Alpha Matte Generation from Single Input for Portrait Matting

Dogucan Yaman[1] Hazım Kemal Ekenel[2] Alexander Waibel[1,3]
[1]Karlsruhe Institute of Technology, [2]Istanbul Technical University, [3]Carnegie Mellon University
{dogucan.yaman, alexander.waibel}@kit.edu, ekenel@itu.edu.tr

Abstract

In the portrait matting, the goal is to predict an alpha matte that identifies the effect of each pixel on the foreground subject. Traditional approaches and most of the existing works utilized an additional input, e.g., trimap, background image, to predict alpha matte. However, (1) providing additional input is not always practical, and (2) models are too sensitive to these additional inputs. To address these points, in this paper, we introduce an additional input-free approach to perform portrait matting. We divide the task into two subtasks, segmentation and alpha matte prediction. We first generate a coarse segmentation map from the input image and then predict the alpha matte by utilizing the image and segmentation map. Besides, we present a segmentation encoding block to downsample the coarse segmentation map and provide useful feature representation to the residual block, since using a single encoder causes the vanishing of the segmentation information. We tested our model on four different benchmark datasets. The proposed method outperformed the MODNet and MGMatting methods that also take a single input. Besides, we obtained comparable results with BGM-V2 and FBA methods that require additional input.

1. Introduction

Image matting has become a popular research topic in the computer vision research area. The main purpose is to distinguish background and foreground to obtain foreground objects as accurately as possible. Therefore, the task is to generate an alpha matte that contains alpha values, namely opacity values, between $[0, 1]$ for each pixel to represent the effect of the foreground over the final image. In addition to this, portrait matting, which is a subtopic of image matting, focuses on generating alpha matte to obtain the subject itself, instead of the generic objects, from an input image or a video frame. There are numerous application areas of portrait matting, such as image/video editing, changing background which is quite common in video conference applications, and video/movie post-production.

There are various challenges in the portrait matting problem due to the complex visual details of a person's body, e.g., the borders around the body, the hair, and the clothes, particularly if the hair flutters and the clothes have some opacity. The matting problem can be formulated as follows:

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i \tag{1}$$

where i represents each pixel in an image I, alpha represents alpha value for the corresponding pixel in the alpha matte α, and F and B are foreground and the new background images, respectively.

In the traditional approach, the alpha matte is generated using an image and a trimap which represents the foreground, background, and unknown areas on the image. The basic idea is to enhance the unknown areas in the trimap, which are generally problematic parts of the subject, e.g., the area around the subject's body, to get a more accurate alpha matte. The predefined foreground and background areas are not changed. On the contrary, several latest works [24,32,38] propose not to use a trimap, since creating trimap is a time-consuming procedure and needs expert annotators. Instead, some works employ original image and coarse annotated segmentation mask to generate a fine-grained alpha matte [38, 48]. Moreover, recent work focuses on using an input image and its background to produce alpha matte without using any other information [29], while other works utilize only the input image to achieve fine-grained alpha matte [24, 32, 51, 52] and predict trimap to use in the alpha matte prediction [43].

One of the crucial challenges is posed by the distribution of the background and foreground of an image. It is an extremely severe case when the background distribution is considerably similar to the foreground distribution. Besides, if the background distribution has a large variance, it is another compelling case to handle the discrimination of background and foreground subject. Yet another challenge arises from the illumination conditions of the input image since the background matting models are sensitive to the illumination distribution. In particular, the alpha matte prediction models are prone to generate coarse, even worse, outputs under the cases of underexposure and overexposure.

In this work, we aim to enhance the quality of the generated alpha matte to extract the person, since fine-grained details of the subjects are the main challenges in the portrait matting task. To alleviate the problem, we handled it using two consecutive stages, which are person segmentation and alpha matte generation. We employed DeepLabv3+ [4] for person segmentation and a generative adversarial network-based (GAN) alpha matte prediction model. While the first network takes an input image and produces the segmentation map, the alpha generation network employs the output of the segmentation network and foreground subject, which is obtained by multiplication of the input image and predicted segmentation map. In the end, the refinement block receives the predicted alpha matte to refine the details. Our contributions can be summarized as follows:

- We propose a two-stage portrait matting network, that consists of a SOTA person segmentation network DeepLabv3+ and subsequently a conditional GAN-based alpha matte prediction module, without using an additional input as trimap, background image, etc.
- We present *segmentation encoding block* to encode the predicted segmentation map and the foreground subject. The idea is to obtain the feature representation of the segmentation map and foreground subject independently of the input, and inject it into the residual block as well as decoder layers along with the depth. We observed that using an independent encoder, instead of encoding concatenation of all inputs with a single encoder, provides better feature representation.
- We propose *border loss* to penalize the errors around the subject more, since it is more likely to have errors in the prediction due to difficulties, such as hair. We also present *alpha coefficient loss* to evaluate only the pixels that have neither 0 nor 1 value in the alpha matte.

2. Related Work

Although person segmentation can be employed to extract the subject from an image as well as replace the background, it is not adequately accurate to eliminate the background and its effects on the subject. Therefore, alpha matte generation is a more accurate approach for background replacement or portrait matting.

Image matting We can divide image matting literature into three main groups which are sampling-based methods [8, 12, 14, 17, 20–23], propagation-based methods [1–3, 6, 16, 26, 27, 41], and deep learning-based methods [5, 7, 10, 13, 18, 24, 28, 31–36, 38, 39, 42, 46, 47, 49, 50, 53–55]. In deep learning-based methods, Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) are proposed to perform alpha matte prediction for image matting [7, 18, 31, 34, 35, 49, 50, 54]. Besides, the attention mechanism increases the matting performance [28, 36]. Recently, trimap-free approaches become more and more important due to the difficulty of obtaining trimap [13, 48, 53].

Portrait Matting In [39], a CNN-based end-to-end system is presented to produce an alpha matte for the portrait matting task. In [5], the key point of Semantic Human Matting (SHM) algorithm is to learn implicit semantic constraints from the data to use. Moreover, the authors provide a new dataset and a novel fusion strategy for the alpha matte. In [47], end-to-end Joint Matting Network (JMNet) benefits from the pose of the human body to produce alpha matte and uses trimap refiner network to improve the sharpness. In [32], the proposed system contains three submodules that are predicting coarse semantic mask, improving the quality of the mask, and generating the final alpha matte. In [38], a trimap-free system takes an input image and the background of the same image without subject to generate alpha matte. To provide generalization, another matting network is trained in relation to the first network. In [9], a light-weight method with two decoders and a single encoder is proposed. The task-specific decoders predict segmentation map and alpha matte using encoded semantic information. In [24], a matting objective decomposition network (MODNet) is proposed and a self-supervision-based strategy is applied to adjust it to the real-world scenario. In [29], the proposed method has two subnetworks and works in real-time with a high accuracy using HR images. While the first network takes an image and its background as input and generates four different outputs —alpha matte, foreground residual, error map, and hidden—, the second network performs refinement. Besides, two large-scale datasets for video and image matting are presented. In [48], the proposed system works without additional input and the task is addressed as self-supervised multi-modality problem. The system utilizes depth-map, segmentation map, and interaction heat-map using three different encoders. A new dataset is also proposed in this work. In [44], the authors propose Consistency-Regularized Graph Neural Network to improve the temporal coherence during the video matting and they also collected a real-world dataset to evaluate the performance. In [52], Cascade Image Matting Network with Deformable Graph Refinement is presented to predict alpha matte automatically without using additional inputs. They predict the alpha matte from low resolution to high resolution. In [51], the proposed system takes a coarse mask to alleviate the alpha generation task. The system does not require a precise trimap but uses a general rough mask to guide the alpha matte prediction. In [43], the authors propose a deep learning-based video matting system and present a novel spatio-temporal feature aggregation module. They also utilized frame-by-frame trimap annotations and contributed to the literature with a large-scale video matting dataset.

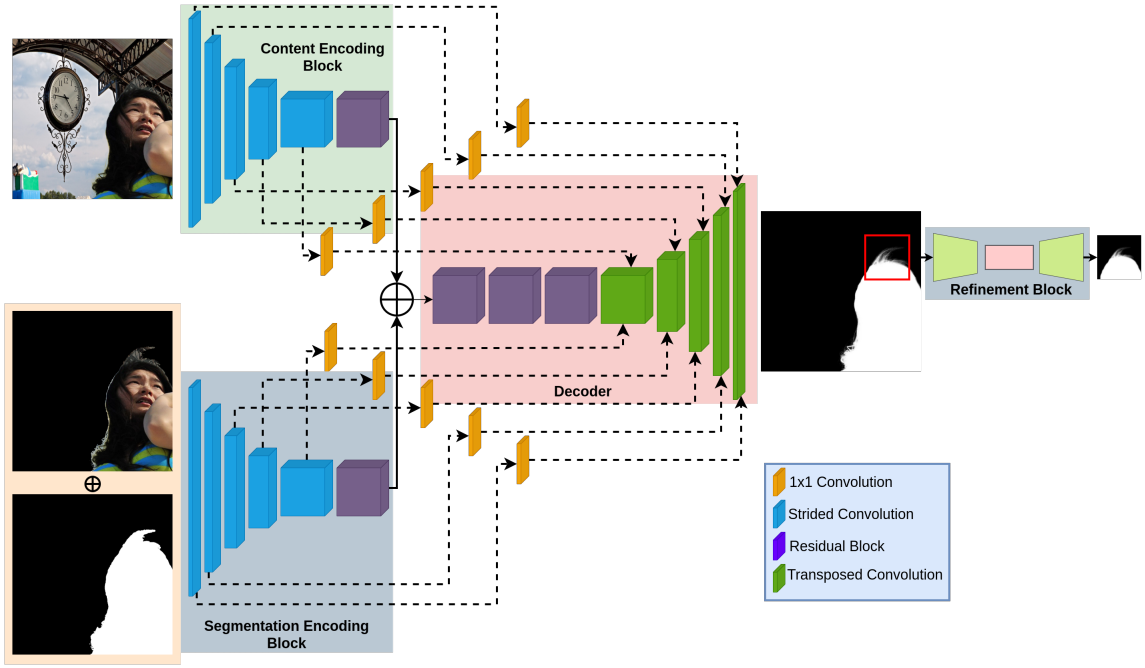

Figure 1. Proposed model. First of all, DeepLabv3+ [4] model works and produces a segmentation map. After that, the visualized system starts to work using the input image and the predicted coarse segmentation map. While the content encoding block encodes the input image to provide feature maps for the decoder network, the segmentation encoding block employs the combination of the predicted segmentation map and foreground subject that is obtained by multiplication of the input image and the predicted segmentation map. Besides, residual connections between encoders' layers and decoder's layers are effective to preserve the information. After each encoders' layer, we passed the extracted feature maps through 1×1 convolutional layers to decrease the depth of the feature maps before concatenating with the decoder's outputs. In the end, the refinement block is responsible for capturing patches from the predicted alpha matte to refine them.

3. Methodology

We propose a two-stage approach to perform portrait matting task. Our model consists of two sub-models which are DeepLabv3+ [4] for person segmentation and alpha matte generation network for alpha matte prediction. While the segmentation model takes a single RGB image to predict the segmentation map, the alpha matte generation network produces alpha matte using the input image as well as the predicted segmentation map. In the alpha generation network, there are two parallel similar encoder blocks, which are the content encoding block and the segmentation encoding block. While the content encoding block provides a feature representation for the input image, the segmentation encoding block encodes the depth-wise concatenation of the predicted segmentation map and foreground subject that is obtained by multiplying the input image and predicted segmentation map. Afterward, the outputs of both encoders are concatenated along with the depth. The concatenated feature representation passes through consecutive residual blocks and the decoder network to obtain the predicted alpha matte. Besides, there are skip connections between the decoder's layers and the encoders' layers. Since the concatenation of three different feature maps makes the feature representation too deep, we pass encoders' outputs through 1×1 convolutions to reduce the dimension before concatenating with the decoder's layers' outputs. In the end, there is a small encoder-decoder network to enhance the predicted alpha matte by taking small patches from the borders of the subject since these regions are more likely to be inaccurate. The proposed system is shown in Figure 1.

Generators While the content encoding block is responsible for encoding the input image to obtain a feature map, the segmentation encoding block provides a feature representation of the predicted segmentation map and foreground subject. The idea behind using separate encoders is to avoid vanishing the features of the segmentation map and the foreground subject. We also found that the segmentation map and extracted foreground subject provide complementary feature representations to the network. According to the experiments, we noticed that using a single encoder causes the vanishing of the feature representation of the additional inputs. Besides, we empirically realize that a less complex encoder is suitable to encode these additional inputs. Our generator is based on the U-Net generator [37] and it contains consecutive convolutional blocks to downsample the

input image. After the encoding blocks, the features are concatenated and the final representation passes through the residual block. Later, the generator has a decoder module to produce an output alpha map by upsampling the residual output. Finally, 64×64 size of consecutive patches in the border of a person's body are extracted from the predicted alpha matte to enhance the details by the refinement network, since the predictions tend to have more errors around the body. The generated image is expected to be the fine-grained alpha matte of the input image for the portrait matting problem. Besides, the skip connections encourage the network to keep information from both encoders.

Multi-scale discriminator network For the multi-scale discriminator [45], we provide an image pyramid using the original image and downsampled versions by a factor of two and a factor of four to obtain the same image on different scales. Therefore, this approach provides us to learn from a general perspective to finer details, since each discriminator has a different receptive field. Please note that all three discriminators are identical, though each discriminator works on a different scale. Since alpha matte does not contain a sufficient amount of useful representation, we decided to use a combination of the alpha matte and the extracted foreground subject, which is obtained by multiplying the alpha matte and the image, as an input to the discriminator network. For the real image, we extracted subjects using the images and the ground truth alpha matte, while we used the images and the predicted alpha matte to obtain fake images for the discriminator. Depth-wise concatenation of the three channels RGB image and one channel alpha matte is the input data of the discriminator.

Loss functions For the training of the alpha generation network, we used adversarial loss [15], perceptual loss [19], alpha loss, border loss, and alpha coefficient loss. In the perceptual loss [19], we utilized the VGG model [40] to extract features. For this, we employed five different layers of the VGG model to obtain features for the generated image and the real image. We followed a similar pipeline as in [19] to decide the layers to extract features. After that, we calculated a weighted sum of the L1 distances between features of the predicted alpha matte and the ground truth alpha matte for all extracted features. Besides, we applied the same loss for generated foreground subject and ground truth foreground subject that we obtained by multiplying the input image with predicted alpha matte and ground truth alpha matte, respectively. Then, we followed the same strategy to extract features and calculate the perceptual loss.

For the alpha loss, we followed a different strategy and calculated the L1 distance between the pixels that have only one or zero values in the pixel domain instead of calculating L1 distance between all pixels. The remaining pixels that have neither one nor zero values are considered by defining another loss based on L1 distance. Thus, we penalized the [0,1] pixels and the pixels between 0 and 1 separately, since they represent different cases and restrain each other when we consider them together. Please note that we also calculated both losses using alpha matte and foreground subjects as in the perceptual loss. Moreover, we proposed the border loss to penalize the area around the subject. For this, we generated border maps by applying morphological erosion and dilation operations separately. Then, we subtracted the eroded segmentation map from the dilated one. The final map represents the border area of the subject. During the training, we utilized this border map to calculate L1 loss for only the corresponding border pixels. The overall loss function is shown in Equation 2

$$\min_{G} \max_{D_1,D_2,D_3} \sum_{k=1,2,3} (L_{cGAN}(G, D_k) + \lambda L_{per}(G) + \beta L_{alpha}(G)) + \gamma L_{border}(G) + \theta L_{ac}(G)) \tag{2}$$

where L_{cGAN} represents conditional adversarial loss, L_{per} shows the perceptual loss, L_{alpha} indicates the alpha loss, L_{border} states the border loss, and L_{ac} expresses the alpha coefficient loss. Besides, $\lambda, \beta, \gamma, \theta$ are coefficients that determine the effect of each losses over the total loss. According to our experiments on validation set, we empirically defined these values as 10, 25, 50, 25.

3.1. Training procedure

During the training, we did not train the segmentation network. Instead, we only trained the alpha generation network and the refinement network end-to-end. During the inference, the framework works end-to-end which means we provide an input image to the whole system and get an alpha matte for the corresponding input image. The input images are resized to 1280×768 resolution before feeding the network. We used 10^{-4} learning rate for the generator and a ten times smaller learning rate (10^{-5}) for the discriminator to slow down the convergence of the discriminator since we empirically realized that discriminator converged too fast. We trained the alpha generation network with batch size of one as using one image in each batch causes better convergence [13]. Besides, we utilized Adam optimizer [25] for the training of both models. We trained the discriminator one step for every five steps for the generator training.

4. Experimental Results

Datasets In order to train our model, we used the combination of Adobe Image Matting (AIM) [49] and Distinctions (D646) [36] datasets to employ more data as well as increase the diversity. Since we focused on the portrait matting problem, we selected all images that contain persons for the training and test by following the same strategy in the portrait matting literature. In the end, there are 201 subjects in the AIM dataset and 363 subjects in the D646

Method	Input	Dataset	MSE	MAE	SAD	Grad	Conn
BGM-V2 [29]	Image, background	AIM	2.12	8.62	9.04	8.32	9.21
FBA [13]	Image, trimap	**AIM**	**0.40**	**3.79**	**3.98**	**1.19**	**3.11**
MODNet [24]	Image	AIM	21.65	32.36	33.93	44.24	35.45
MGM [51]	Image	AIM	1.48	5.96	6.21	4.74	6.55
Ours	Image	**AIM**	**1.06**	**4.93**	**5.04**	**4.22**	**5.39**
BMG-V2 [29]	Image, background	**PM85**	**0.37**	**1.38**	**1.45**	**1.28**	**2.38**
FBA [13]	Image, trimap	PM85	1.01	2.43	2.55	3.50	2.75
MODNet [24]	Image	PM85	2.32	6.90	7.23	12.17	9.48
MGM [51]	Image	PM85	0.38	2.77	2.91	1.32	2.04
Ours	Image	**PM85**	**0.19**	**1.11**	**1.19**	**0.65**	**1.16**
BMG-V2 [29]	Image, background	D646	0.98	4.60	4.83	3.78	5.30
FBA [13]	Image, trimap	**D646**	**0.44**	**3.10**	**3.25**	**1.70**	**2.38**
MODNet [24]	Image	D646	3.51	9.80	10.27	13.54	18.98
MGM [51]	Image	D646	0.88	5.17	5.42	3.40	4.76
Ours	Image	**D646**	**0.71**	**3.84**	**3.99**	**2.74**	**3.84**
FBA [13]	Image, trimap	PPM-100	0.96	2.24	2.41	4.20	2.70
MODNet [24]	Image	PPM-100	4.60	9.70	11.59	12.48	22.16
MGM [51]	Image	PPM-100	1.15	5.07	5.31	5.04	5.29
Ours	Image	**PPM-100**	**0.84**	**4.02**	**4.70**	**3.67**	**4.46**

Table 1. Quantitive evaluation on different datasets. Since PPM-100 dataset contains real-world images, we could not test BGM-V2 due to lack of background images. The corresponding MSE and MAE metrics are scaled by 10^3 to improve the readability.

dataset, making in total 564 subjects for the training set. We created the training set by following the standard strategy in the image matting literature for these datasets. For this, we combined each person in the training set with 100 different images of the MSCOCO dataset [30]. In the end, we have 56400 training images. For the test, we have four different test sets, namely, AIM [49], PhotoMatte85 (PM85) [29], D646 [36], and PPM-100 [24]. We followed the same strategy and combined each person in the test set with 20 different background images of the PASCAL VOC dataset [11]. In the end, AIM contains 220 images (11 different subjects), PM85 includes 1700 images (85 different subjects), and D646 has 220 images (11 different subjects). The images in PPM-100 dataset have real backgrounds and there are 100 images in total. We evaluated our model on these four benchmark datasets and compared our results with the previous works. Please note that the training and test sets do not contain any common subjects, i.e. subject independent setup. The training and test subjects have already been listed for the corresponding datasets.

Evaluation We used mean squared error (MSE), mean absolute error (MAE), sum of absolute difference (SAM), gradient (Grad), and connectivity (Conn) metrics to evaluate our model as in the literature. For comparison, we chose publicly available SOTA methods, namely MODNet [24], BGM-V2 [29], FBA [13], MGM [51] and we tested them on the test sets in order to perform a fair comparison since different backgrounds may change the models' performances. Please note that we calculated these metrics over the whole image, and MSE and MAE scores are scaled by 10^3 to improve the readability. Besides, we performed a user study to compare our results with the other studies. To perform this study, we combined the extracted subjects with a green background and showed these images to the participants.

4.1. Results

In this section, we present the experimental results and compare them with the recent SOTA works in the background matting literature, MODNet [24], FBA [13], BGM-V2 [29], and MGM [51]. Please note that, while our method and MODNet take an input image to generate alpha matte for portrait matting, BGM-V2 requires the original background image without subject as an additional input and FBA expects trimap to identify background, foreground, and unknown areas in addition to the original input image. Besides, MGM [51] requires a segmentation mask as our alpha matte generation network.

Quantitative evaluation Experimental results are shown in Table 1. We evaluated all models under the same conditions, e.g., using background image and resolution. According to the experimental results presented in Table 1, our model surpassed the performance of MODNet and MGM, which do not use any additional inputs, on four different benchmarks. Other methods in the table —BGM-V2 and FBA— benefit from additional input such as the background of the input image and a trimap. These additional

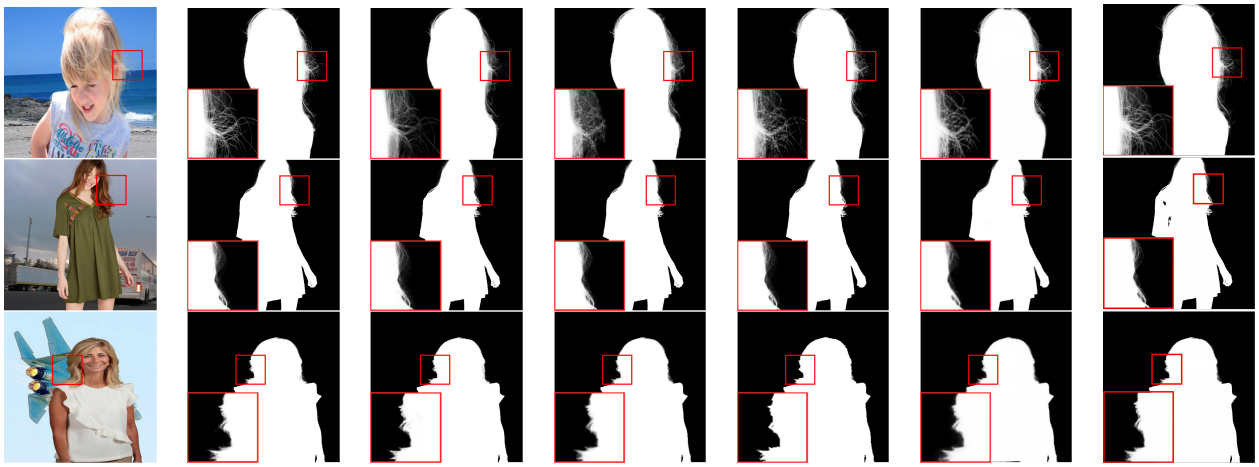

Figure 2. Qualitative comparison. Rows represent AIM, D646, PM85 datasets, respectively.

inputs make the task easier and more accurate results are likely to be obtained, since the background image is the same one as the original input image, and trimap identifies most of the area on the image as foreground and background. On the AIM dataset, our model is found superior to the BGM-V2 in all metrics. However, the FBA method achieves the best performance on this test set. On the PM85 dataset, our proposed model outperforms all methods and gets the SOTA result. In the D646 benchmark, we again outperform the MODNet, MGM, and BGM-V2. The FBA reaches the best performance. However, it is slightly better than our method and our results are quite acceptable when compare with the FBA. Please note that since each study creates the test setup with a different set of background images, the presented scores may show differences.

As previously stated, while our approach does not take any input in addition to the original image, the FBA method takes trimap and the BGM-V2 method takes the background of the original input image that does not contain the subject itself. However, they are too sensitive to these additional inputs. For instance, if there are any dissimilarities in the background image such as translation, BGM-V2 cannot produce a proper output and generates a completely corrupted prediction instead. Similarly, FBA is sensitive to the trimap input. In addition to all these cases, our model and all other models are sensitive to the background of the input image according to the findings of our detailed experiments. It indicates that the alpha matte prediction performance of the models for the same subject can considerably change according to the background of the input image. The illumination conditions, the color distribution, and the existence of multiple subjects on the image affect the alpha matte prediction performance. For the PPM-100 dataset, since the images are real-world images, there are no background images without the subject. Therefore, we could not test the BGM-V2 model on this dataset.

Qualitative evaluation In Figure 2, we present our results, input image, ground truth, and the outputs of the other models for three benchmark datasets; AIM, D646, and PM85. We generated outputs with our model, MGM, and MODNet without additional inputs. However, BGM-V2 method needs the same background of the input image and FBA requires trimap for the corresponding input data. For BGM-V2, we provided the background image that we used during the preparation of the test data. Since D646, PM85, and PPM-100 datasets do not include trimaps, we created different trimaps by using erosion and dilation operations to evaluate FBA and present the best scores. According to the figure, our results are almost the same as the ground truth data, especially for the challenging part, such as hair. Besides, although all models perform quite well, the differences between them are in the details, particularly around the borders of the subjects. Moreover, we randomly collected images from the web and we run our model over them to present the performance of the system on the real-world images. The corresponding outputs are presented in Figure 3. The Alpha column contains the predicted alpha matte and the combined column includes the combination of an arbitrary background image and the extracted subject by using the predicted alpha matte.

We also performed a user study and asked 30 different participants to compare all results according to the quality of the images to measure the matting performance. We used randomly selected sample images from all four benchmark test sets. We present the results in Table 2. We have five different levels of score which are *much better, better, same,*

Score	MODNet	BGM-V2	FBA	MGM
Much better	41.55%	10.45%	4.23%	8.25%
Better	29.22%	32.67%	16.61%	30.27%
Same	19.15%	39.86%	52.44%	41.02%
Worse	8.11%	16.33%	22.47%	18.20%
Much worse	1.94%	0.69%	3.58 %	2.26%

Table 2. User study using all three benchmarks. We compared our model with MODNet [24], BGM-V2 [29], FBA [13], and MGM [51]. The scores demonstrate how much our result is better or worse than the other results.

worse, and *much worse* to compare our results with four different methods. The scores indicate how much the output image of our method is better or worse than the output of other methods. For the comparison, we extracted subjects from the image using predicted alpha matte and combined with a green background to make the details of the subject more visible for the users. During the survey, we showed the original input image and the combination of a green background and outputs of the models. We utilized 8 subjects for each test benchmark, except PPM100 since we could not test the BGM-V2 model on them, for the user study and we made pairs with our results and other results to show them to the participants. In total, we have 24 images for each model to create questions. According to the table, our model overperforms the MODNet and it is slightly better than BGM-V2 and MGM. On the other hand, participants could not easily distinguish our results and FBA results and majority, 52.44%, said they are the same.

4.2. Ablation study

Loss functions We performed an ablation study to evaluate the effects of different parameters on the performance. We first investigated the loss functions and then utilized data type for the losses. In the first part of Table 3, we show used loss functions for the training as well as corresponding MSE values on the AIM test set. It is observed that each employed additional loss contributes significantly to the prediction performance of the model. In the bottom part of Table 3, we present the effect of using the alpha matte and the foreground subject in the loss functions. α means that we only utilized predicted alpha matte and ground truth alpha matte. α and F represent that we extracted the subject from the input image with predicted alpha matte and ground truth alpha matte to obtain predicted and real foreground subjects. Then, we employed these outputs to calculate loss functions for the corresponding case. While using alpha matte helps to penalize the difference between predicted and ground truth alpha matte, using foreground subject provides more information to the network, since it contains much more details and semantic information than the alpha matte. According to the results, MSE scores indicate that using the foreground subject in addition to the alpha matte enables the network to produce a more accurate map.

Loss	MSE
$L_{cGAN} + L_{alpha}$	7.24
$L_{cGAN} + L_{per} + L_{alpha}$	3.78
$L_{cGAN} + L_{per} + L_{alpha} + L_{border}$	1.76
$L_{cGAN} + L_{per} + L_{alpha} + L_{border} + L_{ac}$	1.06
α	3.14
α, F	1.06

Table 3. Ablation study for the loss functions. We repeated the training of the alpha matte generation network using a combination of different loss functions and we present MSE results on AIM test set in the top part of the table. We additionally show the results with all loss functions by using only alpha matte and using foreground subject and alpha matte together in the loss functions.

Cases	MSE
Base model	2.20
Base model + SE block	1.57
Base model + SE block + refinement network	1.06

Table 4. Ablation study for the architecture. We individually investigated the effect of the segmentation encoding block and the refinement module. The experiments are performed on the AIM dataset.

Modules We further examined the effect of the segmentation encoder block and the refinement network. According to the results in Table 4, both the segmentation encoding block and the refinement network are significantly useful to improve the performance of the proposed method. Because, the segmentation encoding block improves the representation of the segmentation area by providing the encoded feature representation to the residual block, while the refinement network enhances the alpha matte prediction performance by focusing on the challenging parts.

Input type We analyzed how using foreground subject in the generator and discriminator as input affects the performance. The results in Table 5 indicate that providing a foreground subject in addition to the segmentation map, which we obtained by multiplying the input and the predicted segmentation map, increases the performance since it provides a more effective feature representation. Similarly, concatenation of the alpha matte and extracted foreground subject provides a more useful representation to the discriminator that yields improvement in the performance. Please note that we evaluated the proposed system on the test set of the AIM dataset.

Limitations Our work is sensitive to the performance

Figure 3. We tested our models on the real images that were collected from the web. In the end, we changed the backgrounds with arbitrary backgrounds using the predicted alpha matte to show the application of the system.

Cases	MSE
Segmentation map	1.86
Segmentation + Foreground	1.41
Alpha matte + Foreground	1.06

Table 5. Ablation study for the input type of the generator and discriminator. While the first part shows the input of the segmentation encoding block in the generator, the second part of the table indicates the input type of the discriminator network.

of the segmentation network. A poor quality segmentation output causes a less accurate outcome at the end of the alpha matte network due to a lack of visual representation of the subject. Besides, due to consecutive residual blocks, the model is not able to run in real-time.

5. Conclusion

In this work, we proposed a conditional GAN-based additional input-free approach to perform the portrait matting task. We addressed the problem as two different subproblems. In the first step, we proposed to use DeepLabV3+ person segmentation model to generate a coarse segmentation map from an arbitrary input image. In the second step, this output and the original image are sent to the alpha generation network to generate the alpha matte. We presented the segmentation encoding block that encodes the combination of the predicted segmentation map and the foreground object. In the end, we have a refinement network to enhance the prediction quality by capturing several patches from the predicted alpha matte in the border area of the subject. Besides, we proposed border loss to penalize challenging parts around the subject and we also presented alpha coefficient loss to measure only the pixels in the alpha matte that the alpha coefficients are neither zero nor one. To handle the domain shift problem, we combined two important training datasets to increase the amount of data as well as the diversity. Experimental results indicate that using border loss and alpha coefficient loss improved the accuracy of the model and combining two datasets increased the generalization capacity. It is also observed that encoding the combination of the segmentation map and the foreground subject by the segmentation encoding block provided more useful features than encoding only the segmentation map. We also found out that the same outcome is also correct for the discriminator. When we provided the prediction output and the foreground subject, the discriminator worked better and was more stable. In future work, it is necessary to focus on the performance to make our model be able to run in real-time with sequential data in order to increase the possibility of usage in the real world. A possible scenario is to utilize the system by eliminating the background to provide privacy in the human-robot interaction.

Acknowledgement. The project on which this report is based was funded by the Federal Ministry of Education and Research (BMBF) of Germany under the number 01IS18040A.

References

[1] 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. 2

[2] Yagiz Aksoy, Tunc Ozan Aydin, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In *Proceedings of the IEEE(CVF Conference on Computer Vision and Pattern Recognition*, pages 29–37, 2017. 2

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

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018. 2, 3

[5] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 618–626, 2018. 2

[6] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2175–2188, 2013. 2

[7] Donghyeon Cho, Yu-Wing Tai, and Inso Kweon. Natural image matting using deep convolutional neural networks. In *European Conference on Computer Vision*, pages 626–643. Springer, 2016. 2

[8] Yung-Yu Chuang, Brian Curless, David H Salesin, and Richard Szeliski. A bayesian approach to digital matting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001. 2

[9] Yutong Dai, Hao Lu, and Chunhua Shen. Towards lightweight portrait matting via parameter sharing. In *Computer Graphics Forum*. Wiley Online Library, 2020. 2

[10] Yutong Dai, Hao Lu, and Chunhua Shen. Learning affinity-aware upsampling for deep image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6841–6850, 2021. 2

[11] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 5

[12] Xiaoxue Feng, Xiaohui Liang, and Zili Zhang. A cluster sampling method for image matting via sparse coding. In *European Conference on Computer Vision*, pages 204–219. Springer, 2016. 2

[13] Marco Forte and François Pitié. f, b, alpha matting. *arXiv preprint arXiv:2003.07711*, 2020. 2, 4, 5, 7

[14] Eduardo SL Gastal and Manuel M Oliveira. Shared sampling for real-time alpha matting. In *Computer Graphics Forum*, volume 29, pages 575–584. Wiley Online Library, 2010. 2

[15] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 4

[16] Leo Grady, Thomas Schiwietz, Shmuel Aharon, and Rüdiger Westermann. Random walks for interactive alpha-matting. In *Proceedings of VIIP*, volume 2005, pages 423–429, 2005. 2

[17] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In *International Conference on Computer Vision and Pattern Recognition*, pages 2049–2056. IEEE, 2011. 2

[18] Hossein Javidnia and François Pitié. Background matting. *arXiv preprint arXiv:2002.04433*, 2020. 2

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 4

[20] 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. 2

[21] 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. 2

[22] Levent Karacan, Aykut Erdem, and Erkut Erdem. Alpha matting with kl-divergence-based sparse sampling. *IEEE Transactions on Image Processing*, 26(9):4523–4536, 2017. 2

[23] Zhanghan Ke, Kaican Li Di Qiu, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *European Conference on Computer Vision*, volume 2, page 6. Springer, 2020. 2

[24] Zhanghan Ke, Kaican Li, Yurou Zhou, Qiuhua Wu, Xiangyu Mao, Qiong Yan, and Rynson WH Lau. Is a green screen really necessary for real-time human matting? *arXiv preprint arXiv:2011.11961*, 2020. 1, 2, 5, 7

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[26] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):228–242, 2007. 2

[27] Anat Levin, Alex Rav-Acha, and Dani Lischinski. Spectral matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1699–1712, 2008. 2

[28] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11450–11457, 2020. 2

[29] 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. 1, 2, 5, 7

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5

[31] Chang Liu, Henghui Ding, and Xudong Jiang. Towards enhancing fine-grained details for image matting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 385–393, 2021. 2

[32] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-Sheng Hua. Boosting semantic human matting with coarse annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8572, 2020. 1, 2

[33] Yuhao Liu, Jiake Xie, Xiao Shi, Yu Qiao, Yujie Huang, Yong Tang, and Xin Yang. Tripartite information mining and integration for image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7555–7564, 2021. 2

[34] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3266–3275, 2019. 2

[35] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. Alphagan: Generative adversarial networks for natural image matting. *arXiv preprint arXiv:1807.10088*, 2018. 2

[36] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020. 2, 4, 5

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015. 3

[38] 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. 1, 2

[39] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *European conference on computer vision*, pages 92–107. Springer, 2016. 2

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[41] Jian Sun, Jiaya Jia, Chi-Keung Tang, and Heung-Yeung Shum. Poisson matting. In *ACM SIGGRAPH 2004*, pages 315–321. 2004. 2

[42] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Semantic image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11120–11129, 2021. 2

[43] 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, 2

[44] Tiantian Wang, Sifei Liu, Yapeng Tian, Kai Li, and Ming-Hsuan Yang. Video matting via consistency-regularized graph neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4902–4911, 2021. 2

[45] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 4

[46] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Hanqing Zhao, Weiming Zhang, and Nenghai Yu. Improved image matting via real-time user clicks and uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15374–15383, 2021. 2

[47] Xian Wu, Xiao-Nan Fang, Tao Chen, and Fang-Lue Zhang. Jmnet: A joint matting network for automatic human matting. *Computational Visual Media*, 6(2):215–224, 2020. 2

[48] Bo Xu, Han Huang, Cheng Lu, Ziwen Li, and Yandong Guo. Virtual multi-modality self-supervised foreground matting for human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 438–447, 2021. 1, 2

[49] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2017. 2, 4, 5

[50] Haichao Yu, Ning Xu, Zilong Huang, Yuqian Zhou, and Humphrey Shi. High-resolution deep image matting. *arXiv preprint arXiv:2009.06613*, 2020. 2

[51] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1154–1163, 2021. 1, 2, 5, 7

[52] Zijian Yu, Xuhui Li, Huijuan Huang, Wen Zheng, and Li Chen. Cascade image matting with deformable graph refinement. *arXiv preprint arXiv:2105.02646*, 2021. 1, 2

[53] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7469–7478, 2019. 2

[54] Yijie Zhong, Bo Li, Lv Tang, Hao Tang, and Shouhong Ding. Highly efficient natural image matting. *arXiv preprint arXiv:2110.12748*, 2021. 2

[55] 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. 2