Deep Residual Network for High-Resolution Background Matting

Said EL ABDELLAOUI[1,2]*, Ilham KACHBAL[1]

[1] LAPSSII, High School of Technology, Cadi Ayyad University, B.P. 89, Safi, Morocco
Said.elabdellaoui@uca.ac.ma (**Corresponding author*)

[2] LRIT Laboratory Associate Unit to CNRST (URAC 29), Faculty of Sciences,
Mohammed V-Agdal University, Rabat, Morocco
Ilhamkachbal@gmail.com

Abstract: Image matting is one of the most important tasks in the computer vision community whose popularity has increased in recent years. This is a highly critical method in video and image editing applications, which involves the separation of the foreground from the background of an image. The previous methods provide a low accuracy when the background and foreground of an image are similar. This paper proposes an effective matting method that integrates the combination of a supervised deep learning matting network generator and a self-supervised refinement network. The generator uses a supervised encoder-decoder network for the extraction of the foreground and alpha matte from the original input image. The results obtained by this network are employed by the self-supervised refinement network to evaluate the newly created composite images and ultimately improving the matting process. The proposed method has obtained better results in comparison with other methods, which makes it more reliable.

Keywords: Matting, Alpha matte, KNN, Soft-Segmentation, GAN.

1. Introduction

Matting, as a way of exploiting images, is gaining importance especially with the help of the most powerful cameras. One of the tasks that either professionals or consumers perform mostly when exploiting images, is object extraction from an image or object composition onto another background. It becomes frequently used in various domains such as promotion image composition, film production, etc. Background matting has taken an important step towards a new style of marketing via the composition of a product image on a human body. Image matting aims to predict the alpha matte at each pixel of the foreground object. Mathematically, an input image Ii is modelled as a linear combination of the foreground and the background color by the equation:

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

where Fi, Bi, and αi indicate the foreground, the background color, and the desired alpha matte at pixel (i), respectively. The problem of image matting is ill-posed because it tries to determine seven values (α_i, F_i and B_i) based on only three known values (I_i) for RGB image, as it can be seen in equation (1). For many of the prevailing approaches (Grady et al., 2005; Levin et al., 2007), the task of image matting always requires a trimap that denotes the known foreground, known background, and unknown region that contains the combination of background and foreground colors. The limitation of trimap lies in the fact that it does not work as expected if there is more than one object to extract in an image. It chooses one of the objects to extract and considers the others as background.

The task of image matting faces several challenges. Firstly, in natural images, foreground and background colors are often similar which makes traditional color-based techniques very weak. Secondly, the majority of natural images contain microscopic details like cells, fingers, or hair which create a huge challenge for accurate extraction. It is easy to assume that the first challenge is more associated with the contextual features, while the other is more associated with textural quality and contains more spatial details. As a solution to those challenges, in recent years, researchers have proposed deep learning algorithms, which in addition to the image colors, use contextual information of the picture as well. This technique solves the first problem almost perfectly. However, for the second one still needs a powerful model that can detect microscopic structure in images. This paper proposes an effective matting method that merges two different algorithms. The first one uses a deep learning matting network generator estimation for the extraction of the foreground and alpha matte from the original input image, while leaving the refinement work to the second

https://doi.org/10.24846/v30i3y202105

one which is a self-supervised refinement network that uses the output of the first model and tries to make it more accurate.

This work is structured as follows. First, Section 2 discusses the existing natural image matting techniques. Then, Section 3 describes each individual step of the proposed background matting algorithm. Next, Section 4 describes the experiments carried out by showing results of applying the proposed background matting algorithm to an image taken by a fixed camera or a selfie camera. Further on, Section 5 presents a few aspects of the proposed algorithm in comparison with the previous algorithms. Finally, Section 6 provides a summary of this work.

2. Related Works

Deep learning brings about huge progress for background replacement with the high-quality illustration of the foreground structure. This section briefly reviews background replacement based on three categories: segmentation, trimap-based matting approaches and deep-learning based methods (see Figure 1).

Segmentation. Segmentation is the process of partitioning an image into multiple segments where each pixel may belong partially to more than one segment. It is typically employed in order to locate boundaries and objects in images. Mask R-CNN (He et al., 2020) is still an excellent solution of segmentation, however DeepLabV3 (Yurtkulu et al., 2019) is considered a state-of-the-art model of human semantic segmentation.

Trimap-based matting. The existing matting methods mostly attain foreground extraction from additional input: scribbles or trimaps. The scribbles indicate various scribbles generated by the user in different areas, while the trimap consists of known background, known foreground, and unknown regions mapped onto the RGB image. The unknown region contains mixtures of background and foreground colors which are considered as the key point for image matting. Therefore, a majority of methods use trimaps as an important input with the original image to recognize the foreground structure. Trimap-based approaches aim to detect details selected from the input image in order to determine unknown areas. These approaches can be classified into two classes: sampling-based and affinity-based approaches.

Sampling-based methods, like those described in (He et al., 2011; Fuertes-Camacho et al., 2019), sample the colors by collecting a set of pixels inside the known foreground and background regions, and the corresponding alpha value of the unknown pixels is estimated by using equation (1). Recently, Li et al. (2020) proposed a learning-based sampling technique that featured a neural network that achieved a good performance. In affinity-based methods (Cho et al., 2019; Grady et al., 2005) alpha values are propagated from the unknown region to the background. These techniques use a set of adjoining pixels to spread alpha values from the known regions to the unknown ones.

Deep-learning matting. The extraction of an object is represented by high-quality semantic segmentation. Most famous approaches in image matting are based on deep learning algorithms

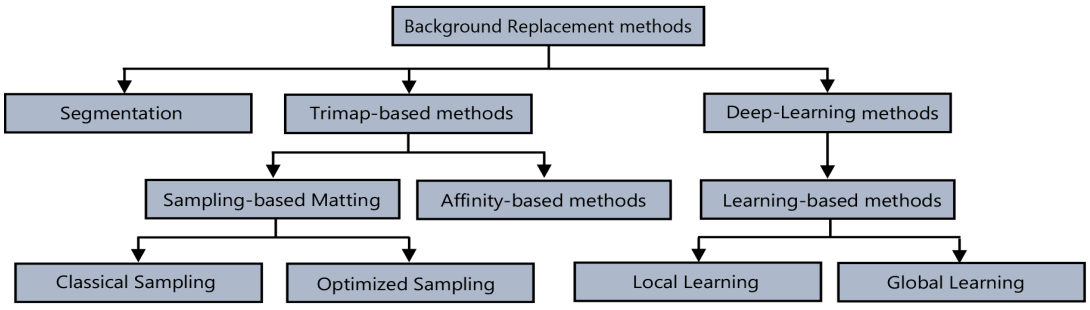

Figure 1. Background Replacement methods

Table 1. Image Matting Techniques used in some literature

References	Image Matting Methods used	Image Matting Techniques used	Trimap Generation	Segmentation
(Sindeyev et al., 2007)	Sampling based method	Bayesian Matting	✓	✓
(Zheng & Kambhamettu, 2009)		Short Sampling Matting	✓	✗
(He et al., 2019)		Global Sampling Matting	✗	✓
(Chen et al., 2013)	Affinity-based methods	KNN Matting	✓	✓
(Sun et al., 2004)		Poisson Matting	✗	✓
(Grady et al., 2005)		Random Walk Matting	✓	✓
(Levin et al., 2007)	Learning-based method	Closed-Form Matting	✗	✓
(Li & Lu, 2020)		CGA Matting	✓	✓
(Hou & Liu, 2019)		Context-Aware Matting	✓	✓

(Sun et al., 2004; Chen et al., 2013). Deep learning matting can be defined as the merger between sampling-based and affinity-based matting. These methods can be classified into two categories: Local learning (Peng et al., 2009) and Global learning (He et al., 2011). The local learning-based method uses neighboring pixels of the estimated pixels to learn the alpha color model. This frequently fits the scribble-based matting. The global learning-based method uses a set of near-labeled pixels to learn the model and fit the trimap-based matting. Recently, deep learning techniques are aimed at separating the matting tasks into trimap adaptation and alpha estimation tasks. Therefore, a large amount of research has been carried out to approach matting color problems. Xu et al. (2017) first proposed an artificial Adobe Deep Image Matting dataset for deep learning-based matting behind a 2-step end-to-end image matting neural network (Lu et al., 2019). Recently, deep learning has shown impressive matting technique results (Chen et al., 2018; Shen et al., 2016). Shen et al. (2016) propose a generation of the trimap from portrait image using a deep neural network and also propose a matting layer (Li et Lu., 2020) which uses the forward and backward propagation strategy. Chen et al. (2018) introduced an automatic human matting algorithm without feeding trimaps. These methods employ image segmentation to predict the alpha mattes from an RGB image input. IndexNet Matting (Yurtkulu et al., 2019) employs encoder and decoder network in order to learn index pooling and un-pooling. Context-Aware Matting (Hou & Liu, 2019) introduces double decoders to estimate alpha and foreground map. GCA Matting (Li et Lu., 2020) uses deep learning to take advantage of the user-generated trimap by employing a trimap-guided attention mechanism. Sengupta et al. (2020) introduced Background Matting where a further background image is captured to serve as a big cue for predicting the alpha matte and therefore the foreground layer. Although this method showed high-quality matting results, Lutz et al. (2018) used a discriminator to make the alpha matte for realistic image input. These works are summarized in Table 1.

In this study, differently from all techniques introduced before, the proposed approach aims to solve image matting problems without user interaction to generate trimaps or scribbles.

3. The Proposed Approach

The proposed technique uses Deep Learning Matting Network Generator (DLMN-G) prediction for the extraction of the alpha value with the foreground color for a given input original image (Figure 2) and a Self-Supervised Refinement Network (SSRN) to refine and recover a high-quality matting detail (Figure 3). The steps of the proposed method are summarized in Figure 4.

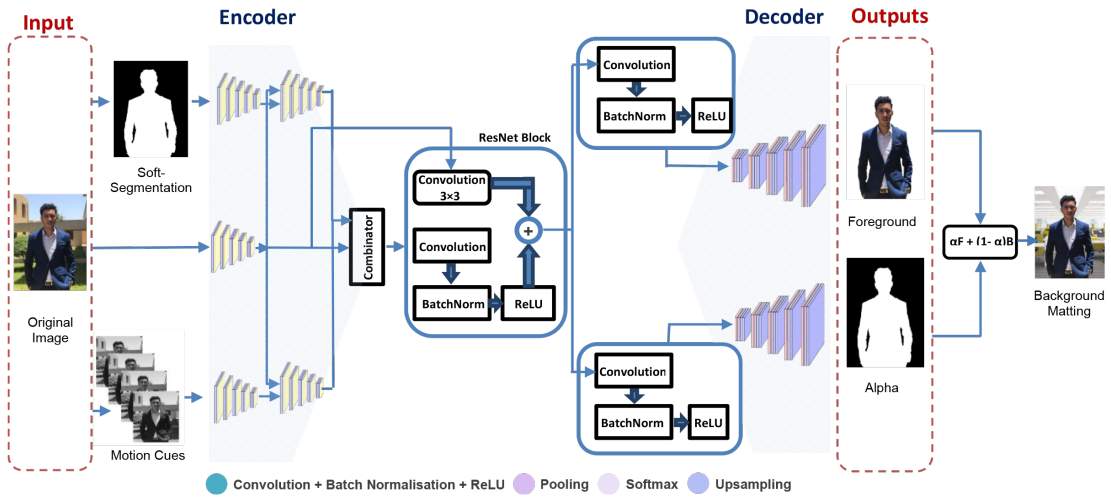

Figure 2. The proposed Deep Learning Matting Network Generator Approach

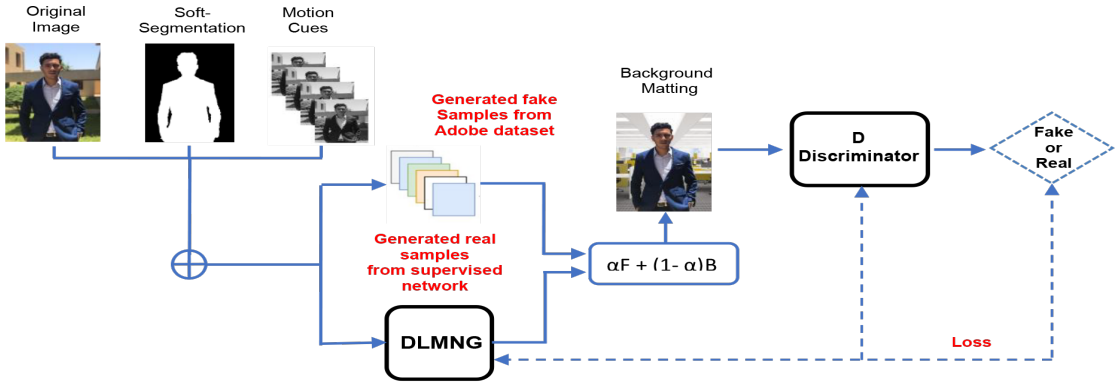

Figure 3. The proposed Self-Supervised Refinement Network Approach

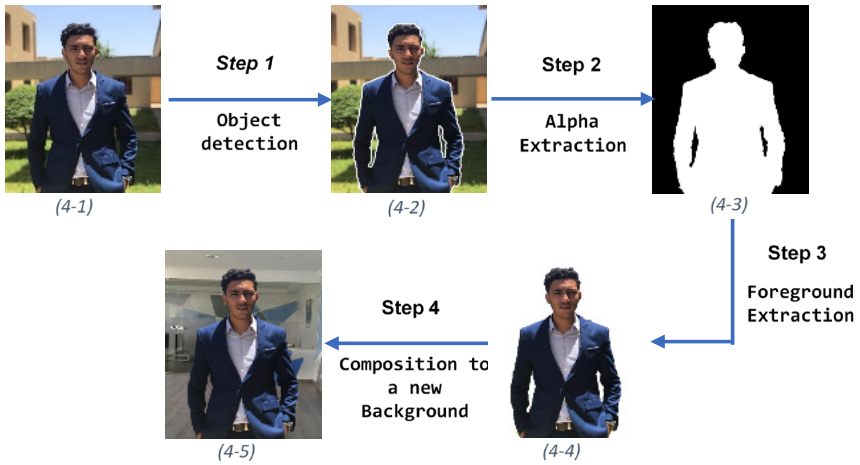

Figure 4. Steps of the proposed contribution

Deep Learning Matting Network Generator

In the first part of the proposed network (Figure 2), the Deep Learning Matting Network (DLMN-G) is introduced that takes as input an original image with a person in the foreground. Firstly, atrous convolution is applied in order to generate soft-segmentation using DeepLabV3 (Yurtkulu et al., 2019). It should be noted that soft-segmentation has been used inside this network in order to automatically generate a trimap while applying erosion and dilatation. Secondly, the motion cues are also generated from the same original image by simply concatenating four original images converted to grayscale. The proposed supervised network consists of multiple encoders, applies 3x3 convolutional layers followed by Batch

Normalization and Rectified Linear Unit (ReLu) activation at each step; each encoder generates 256 feature maps. These feature maps are combined in the combinator bloc to improve the matting process, which generates for each pair 64 features that are passed to the rest of this network. Finally, these three pairs are feeding the residual blocks and decoders. The ResNet Block combines the results of the combinator with that of the original image encoder and its output is then provided as input for two separated ResNet Blocks, one for the alpha value and the other one for the foreground.

The proposed DLMN-G approach (Figure 2), is still not able to handle all problems and difficulties that are present in real data which may be related to the background traces such as those around arms, fingers, hair, and significant parts of the foreground color that match the background color. For solving these, one decided to learn from the self-supervision refinement network.

Self-Supervised Refinement Network

In the second part of the proposed network (Figure 3), a Self-Supervised Refinement Network (SSRN) is illustrated in order to improve the matting quality (Abdellaoui & Kachbal, 2021) as it can be seen in Figure 5 and Figure 6. In order to train this network, the Adobe dataset is used which for each corresponding 100 backgrounds of original image encoder and then provided as input selected from (Xu et al., 2017), implements 431 foreground images with their ground truth alpha mattes.

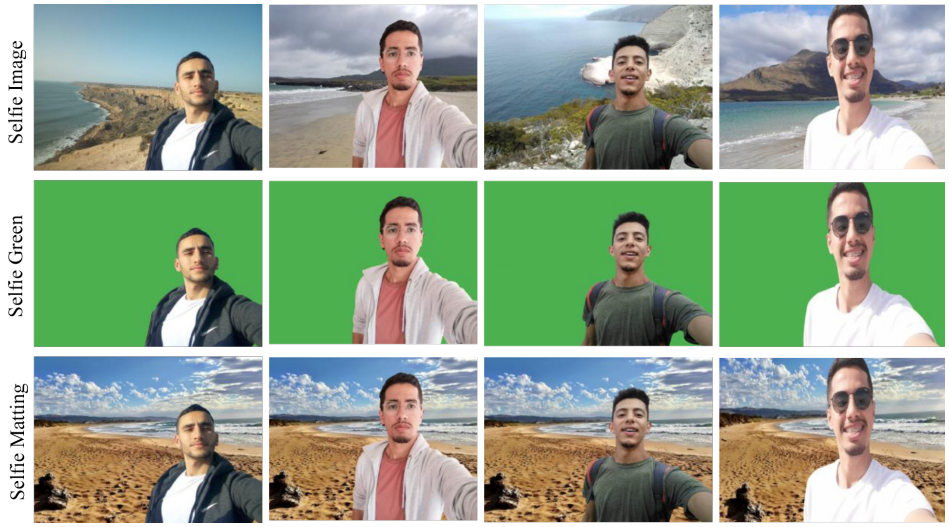

Figure 5. Visual results of the proposed method on images taken by selfie camera

Figure 6. Visual results of the proposed method

One completely followed the composition order employed by Xu et al. (2017) while using the dataset. It has various advantages; firstly, it covers numerous matting instances including fingers, hair, fur, and secondly, many composite images have complicated background textures, making the Adobe dataset robust and practical. In particular, to generate imperfect segmentation, the value of alpha matte is thresholded and morphological operations like erosion, dilation and Gaussian blurring are applied. For the motion cues, one generated a random set of foreground and alpha converted to grayscale. To calculate image input and motion cues, one used the matting equation with B as background. SSRN uses a generative adversarial network which includes a copy of DLMN-G results. The generator of the proposed SSRN tries to estimate a fake matte similar to the Adobe dataset generator while the discriminator network D checks if the compositing result is real or fake. The presented generator G (Lutz et al., 2018) is an encoder-decoder network that can use three-feature maps for a RGB image, single features for the grayscale motion cues information, and single features for the soft-segmentation. The encoder architecture is based on the ResNet101 network where one uses dilated convolution with 2 and 4 rates, respectively instead of normal convolutions in the third and fourth block. The decoder is also a ResNet network that contains a group of convolutional layers and skip connections (Lutz et al., 2018).

The discriminator D tries to enforce G to output more accurate results by differentiating fake from real input. The real input contains original image, soft-segmentation, and motion cues, while the fake input contains the new composition using the alpha matte generated by G.

The real matting network generator DLMN-G and discriminator network D are trained on real inputs with a standard discriminator loss. However, in case that the input image is surely real, DLMN-G chooses to pick setting $\alpha = 1$ everywhere, which ends the process of refinement. The goal of this self-supervised network is to estimate the real alpha, to minimize the alpha prediction loss ζ_{GAN} which is defined as:

$$\zeta_{GAN} = \log D(x) + \log\left(1 - D\left(C\left(G\left(x\right)\right)\right)\right) \quad (2)$$

where x may be defined as a composite image from alpha and foreground. C is a composition function that uses the predicted alpha from G to generate fake image composition. The generator G tries to generate alphas that are close to the real one by attenuating ζGAN while the discriminator D tries to differentiate the real image from a fake composite image by maximizing ζGAN. The objective of the proposed network is completed by the combination of the above losses (Lutz et al., 2018):

$$\zeta_{AlphaGAN}(G,D) = \zeta_{alpha}(G) + \zeta_{comp}(G) + \zeta_{GAN}(G,D) \quad (3)$$

To approve the merger between supervised deep learning matting network generator and self-supervised refinement network, the proposed network was compared to the residual encoder-decoder network (He et al., 2017). It was found that the supervised block uses both segmentation and color difference cues which makes the proposed network more robust therefore enabling it to perform better on portions of the foreground that are quite similar to the background.

To sum up, the proposed methodology works in 5 steps as it is shown in Figure 4. At first, an original image is fed to the network as it is shown by (4-1); then the supervised model will use soft-segmentation to detect objects that should be extracted (4-2); after that, the motion cues and soft-segmentation are employed in order to calculate the alpha values (4-3); these values are employed for detecting and extracting the foreground (4-4); finally, the extracted foreground is applied on on another background (4-5).

4. Results

In order to validate the performance of the proposed approach, it was tested on real data images that were taken by Smartphone or professional camera. All frames were captured in HD (1920×1080). Figure 6 shows the results of the presented approach. It can be noticed that it was possible to produce high-quality matting in natural settings even with complicated backgrounds. Note that the hair details of the girl image in Figure 6 were well recovered by employing this method. Composition examples of selfie images captured by a front-facing camera are illustrated in Figure 5 and, as it can be seen, these compositions have high visual quality in different matting cases like hair and semi-transparency.

5. Discussion

Matting is an important task in both image and video editing, which poses a big challenge

to computer vision. This process needs user interaction and it deteriorates rapidly for most of the existing algorithms. Based on the comparisons with state-of-the-art image matting models, it is obvious that the presented matting method achieves superior performance both quantitatively and qualitatively.

In order to conduct further experiments on the efficiency of the proposed approach, it shall be compared with other existing affinity-based matting techniques like k-Nearest Neighbour (KNN) and Random Walk (RW). The employed method certainly provides a good refinement of the composite image, especially when the foreground and background colors are similar.

The results obtained when using propagation methods (KNN /Random Walk) depend on the complexity of the picture, as it is shown in Figure 7. When the object of the selected image is captured on a solid color background, the two methods offer good results in terms of matting in which the KNN method performs better in the case of the first series of pictures (Alpha (a), Green Screen (a) and Background matting (a)). However, when the object is on a background whose colors are similar to those of the foreground, as in Figure 8, the two methods (KNN / Random Walk) cannot distinguish between background colors and the object colors (Alpha (b), Green Screen (b), Background matting (b)). The KNN method also fails for complex body features and fine features like hair, as it can be seen in Figure 8.

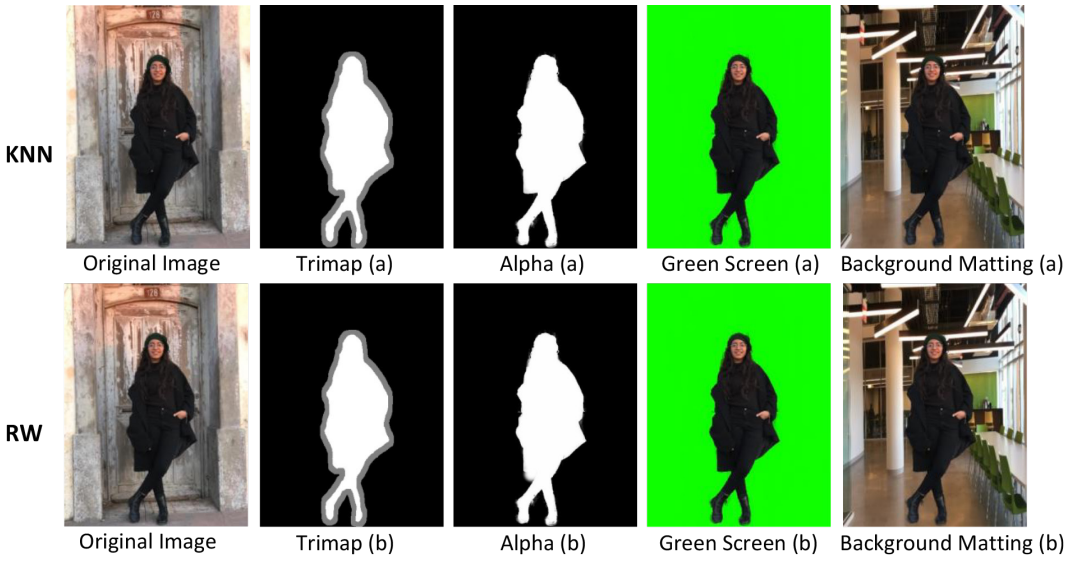

Figure 7. Comparison of KNN and Random Walk matting when the background and the foreground colors are different

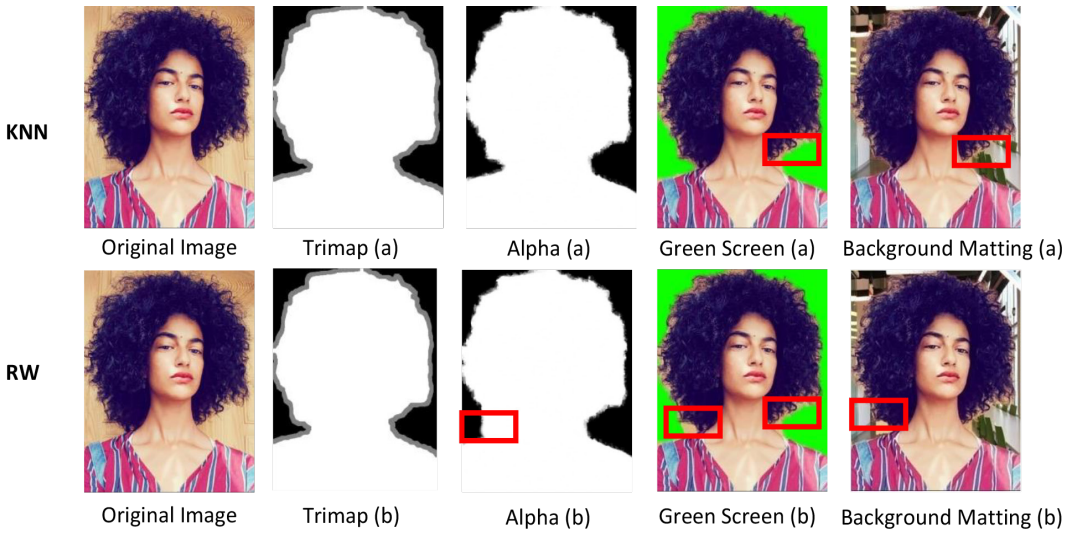

Figure 8. Comparison of KNN and Random Walk matting when the background and the foreground colors are similar

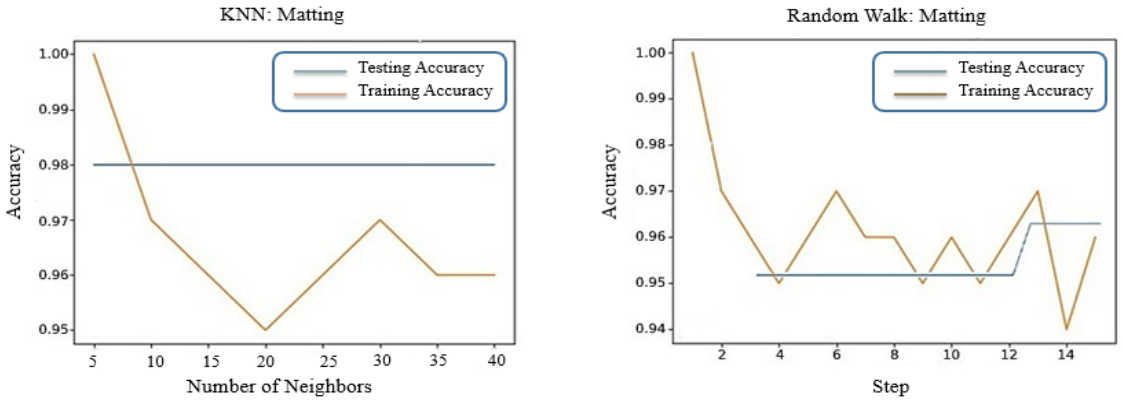

Figure 9. Testing and Training Accuracy for the KNN and Random Walk methods

In Figure 9, it can be noticed that testing accuracy obtained for KNN matting is higher than that of Random Walk method, but it is still failing with image details such as hair as it is illustrated in Figure 8. The advantage of the proposed technique lies in its capacity to estimate a refined alpha matte even with complicated background colors (foreground and background colors are quite similar), which is essential for the accuracy of the estimation. For example, in Figure 8, the input image for KNN and Random Walk methods shares very similar colors with the background, especially in the hair region. Clearly, as it can be seen in Figure 5 and Figure 6, the employed method obtained much better training and validation results for each epoch (Figure 10) in comparison with previous methods (Figure 7 and Figure 8) for which visible mistakes appear in complex regions. Therefore, the obtained results prove the capability of the employed deep learning model to understand complicated images and to predict more details such as those in Figure 6 in comparison with the KNN and Random Walk methods.

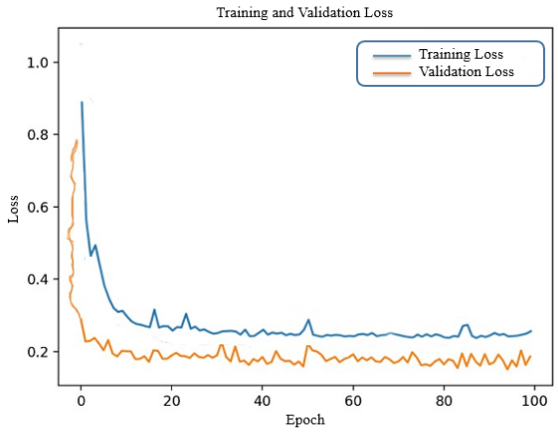

Figure 10. Training and Validation Loss for the presented approach

6. Conclusion

This paper focused on the background image matting problem that is highly important for a lot of applications. It proposes a combination between a supervised and a self-supervised network. The first one aims to extract foreground and alpha values, while the second one is a self-supervised refinement network which performs alpha refinement. Furthermore, the proposed method shows high-quality performance which could contribute to a new state of the art for background matting.

REFERENCES

Abdellaoui, S. E. & Kachbal, I. (2021). Apparel E-Commerce Background Matting, *International Journal of Advanced Research in Engineering and Technology (IJARET), 12*(3), 421-429.

Chen, Q., Li, D. & Tang, C. K. (2013). KNN matting, *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(9), 2175-2188.

Chen, Q., Ge, T., Xu, Y., Zhang, Z., Yang, X. & Gai, K. (2018). Semantic Human Matting. *In Proceedings of the 26th ACM International Conference on Multimedia* (pp. 618-626).

Cho, D., Tai, Y. W. & Kweon, I. (2019). Deep convolutional neural network for natural image matting using initial alpha mattes, *IEEE Transactions on Image Processing*, *28*(3), 1054-1067.

Grady, L., Schiwietz, T., Aharon, S. & Westermann, R. (2005). Random walks for interactive alpha-matting. In *Proceedings of VIIP`05* (pp. 423-429).

He, K., Rhemann, C., Rother, C., Tang, X. & Sun, J. (2011). A global sampling method for alpha matting. In *CVPR 2011* (pp. 2049-2056).

He, K., Gkioxari, G., Dollar, P. & Girshick, R. (2017). Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV),* (pp. 2980-2988).

Hou, Q. & Liu, F. (2019). Context-Aware Image Matting for Simultaneous Foreground and Alpha Estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 4129-4138).

Levin, A., Lischinski, D. & Weiss, Y. (2007). A Closed-Form Solution to Natural Image Matting, *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(2), 228 - 242.

Li, Y. & Lu, H. (2020). Natural Image Matting via Guided Contextual Attention. In *Proceedings of the AAAI Conference on Artificial Intelligence, 34* (pp. 11450-11457).

Lu, H., Dai, Y., Shen, C. & Xu, S., (2019). Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3266-3275).

Lutz, S., Amplianitis, K. & Smolic, A. (2018). AlphaGAN: Generative adversarial networks for natural image matting. In *British Machine Vision Conference* (p. 259).

Peng, H., Chen, S. & Zhang, D. (2009). Local Learning Approach for Natural Image Matting, *Journal of Software, 20*(4), 834-844.

Sengupta, S., Jayaram, V., Curless, B., Seitz, S. & Kemelmacher-Shlizerman, I. (2020). Background Matting: The World Is Your Green Screen. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 2291 - 2300).

Shen, X., Tao, X., Gao, H., Zhou, C. & Jia, J. (2016). Deep Automatic Portrait Matting. In *European Conference on Computer Vision* (pp. 92 - 107).

Sindeyev, M., Konushin, V. & Vezhnevets, V. (2007). Improvements of bayesian matting. In *Proceedings of GraphiCon* (pp. 88 - 95).

Sun, J., Jia, J., Tang, C.-K. & Shum, H.-Y. (2004). Poisson matting, *ACM Transactions on Graphics (TOG), 23*(3), 315-321.

Xu, N., Price, B., Cohen, S. & Huang, T. (2017). Deep image matting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2970 - 2979).

Yurtkulu, S., Sahin, Y. & Unal, G. (2019). Semantic Segmentation with Extended DeepLabv3 Architecture. In *2019 27th Signal Processing and Communications Applications Conference (SIU),* (pp. 1 - 4).

Zheng, Y. & Kambhamettu, C. (2009). Learning based digital matting. In *2009 IEEE 12th International Conference on Computer Vision* (pp. 889 - 896). IEEE.