Image Matting using Neural Networks

Nrupatunga J, Swarnalatha K S
Dept. of Information Science and Engineering
Nitte Meenakshi Institute of Technology, Karnataka, India-560064

Abstract—**Image matting, also refers to picture matting in the article, is the task of finding appealing targets in a picture or sequence of pictures i.e., video, and it has been used extensively in many photo and video editing applications. Image composition is the process of extracting an eye-catching subject from a photograph and blending it with a different background. a) Blue/Green screen (curtain) matting, where the backdrop is clear and readily distinct between the foreground (frontal area) and background (foundation) portions. This approach is now the most used type of image matting. b) Natural picture matting, in which these sorts of photos are taken naturally using cameras or cell phones during everyday activities. These are the present known techniques of picture matting. It is difficult to discern the distinction between the frontal area and the foundation at their boundaries. The current framework requires both the RGB and trimap images as inputs for natural picture matting. It is difficult to compute the trimap since additional framework is required to obtain this trimap. This study will introduce the Picture Matting Neural Net (PMNN) framework, which utilizes a single RGB image as an input and creates the alpha matte without any human involvement in between the framework and the user, to overcome the drawbacks of the prior frameworks. The created alpha matte is tested against the alpha matte from the PPM-100 data set, and the PSNR and SSIM measurement index are utilized to compare the two. The framework works well and can be fed with regular pictures taken with cameras or mobile phones without reducing the clarity of the image.**

Keywords*—*Picture matting; RGB picture; Blue/Green screen; foreground; background

I. INTRODUCTION

With the commonness of cell phones [24], image sensors have expanded strongly, and accordingly expanding the quantity of pictures [19] being captured through electronic devices. Hence, to handle this immense measure of visual information into helpful visual data, image processing has formed into a need in the ongoing moment capturing situation. Image processing has many applications and few of them are up-gradation of image, rebuilding of image, and getting relevant information from an image. Picture matting (or Alpha matting) is a sub-space of image processing that can remove the forefront from a picture, as an alpha matte. It has an assortment of utilization [2][3], like segmentation based on colour in an image [28], removal of reflectional lights, picture colorization, deblurring (bokeh effect), fashion e-commerce [14, 29, 30] and denoising to give some examples.

Numerically, picture matting issue can be demonstrated as:

$$I = \alpha F + (1 - \alpha) B$$

where *I*, *F* and *B* signify input picture, frontal area of output and foundation picture separately. α ranges from 0 to 1 which denotes frontal area opacity. 0% opacity of frontal area and 100% opacity of foundation is obtained when $\alpha = 0$ and vice versa when $\alpha = 1$. For every one of the fragmentary estimations of α, these pixels lie in the blended or obscure locales.

In picture matting, most pixels have α esteems either 1 or 0, the essential issue is to gauge the exact α values for pixels in blended locales, obviously to isolate frontal area and foundation locales. Given an information picture *I*, we need to appraise F, B and α at the same time which is a poorly presented issue. To resolve this issue the most generally utilized technique is a pre-characterized trimap [26] as deduced above. Notwithstanding, for humans it takes more time to get the trimap manually like annotating the edges and will be less precision even if captured through a depth camera or else [17, 18] there arises for another framework for getting a trimap. As a result, a few recent studies [16] have attempted to eliminate model reliance upon this trimap by developing trimap-free techniques.

The proposed PMNN, is a lightweight neural net which disintegrates the representation matting process in three corresponding sub-undertakings and streamlines them at the same time through unambiguous constraints, in order to predict a precise alpha matte from just a single RGB picture. There are two bits of knowledge behind PMNN. In the first place, neural networks are better at learning a bunch of straightforward targets as opposed to a mind boggling one. In this manner, addressing the series matting of sub-targets can accomplish better execution. Secondly, by supervising every sub-target allows different components of the system to know decoupled knowledge, allowing the sub-targets being addressed in one framework.

Due to the removal of the trimap input, the tests suggest that PMNN is more robust in practical scenarios. The technique is attempting to determine whether or not a green screen is required for ongoing picture matting. In rundown, will introduce a neural net framework design called PMNN achieving progressively without trimap representation matting.

II. LITERATURE SURVEY

There are various picture matting approaches [1] (Learning-based matting, propagation-based matting (alpha propagation-based matting), sampling-based matting) have been used to demonstrate the importance of exact alpha matte computation, pixel sample selection, and trimap generation. The interval line based picture matting approach is used to speed up the matte computation. Learning based matting solutions are becoming more popular among established procedures. To improve Matte precision, propagation based techniques and sampling based strategies can be combined, and learning based matting approaches can be used to modify the matte result.

In [4], the author provides a complete semantic matting in the absence of trimap as supplementary information by combining a data set of coarse annotations [20, 25] as information with fine annotation data. To create a prediction of a mask [21, 27] by a network using hybrid information to gauge the mask of coarse semantics, and then introduce a qualitative unification neural net that can bring together the nature of previous output of coarse mask. To estimate the final alpha matte, a matting refinement neural net takes in the combined mask and information picture. By this it expects to predict the opacity of the per-pixel of the frontal area of human locale which is very difficult and for the most part requires trimap and a lot of excellent annotations on the information. Annotation on such information is work concentrated and requires extraordinary abilities beyond the ordinary use, particularly taking into account the exceptionally point by point hair part of people [13].

By incorporating neural net into the process of learning an alpha matte principal propagation, [5] proposes a deep proliferation based picture matting structure. The deep component extraction module, the propagation of matte module, and the learning affinity [15] module are connected to create the deep learning engineering. By using the training process from end to end, these three components can be separated and streamlined in relation to one another. By learning deep picture depictions tuned to propagation matte, the structure creates a semantic-level sequence of comparability of pixels for proliferation. It consolidates the force of deep learning and matte proliferation. The complex of training was approved by the exploratory outcomes from 243K pictures made in light of two benchmark matting data sets. In order to understand [23] deep picture representation with an adaptation to propagation of alpha matte and create more appealing pairwise propagation compatibility, for the design of a DeepMattePropNet.

Past calculations have terrible showing when a picture has comparable frontal area and foundation tones or textures of complication. The main reasons are due to older methodologies. a) using only features of low-level, b) lack of context from high level. In this [6] study, presents a unique deep neural net-based approach that addresses both of these problems. There are two parts to the deep framework. The first section is a convolutional neural encoder-decoder model that predicts an alpha matte of a picture using a picture and the comparative trimap as information sources. The next section is a small convolutional neural net that enhances the prior network's alpha matte forecasts to see more precise alpha quality and finer borders. Moreover, they've also created a massive scope picture matting data - set, which includes training pictures of 49300 and 1000 test pictures. And evaluated the calculation using a picture matting benchmark, a trial set, and a variety of real-life images. The calculation clearly outperforms previous strategies in the trials.

By thoroughly examining numerous differences between the foreground and background images, Jizhizi Li et al.[31] identified the domain gap issue between composite images and real-world photos. They discover that a properly planned composition route RSSN that seeks to lessen the disparities can result in a superior model with impressive generalization ability. Additionally, they offer a benchmark that includes 10,000 portrait photographs with their manually labelled alpha mattes and 2,000 high-resolution real-world animal images, to be used as a test set for determining how well the matting model generalises to real-world images. To fully utilize the trimap information in the transformer block, GyuTae Park et al. [32] suggest a transformer-based image matting model called MatteFormer. The initial step in our procedure is the introduction of a prior-token, which is a global representation of each trimap region (e.g., foreground, background and unknown). As global priors, these prior-tokens take part in each block's self-attention process. PAST (Prior-Attentive Swin Transformer) blocks, which are used at each stage of the encoder and are based on Swin Transformer blocks but differ in a few ways.

III. Data Collection

It's interesting that the offer the Photographic Portrait Matting benchmark (PPM-100) includes 100 representative photographs with meticulous annotations and various foundations. In order to modify the example varieties in PPM-100 and assure test variety, need to characterise a few ordering rules. For illustration, (a) if the entire human body has been included; (b) is the picture foundation occluded; (c) if the person is carrying any other items. Since this is more in line with the practical uses, need to respect small items that people hold as part of the closer view. The examples in PPM-100 do have flamboyant postures and more regular foundations, as shown in Fig. 1. In this approach, the PPM-100 standard is a more thorough benchmark.

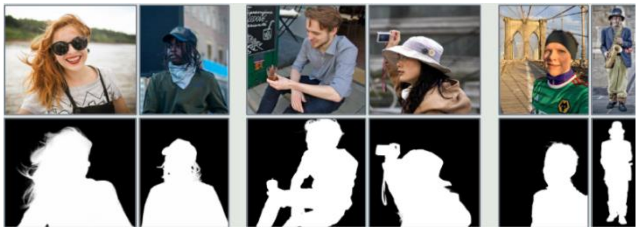


Fig. 1. The PPM-100 data set.

Fig. 1, has more variation in the frontal areas and unique picture foundations. To display experiments having thin hair [10], including more elements [11], and in the absence of bokeh or in the presence of full-body.

IV. Methodology

PMNN comprises three branches, which gain disparate sub-targets through unambiguous requirements. A low-goal branch for measuring human semantics is present in PMNN and is controlled by the base truth matte's thumbnail. As a result, a high-goal branch that is familiar with the center around the representation limits is controlled by the progress area (α ε (0, 1)) in the base truth matte. In order to predict the final alpha matte, a combination branch (controlled by the full base truth matte) is introduced toward the end of PMNN. So that it can examine the branches employed to address each sub-objective in the subsections that follow.

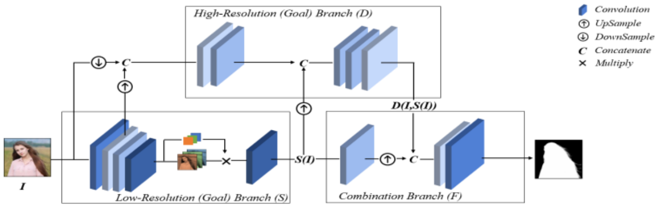


Fig. 2. Block diagram of PMNN.

Fig. 2 uses three associated branches, S, D, and C. PMNN predicts human semantics s_p, edge subtleties d_p, and last alpha matte p given an information image I. So that it is easy to simplify PMNN from beginning to end because the disintegrating sub-targets are correlated and help reinforce one another.

1) Semantic assessment: The first phase of PMNN is to locate the person information Picture *I*, much like other existing frame-work techniques. It is essential that to only use an encoder, or the low-goal branch *S* of PMNN, to extract the high-level meanings, it has two primary benefits. First, since the decoder is no longer used to perform semantic assessment, it will be more effective. Finally, resulting branches and joint streamlining benefit from the high level depiction *S(I)*. As a backbone for *S*, any CNN [9] [12] can be used. Framework chose the MobileNetV2 [7] technology and [8], a great model created for cell phones, as the *S* in order to operate with continuous collaboration.

S(I) is through into a convolution layers that is initiated by the sigmoid activation function to bring down the channel count to 1 in order to estimate the s_p coarse semantic mask. And to control s_p using a small portion of the fundamental truth matte α_g. So the block employ loss of *L2* here because s_p should be smooth,

like in:

$$L_s = \frac{1}{2}||s_p - G(\alpha_g)||_2$$

Where *G* represents Gaussian blur and also represents the 16x down sampling. It eliminates the fine designs (like hair) that are not fundamental for human semantics.

2) Detail assessment: A high-goal branch *D* that uses *I*, *S(I)*, and the low-level components of *S* as information sources surrounds the frontal area representation in the progress. Reusing the low-level spotlights is done in order to reduce the computational burden on *D*. Additionally, enhancing *D* in the three following perspectives: (a) Compared to *S*, *D* has less convolution layer; (b) The convolution layer in *D* are chosen with lower channels; (c) Through *D*, the block doesn't maintain the initial information resolution. In actuality, *D* has a max channel count of 64 and comprises 12 convolution layer. In the first level, the feature resolution of the map is down-scaled to 1/4 of *I*, and in the following two levels, it is recovered.

The *D* outcome results as *D(I, S(I))*, the intricacy between the sub target human semantic *S(I)* of high level is needed in prior for detail assessment. From *D(I, S(I))*, determine the maximum detail matte d_p and grip it along loss of *L1*, as follows:

$$L_d = m_d||d_p - \alpha_g||_1$$

where md serves as a binary mask for L_d to emphasise the representational boundary. Erosion and dilation on α_g result in the production of m_d. If the pixels are inside the transitional area, its attributes are 1, otherwise they are 0. In actuality, the pixels in the dim area of the trimap are those with $m_d = 1$. Despite the possibility of inaccurate quality in d_p for pixels with $m_d = 0$, it has great accuracy for pixels having $m_d = 1$.

3) Semantic-detail combination: A direct CNN module, the combination branch *C* in PMNN combines semantic and detail. Firstly increase the sample size, to align *S(I)* from with *D(I, S(I))*. At that point, *S(I)* and *D(I, S(I))* is added to estimate alpha matte α_p, which is anticipated by:

$$L_\alpha = ||\alpha_p - \alpha_g||_1 + L_c$$

where L_c is the loss of composition. The background picture ground truth, the frontal area ground truth and the outright difference between the information picture *I* and the composited picture obtained from α_p are all estimated.

Using the appropriate amounts of L_s, L_d, and L_α PMNN is ready from beginning to end.

$$L = \lambda_s L_s + \lambda_d L_d + \lambda_\alpha L_\alpha$$

where, adjusting the three loss: λ_s, λ_d, and λ_α are hyper parameters. The process of training the hyper parameter is robust. Which lay down $\lambda_s = \lambda_\alpha = 1$ and $\lambda_d = 10$.

V. Result

The below Fig. 3 represents the outcome from the framework and the Fig. 4 represents the sample comparison between the generated alpha matte and the dataset alpha matte.

The output of a model is tested using an image similarity measures of PSNR - peak signal-to-noise ratio using the formulas as shown,

$$MSE = \frac{\Sigma_{M,N}\,[I_1(m,n) - I_2(m,n)]^2}{M * N}$$

MSE - Mean Square error. The information picture has *M* row and *N* column. The following formula is used to determine the PSNR:

$$PSNR = 10log_{10}(\frac{R^2}{MSE})$$

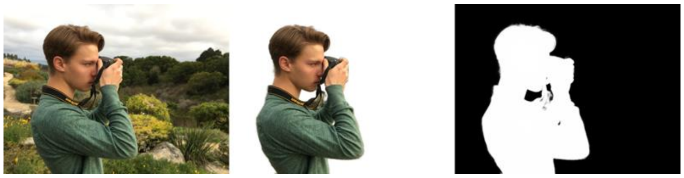
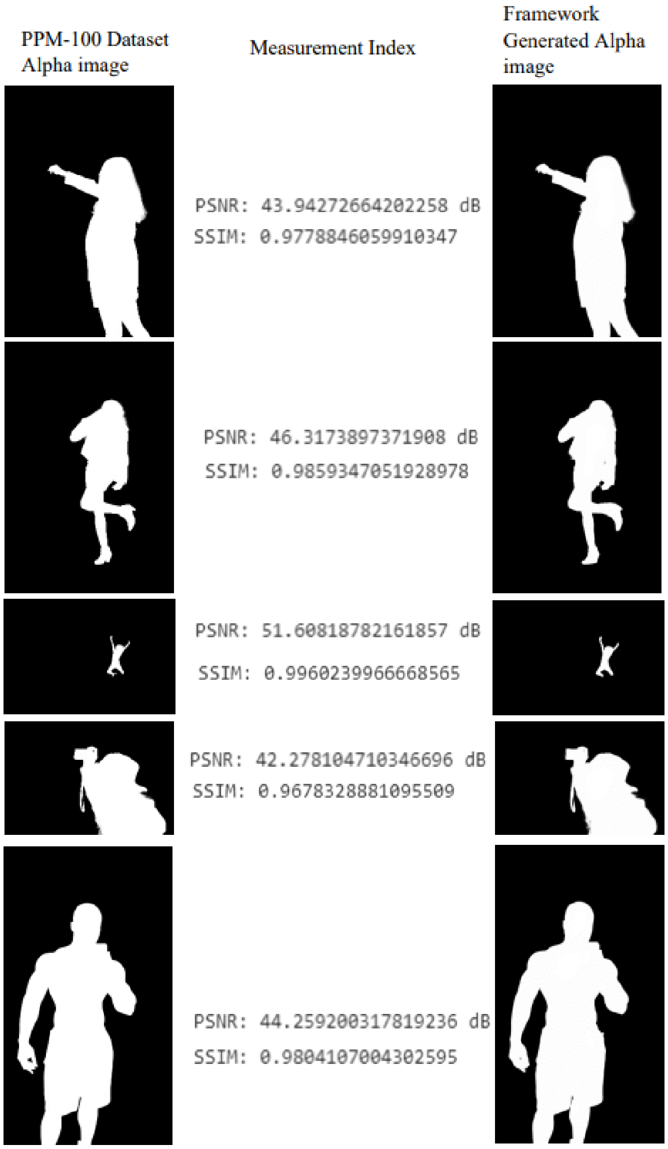


Fig. 3. Output of the framework consisting of RGB Image, extracted frontal area and alpha matte (From left to right).



```
PSNR: 43.94272664202258 dB
SSIM: 0.9778846059910347

PSNR: 46.3173897371908 dB
SSIM: 0.9859347051928978

PSNR: 51.60818782161857 dB
SSIM: 0.9960239966668565

PSNR: 42.278104710346696 dB
SSIM: 0.9678328881095509

PSNR: 44.259200317819236 dB
SSIM: 0.9804107004302595
```

Fig. 4. The Comparison result between the generated and PPM alpha matte.

R is the highest pixel value that the image can contain. Notice that to get decibel value logarithm function is used.

SSIM - Structural Similar Index Measure, is based on structures that may be seen in the photograph. In other words, SSIM assesses the perceived variation between two analogous images. Between -1 and 1, the SSIM score denotes complete structural similarity, where 1 being perfect similarity. The distance X-Y across two windows:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

In Fig. 4, the first column displays the alpha matte from the PPM - 100 data - set. The second third column picture displays the PMNN generated output of alpha matte. The PSNR and SSIM calculated values are displayed in the second column for the generated picture along with the standard data set of PPM - 100.

The framework not only performs well with the data set but also outperforms well along with the real picture captured with the cell phones or camera irrespective of the dimension or pixels or size of the picture.

VI. Conclusion

A quick, easy, and interesting PMNN is introduced as a result of the framework's aim to avoid the use of a green screen [22]. By relying just on RGB photos as data, the technique permits the forecasting of alpha mattes under changing landscape. The framework operates effectively with the PPM-100 data set, as stated in the result section, and also with captured images, regardless of the file size and dimension of the captured image. The frontal region may now be extracted from the image more easily and without the need for human interaction thanks to the model editing.

PMNN is exhibited to have excellent exhibitions on the meticulously sketched PPM-100 benchmark and various here and now facts. The model's future application must include even more accurate and exact extraction of fine hair and fine details in the image with a move toward zero error. The one significant feature that may be introduced is the user's choice of extraction in the image, similar to how the green/blue screen offers the option for the fine extraction of required frontal area from the device captures. This extraction could be an object in the image, a person with some objects, or a person alone.

References

[1] Qingsong Zhu, Pheng Ann Heng, Ling Shao and Xuelong Li, "What's the Role of Image Matting in Image Segmentation?", IEEE International Conference on Robotics and Biomimetics (ROBIO) Shenzhen, 2013.

[2] Jagruti Boda and Dhatri Pandya, "A Survey on Image Matting Techniques", International Conference on Communication and Signal Processing (ICCSP), 2018.pp vol.

[3] Anil singh Parihar, "A Study on Image Matting Techniques", 5th IEEE International Conference on Recent Advances and Innovations in Engineering- ICRAIE 2020.

[4] J. Liu et al., "Boosting Semantic Human Matting With Coarse Annotations," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8560-8569, doi: 10.1109/CVPR42600.2020.00859.

[5] Yu Wang, Yi Niu, Peiyong Duan1, Jianwei Lin, Yuanjie Zheng, "Deep Propagation Based Image Matting", International Joint Conferences on Artificial Intelligence Organization, Twenty-Seventh International Joint Conference on Artificial Intelligence, Pages 999-1006.

[6] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang, "Deep Image Matting", IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[7] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks", CVPR, 2018.

[8] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132-7141, doi: 10.1109/CVPR.2018.00745.

[9] https://towardsdatascience.com/a-comprehensive-introduction-to-different-types-of-convolutions-in-deep-learning-669281e58215.

[10] Chang Liu, Henghui Ding and Xudong Jiang, "Towards Enhancing Fine-grained Details for Image Matting", IEEE Winter Conference on Applications of Computer Vision (WACV), 2021.

[11] Lei Liu and Yingyun Yang, "Image Matting Based On Deep Learning And Image Dissection", IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC) 2019.

[12] Donggeun Yoon, Jinsun Park, and Donghyeon Cho, “Lightweight Deep CNN for Natural Image Matting via Similarity-Preserving Knowledge Distillation”, IEEE SIGNAL PROCESSING LETTERS, VOL. 27, Page(s): 2139 - 2143, 2020.

[13] Xuqian Ren, Yifan Liu and Chunlei Song, “A Generative Adversarial Framework For Optimizing Image Matting And Harmonization Simultaneously”, IEEE International Conference on Image Processing (ICIP), 2021.

[14] Rishab Sharma, Rahul Deora, and Anirudha Vishvakarma, “AlphaNet: An Attention Guided Deep Network for Automatic Image Matting”, International Conference on Omni-layer Intelligent Systems (COINS) 2020.

[15] Guilin Yao, Dongai Jiang, and Jianliang Sun, “An Affinity Based Matting Method Based on Multi-Scale Space Fusion”, 33rd Chinese Control and Decision Conference (CCDC) 2021.

[16] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou2 and et.al., “Attention-Guided Hierarchical Structure Aggregation for Image Matting”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020.

[17] Bo Liu, Haipeng Jing, Guangzhi Qu and Hans W. Guesgen, “Cascaded Segmented Matting Network for Human Matting”, IEEE Access Volume: 9, Page(s): 157182 – 157191, 2021.

[18] YanLong Xu, Hui Fan∗, Jinjiang Li, “DenseNet Matting Algorithm Based on Embedded Improved SKNet”, IEEE International Conference on Progress in Informatics and Computing (PIC) 2020.

[19] Haichao Yu, Ning Xu and et.al. “High-Resolution Deep Image Matting”, The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), 2021.

[20] Yaoyi Li, Jianfu Zhang, Weijie Zhao, Weihao Jiang and Hongtao Lu, “Inductive Guided Filter: Real-Time Deep Matting with Weakly Annotated Masks on Mobile Devices”, IEEE International Conference on Multimedia and Expo (ICME) 2020.

[21] Qihang Yu, Jianming Zhang and et.al., “Mask Guided Matting via Progressive Refinement Network”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[22] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman, “Background matting: The world is your green screen”, CVPR, 2020.

[23] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia, “Deep automatic portrait matting”, ECCV, 2016.

[24] R. Dong, B. Wang, Z. Zhou, S. Li and Z. Wang, "Design and Implementation of an Image Matting System on Android Phones," 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, 2015, pp. 522-526, doi: 10.1109/IHMSC.2015.53.

[25] X. Fang, S. -H. Zhang, T. Chen, X. Wu, A. Shamir and S. -M. Hu, "User-Guided Deep Human Image Matting Using Arbitrary Trimaps," in IEEE Transactions on Image Processing, vol. 31, pp. 2040-2052, 2022, doi: 10.1109/TIP.2022.3150295.

[26] V. Gupta and S. Raman, "Automatic trimap generation for image matting," 2016 International Conference on Signal and Information Processing (IConSIP), 2016, pp. 1-5, doi: 10.1109/ICONSIP.2016.7857477.

[27] C. Orrite, M. A. Varona, E. Estopiñán and J. R. Beltrán, "Portrait Segmentation by Deep Refinement of Image Matting," 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 1495-1499, doi: 10.1109/ICIP.2019.8799367.

[28] B. Qian and X. Gu, "Automatic ID Photos Matting Based on Improved CNN," 2019 18th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), 2019, pp. 96-99, doi: 10.1109/DCABES48411.2019.00031.

[29] D. Shin and Y. Chen, "Deep Garment Image Matting for a Virtual Try-on System," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 3141-3144, doi: 10.1109/ICCVW.2019.00384.

[30] B. Yuan, Z. Lu, J. -H. Xue and Q. Liao, "A New Approach to Automatic Clothing Matting from Mannequins," 2019 IEEE International Conference on Multimedia and Expo (ICME), 2019, pp. 880-885, doi: 10.1109/ICME.2019.00156.

[31] Li, J., Zhang, J., Maybank, S.J. et al. Bridging Composite and Real: Towards End-to-End Deep Image Matting. Int J Comput Vis 130, 246–266 (2022). https://doi.org/10.1007/s11263-021-01541-0.

[32] GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, Nojun Kwak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11696-11706.