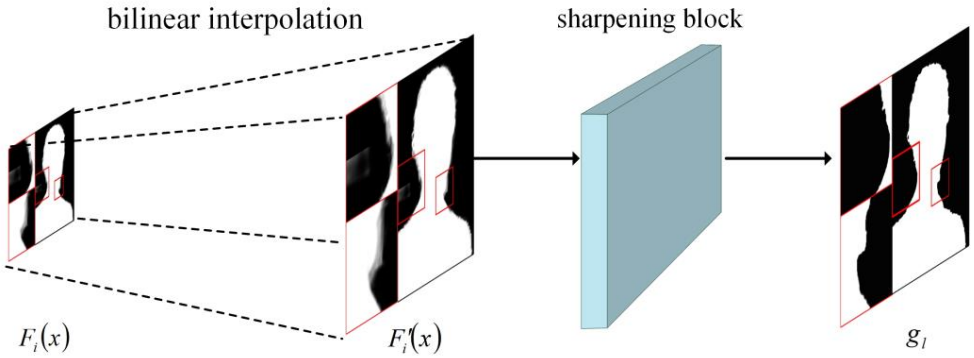

Figure 3. Contour sharpness refinement network.

The contour sharpness refinement network devised in this paper consists of two components as follows:

(1) An upsampling using bilinear interpolation to upsample each layer of the initial feature map progressively by $F_i(x) \in R^{H_i,W_i,C_{in}}$ $(i = 1, 2 \cdots 6)$ into a high-resolution feature map that matches the size of the original image $F_i'(x) \in R^{H,W,C_{in}}$ $(i = 1, 2 \cdots 6)$.

(2) A sharpening block sharpening the generated high-resolution feature maps $F_i'(x)$ by sharpening $g_l = F_i'(x) \cdot f_{gl}$ $(i = 1, 2 \cdots 6)$ to generate a new mask, where the formula for generating a new mask is shown in Equation (2).

$$f_{gl} = \begin{cases} 0 & if\ 0 < \alpha_{l-1} \\ 1 & otherwise \end{cases} \tag{2}$$

As a common upsampling method in image processing, the bilinear interpolation method has the characteristics of a simple calculation and smooth scaling of the image. Since the bilinear interpolation method only considers the effect of the grey values of the four direct neighbors around the sample point to be measured, but not the effect of the rate of change of the grey values between neighbors, if upsampling is performed after sharpening the network, it will inevitably lead to a blurring of the scaled image in terms of edge details. Since the final prediction is a fused image with multiple levels of sampling, adding a contour sharpening network after only one layer of the decoder will result in a poorly sharpened figure edge and will not allow it to be used to its maximum effect. Adding the network after two or more decoder layers may result in over-sharpening and over-complicated calculations, affecting the final fusion result. Therefore, this paper is designed to prevent this problem by adding the network after two decoder layers.

The external layer network can preserve more detailed information in the convolutional neural network. In contrast, the information in the deep layer network is more abstract. Its primary role is to determine the approximate location of the foreground and background, so the detail processing of the edges of the portrait is lacking, which will cause detail loss and generate a lot of noise when image fusion is performed after upsampling dimensional matching [28]. In convolutional neural networks, the decoders of different layers handle different tasks. Specifically, the different layers in U2-Net have the following functions: The shallow convolution layers in the encoder are used to extract low-level features from images, such as edge and texture information. The deep convolution layers in the encoder are used to extract high-level features from images, such as the shape and texture information of objects. The attention module in the encoder adjusts the weights of feature maps to enhance the model's focus on target regions. By combining these different layers, U2-Net can accurately segment input images and has good processing capabilities for complex backgrounds and small targets.

Considering the different effects produced by different layers, the contour sharpness refinement network is placed in the fifth layer decoder and sixth layer decoder, the third layer decoder and fourth layer decoder, and the first layer decoder and second layer decoder,

respectively, for testing in this paper, which is to improve the accuracy of the network model. As shown in Figure 4, the fifth layer decoder and sixth layer decoder need to judge the foreground and background information of the image. When judging the part of the human portrait where the hair is in contact with the body, the hair belongs to continuous semantic information. Therefore, the shoulder and the hair are not continuous with each other after the addition of the contour sharpness refinement network. Adding the contour sharpness refinement network after the fifth layer decoder and sixth layer decoder may lead to errors in judging the foreground and background image, and the complete foreground information cannot be obtained; adding the contour sharpness refinement network after the first layer decoder and second layer decoder will lead to the problem of the low refinement of edges; adding the contour sharpness refinement network after the third layer decoder and fourth layer decoder can obtain the complete foreground information and the shoulder and hair outline of the character are extremely clear. Moreover, the addition of the contour sharpness refinement network after the third layer decoder and fourth layer decoder prevents over- or under-fitting. Thus, this paper proposes to add the contour sharpness refinement network after the third layer decoder and fourth layer decoder of the network to achieve a finer matting of the contour edges of the bust portrait while obtaining a complete foreground.

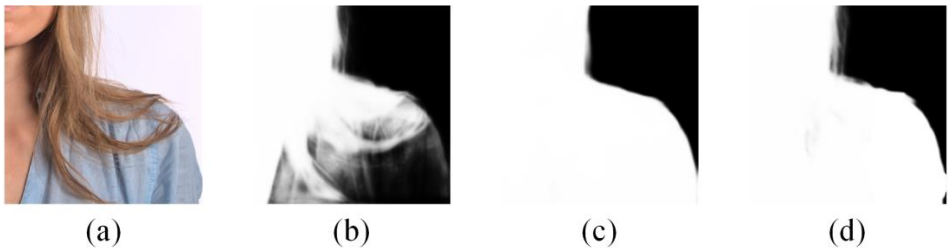

Figure 4. Effect of contour sharpness refinement network placed on different layers: (**a**) original image; (**b**) placed on layers 5 and 6; (**c**) placed on layers 1 and 2; and (**d**) placed on layers 3 and 4.

As shown in Figure 5, to verify the effectiveness of this approach, the contour sharpness refinement network placed on different layers was tested on the PPM-100 dataset and the RealWorldPortrait-636 dataset, and the results found that placing the contour sharpness refinement network on the third and fourth layers still gave the best results on the different datasets.

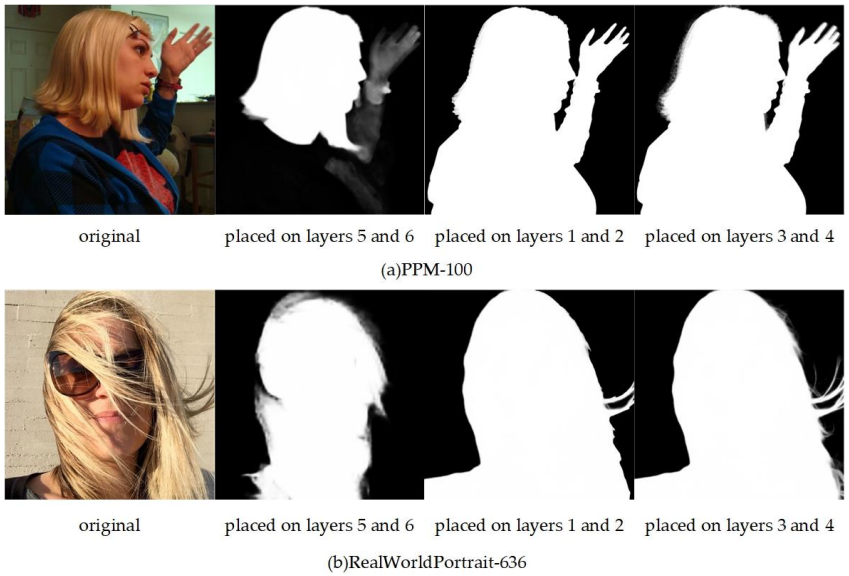

Figure 5. Experimental effects of different layers of contour sharpness refinement networks placed on different datasets.

4. Experiments and Analysis

4.1. Experimental Environment

In this paper, the deep learning framework based on Pytorch is used for training and testing, the interface language is Python 3.6.6, the operating system is Windows 10, the CPU model is Intel Xeon W-2123, and the GPU model is Quadro p4000. The batch size is set to 10 and the epoch is 1000.

4.2. Experimental Results and Analysis

To test the performance of the designed network, the proposed method is trained and tested with U2-Net [14], Glance and Focus Matting Network [25], and MOD Net [26], respectively, on the PPM-100 datasets, RealWorldPortrait-636 dataset, and self-built dataset.

In this paper, the quantitative comparison uses the evaluation metrics of sum of absolute difference (SAD), mean squared error (MSE), gradient error, and connectivity error [29] between the pixels of the predicted alpha map and ground truth proposed by Rhemann et al. The lower value of the four indicators represents a better quality of the generated alpha map. The comparison results in the PPM-100 test set are shown in Table 1, in the RealWorldPortrait-636 test set in Table 2, and in the self-built test set in Table 3.

Table 1. Comparison of the results of several methods on the PPM-100 test set.

Method Name	SAD	MSE(10^{-3})	Gradient	Connectivity
U2-NET [13]	114.62	12.26	103.16	73.87
GF Matting [22]	142.36	6.92	124.35	227.27
MOD Net [23]	125.85	11.25	114.96	87.52
Ours	66.26	12.18	97.70	72.86

Table 2. Comparison of the results of several methods on the RealWorldPortrait-636 test set.

Method Name	SAD	MSE(10^{-3})	Gradient	Connectivity
U2-NET [13]	72.79	11.99	71.08	51.35
GF Matting [22]	96.37	8.94	84.21	60.63
MOD Net [23]	50.34	9.77	58.36	30.19
Ours	45.14	9.49	44.72	26.10

Table 3. Comparison results of several methods on self-built test sets.

Method Name	SAD	MSE(10^{-3})	Gradient	Connectivity
U2-NET [13]	72.78	11.56	28.81	50.18
GF Matting [22]	108.11	7.26	34.51	43.86
MOD Net [23]	85.71	10.27	27.55	38.05
Ours	62.09	10.57	20.46	32.64

As can be seen from Table 1, the bust-matting network with the contour sharpness refinement network proposed in this paper is significantly lower than the comparison methods in the SAD evaluation index, the Gradient and Connectivity evaluation indices are lower than the comparison methods, and the MSE evaluation index has no obvious advantage. As can be seen from Table 2, the method proposed in this paper outperforms the comparison method in the SAD, Gradient, and Connectivity evaluation indices, and is second only to GF Matting in the MSE evaluation indices. As can be seen from Table 3, although the method proposed in this paper outperforms only U2-NET in the MSE evaluation indices, it is significantly lower than the comparison method in the SAD and Gradient indices, and the reason for the lack of an advantage in the MSE metrics may be that the contour sharpness refinement network designed in this paper reduces the pixel value of the foreground uncertainty region after superimposition, thus affecting the MSE evaluation metrics.

To more intuitively verify the effectiveness of the network designed in this paper, the qualitative comparison results of the relevant methods are presented in this paper. The experimental results of the PPM-100 test set are shown in Figure 6, the experimental results of the RealWorldPortrait-636 test set are shown in Figure 7, and the experimental results of the self-built Matting test set are shown in Figure 8.

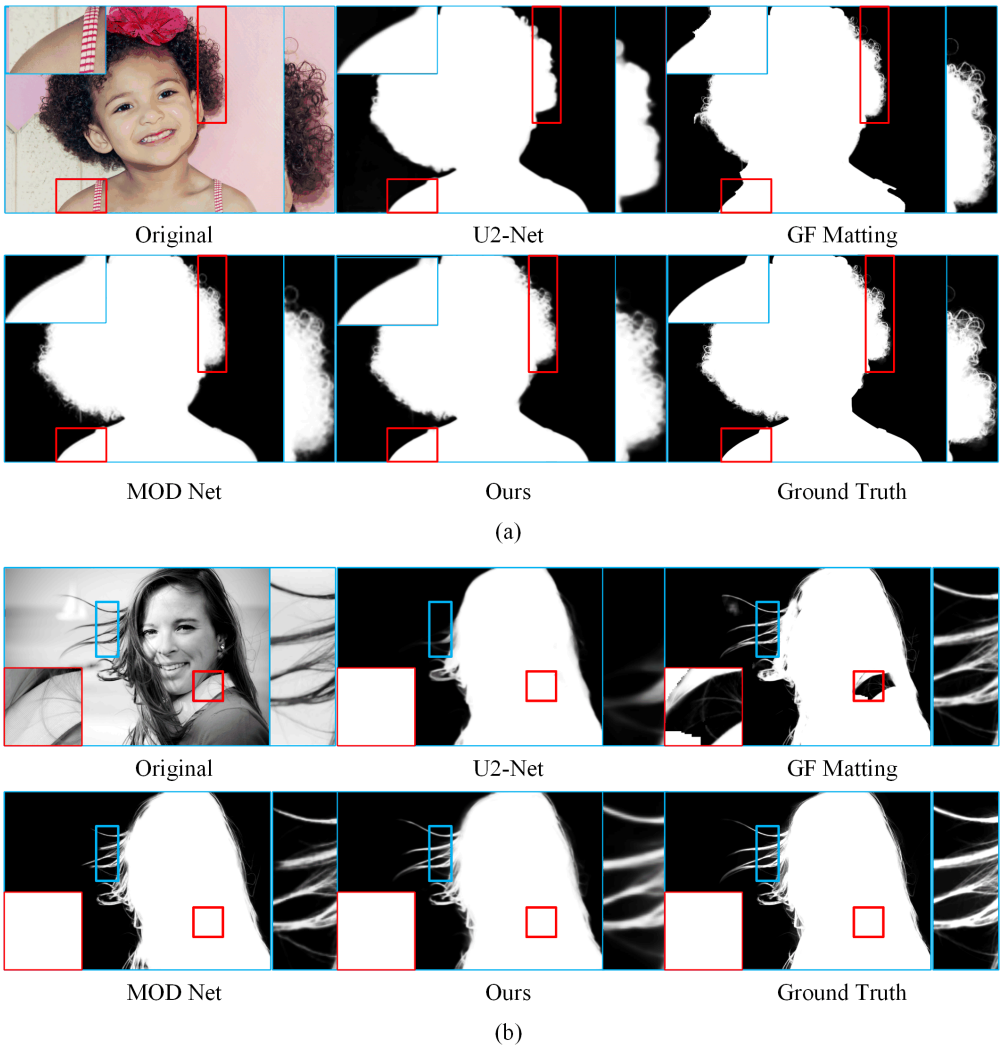

Figure 6. PPM-100 test set experimental results: (**a**) Chart 54 in the dataset; (**b**) Chart 21 in the dataset.

The foreground part of the portrait is accurately predicted by U2-Net, but the prediction of the hair is incomplete; for example, in (a), the edges of the portrait are very blurred, and in (b), the uppermost part of the hair is not predicted. Although GF Matting and MOD Net have some improvement in the detail of the character's hair, there is the problem of an incomplete foreground in global prediction. For example, MOD Net predicts the shadow behind the portrait as the foreground in (a); GF matting gets an incomplete prediction of the shoulder part in (a) and forecasts the collar as the background in (b), resulting in a gap in the final generated image. The method devised in this paper provides a more complete global prediction of the portrait foreground and has some advantages in terms of hair detail.

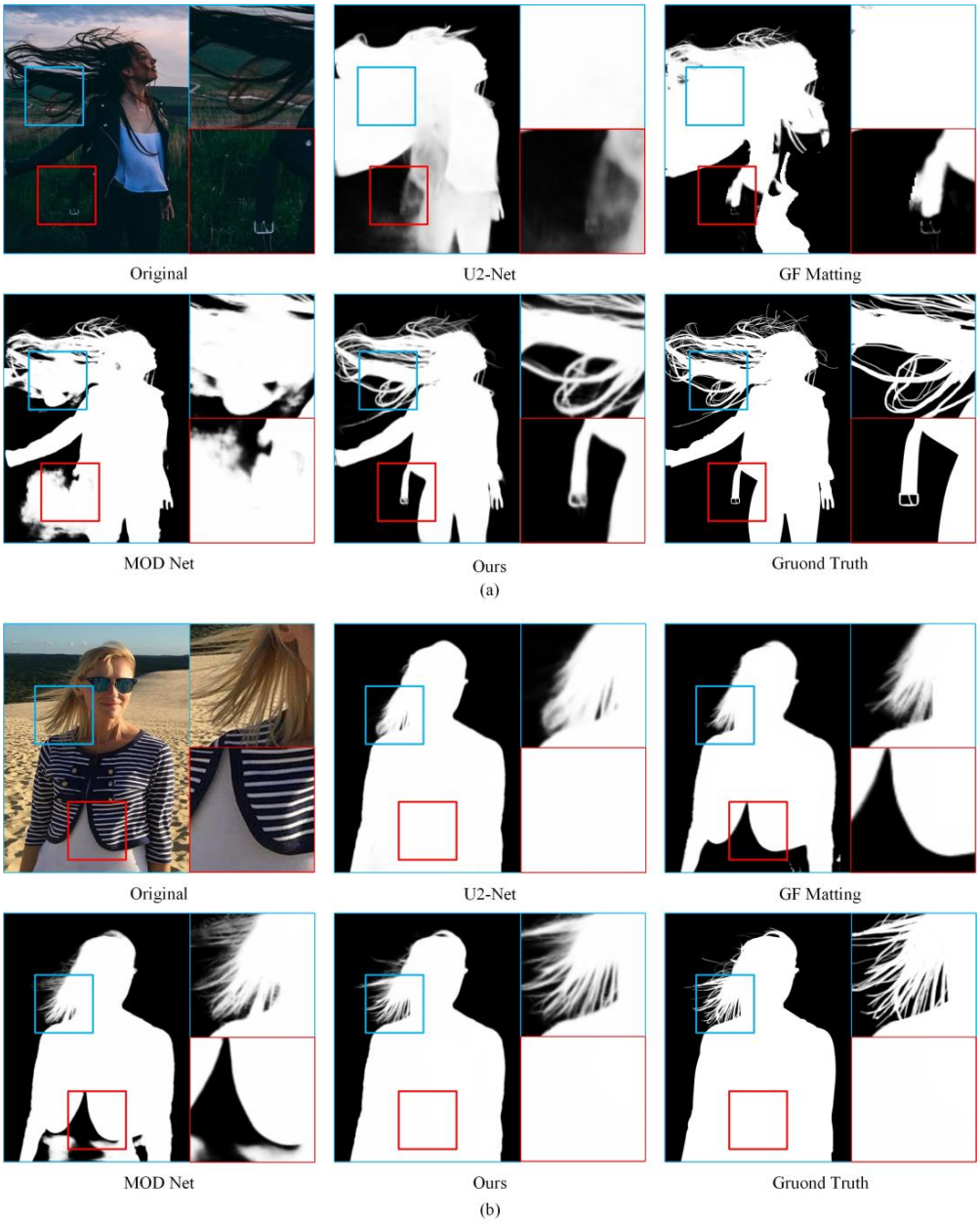

Figure 7. RealWorldPortrait-636 test set experimental results: (**a**) Chart 612 in the dataset; (**b**) Chart 295 in the dataset.

As seen in Figure 7, several of the comparison methods have incomplete predictions of the portrait in (a) and unclear predictions of the hair in the portrait, with all of them predicting the grass in the background as hair. In (b), although the U2-Net prediction of the foreground is more complete, there is a blurring of the edges of the hair. GF Matting and MOD Net have a clearer prediction of the character's hair but an incomplete prediction of the character as a whole. In contrast, the method in this paper produces better prediction results for the detailed part of the edge of the portrait while obtaining a complete foreground image. The method in this paper has specific progress in the detail processing of the part of the edge of the portrait compared to other comparison methods.

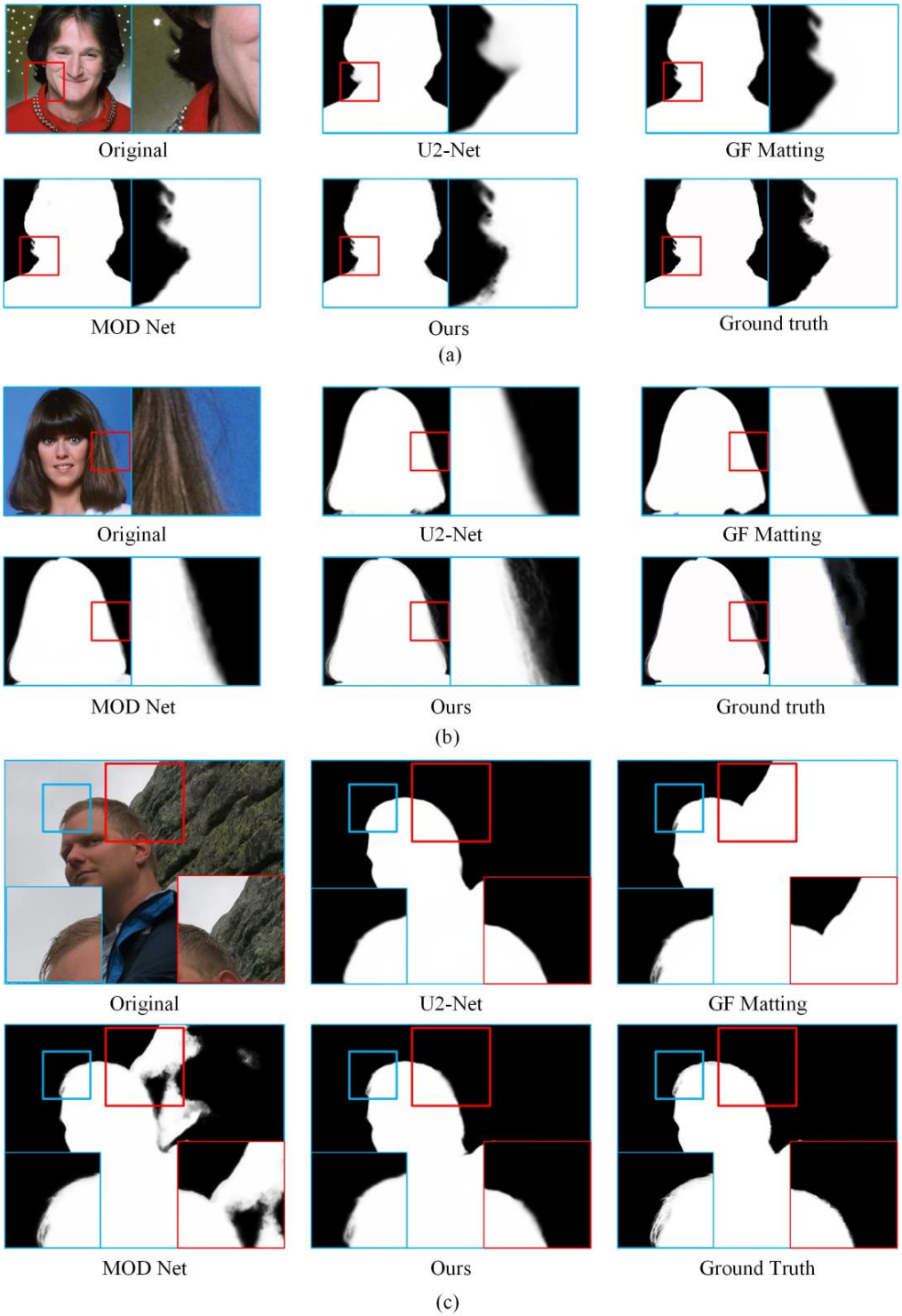

Figure 8. Experimental results of the self-built matting test set: (**a**) Chart 28 in the dataset; (**b**) Chart 67 in the dataset; (**c**) Chart 50 in the dataset.

As can be seen in Figure 8, several of the comparison methods have good results in predicting busts, but there are still cases where the details on the edges of the portrait are missing and the foreground areas of the portrait are misjudged. In (a) and (b), U2-Net, GF Matting, and MOD Net have good results in global prediction, but the local details of the portrait are blurred and the hair is not well-processed, not to the extent that the hair is visible. In (c), GF Matting and MOD Net both judge the mountain behind the portrait as the foreground, thus leading to a serious foreground prediction error.

In summary, the network designed in this paper outperforms the comparison methods in quantitative comparisons for the SAD, Gradient, and Connectivity indices. Meanwhile, in qualitative comparisons, it can have more refined predictions of local details on the premise of complete global predictions for busts, verifying the feasibility of the method designed in this paper.

5. Conclusions

This paper proposes a contour sharpness refinement network for the edge detail processing of busts and designs a matting method based on a combination of U2-Net and contour sharpness refinement network. The method is characterized by accurate global foreground prediction, high-quality local edge detail, and no additional manual annotation of thirds. Based on the high level of edge detail in the bust, the edge detail of the figure is refined by the contour sharpness refinement network. The network used in this paper is not trained with trimaps, but still has better results and relatively more precise edge detail in the portrait. This is mainly due to the fact that the proposed contour sharpness refinement network refines the edges of the bust and removes some of the noise in the foreground. The experimental results show that the method improves the performance on the PPM-100 datasets, the RealWorldPortrait-636 dataset, and the self-built dataset, and that the edge details of the bust can be seen through intuition, which validates the effectiveness of the method in this paper. The sharpening process carried out by this method to obtain images with sharp edges has a large and abrupt change to the boundaries, and therefore has certain drawbacks for the processing of defocused images, for which we will make further modifications to the model in the next step.

Author Contributions: Conceptualization, H.X. and K.H.; methodology, K.H.; software, K.H.; validation, H.X., K.H. and D.J.; formal analysis, W.M.; investigation, K.H.; resources, H.X.; data curation, D.J.; writing—original draft preparation, K.H.; writing—review and editing, H.X.; visualization, K.H.; supervision, D.J.; project administration, W.M.; funding acquisition, H.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are available upon request due to restrictions. The datasets used in the text are all publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ruzon, M.A.; Tomasi, C. Alpha Estimation in Natural Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2000 (Cat. No. PR00662), Hilton Head Island, SC, USA, 13–15 June 2000; IEEE: Piscataway, NJ, USA, 2000; pp. 18–25.
2. Sun, J.; Jia, J.; Tang, C.-K.; Shum, H.-Y. Poisson matting. In Proceedings of the ACM SIGGRAPH 2004 Papers, Los Angeles, CA, USA, 8–12 August 2004; pp. 315–321.
3. Chuang, Y.-Y.; Curless, B.; Salesin, D.H.; Szeliski, R. A Bayesian Approach to Digital Matting. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001; IEEE: Piscataway, NJ, USA, 2001; pp. 264–271.
4. Islam, M.A.; Kalash, M.; Bruce, N.D. Revisiting Salient Object Detection: Simultaneous Detection, Ranking and Subitizing of Multiple Salient Objects. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7142–7150.
5. Porter, T.; Duff, T. Compositing Digital Images. In Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques, Detroit, MI, USA, 25–29 July 1984; pp. 253–259.
6. Ruan, Q.; Wu, Q.; Yao, J.; Wang, Y.; Tseng, H.-W.; Zhang, Z. An Efficient Tongue Segmentation Model Based on U-Net Framework. *Int. J. Pattern Recognit. Artif. Intell.* **2021**, *35*, 2154035. [CrossRef]
7. Qiao, Y.; Liu, Y.; Yang, X.; Zhou, D.; Xu, M.; Zhang, Q.; Wei, X. Attention-Guided Hierarchical Structure Aggregation for Image Matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13676–13685.
8. Yihui, L.; Fujian, F.; Zhaoquan, C. Pyramid Matting: A resource-adaptive multi-scale pixel pair optimization framework for image matting. *IEEE Access* **2020**, *8*, 93487–93498. [CrossRef]

9. Wang, Y.; Niu, Y.; Duan, P.; Lin, J.; Zheng, Y. Deep Propagation Based Image Matting. In Proceedings of the IJCAI, Stockholm, Sweden, 13–19 July 2018; pp. 999–1006.
10. 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, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4130–4139.
11. Liu, B.; Jing, H.; Qu, G.; Guesgen, H.W. Cascaded Segmented Matting Network for Human Matting. *IEEE Access* **2021**, *9*, 157182–157191. [CrossRef]
12. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 11696–11706.
13. Cai, H.; Xue, F.; Xu, L.; Guo, L. TransMatting: Enhancing Transparent Objects Matting with Transformers. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part XXIX. Springer: Berlin/Heidelberg, Germany, 2022; pp. 253–269.
14. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [CrossRef]
15. Xie, E.; Sun, P.; Song, X.; Wang, W.; Liu, X.; Liang, D.; Shen, C.; Luo, P. Polarmask: Single Shot Instance Segmentation with Polar Representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12193–12202.
16. Xiao, J.; Guo, H.; Yao, Y.; Zhang, S.; Zhou, J.; Jiang, Z. Multi-Scale Object Detection with the Pixel Attention Mechanism in a Complex Background. *Remote Sens.* **2022**, *14*, 3969. [CrossRef]
17. Luo, G.; Zhou, Y.; Sun, X.; Cao, L.; Wu, C.; Deng, C.; Ji, R. Multi-Task Collaborative Network for Joint Referring Expression Comprehension and Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10034–10043.
18. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
20. Zhou, F.; Tian, Y.; Qi, Z. Attention transfer network for nature image matting. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 2192–2205. [CrossRef]
21. Fang, X.; Zhang, S.-H.; Chen, T.; Wu, X.; Shamir, A.; Hu, S.-M. User-Guided Deep Human Image Matting Using Arbitrary Trimaps. *IEEE Trans. Image Process.* **2022**, *31*, 2040–2052. [CrossRef] [PubMed]
22. Wang, W.; Tu, A.; Bergholm, F. Improved Minimum Spanning Tree based Image Segmentation with Guided Matting. *KSII Trans. Internet Inf. Syst.* **2022**, *16*, 211–230.
23. Xu, N.; Price, B.; Cohen, S.; Huang, T. Deep Image Matting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2970–2979.
24. Lin, S.; Ryabtsev, A.; Sengupta, S.; Curless, B.L.; Seitz, S.M.; Kemelmacher-Shlizerman, I. Real-Time High-Resolution Background Matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8762–8771.
25. Li, J.; Zhang, J.; Maybank, S.J.; Tao, D. Bridging composite and real: Towards end-to-end deep image matting. *Int. J. Comput. Vis.* **2022**, *130*, 246–266. [CrossRef]
26. Ke, Z.; Sun, J.; Li, K.; Yan, Q.; Lau, R.W. Modnet: Real-Time Trimap-Free Portrait Matting via Objective Decomposition. In Proceedings of the AAAI Conference on Artificial Intelligence, Virutal, 22 February–1 March 2022; pp. 1140–1147.
27. Yu, Q.; Zhang, J.; Zhang, H.; Wang, Y.; Lin, Z.; Xu, N.; Bai, Y.; Yuille, A. Mask guided matting via progressive refinement network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1154–1163.
28. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
29. Rhemann, C.; Rother, C.; Wang, J.; Gelautz, M.; Kohli, P.; Rott, P. A perceptually motivated online benchmark for image matting. In Proceedings of the 2009 IEEE conference on computer vision and pattern recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 1826–1833.