Deep Garment Image Matting for a Virtual Try-on System

Dongjoe Shin
University of Portsmouth, UK
dongjoe.shin@port.ac.uk

Yu Chen
Metail Ltd. Cambridge, UK
yu@metail.com

Abstract

To improve online shopping experience, many fashion retailers try to provide high quality garment images, capturing fine details as well as various opacities. A skilled operator can deliver a satisfactory result using manual segmentation tools, but it is challenging to scale up this process to address seasonal demands. To balance the quality and the processing cost, we investigate the use of a deep learning based matting technique that can produce a high quality alpha map from an approximate garment segmentation. The proposed model adopts the deep image matting model [10], but we replace the refinement network with a sequence of recursive convolutional network (RCN) units. Our main motivation for this modification is that the fine garment details created by different materials are represented better with the mixture of the image features from different scales. Therefore, we need to construct deeper convolutional layers for better scale analysis but we also need to maintain the number of unknowns low as producing training data is expensive. The proposed RCN based refinement network can address these conflicting restrictions well and our experiments demonstrate that it can achieve a lower training loss and produce better prediction results than the baseline refinement model under the same training condition.

1. Introduction

Realistic garment visualisation is a fundamental functionality requested by most online fashion retailers. Conventional approaches used in many e-commerce platforms heavily rely on a static garment image dressed on a fixed mannequin. This passive visualisation is often considered as a limitation for delivering a better user experience, and various immersive visualisation techniques have been tried recently [4]. If we adopt simple image-based rendering for photorealistic garment visualisation, e.g. producing a final rendering result by overlaying a warped garment image on a rendered 3D body model (see Fig. 1(b) and (c)), it is necessary to prepare a garment dataset extracted from initial studio photographs (see Fig. 1(a)). However, populating such a dataset is not trivial due to the complex shape of a garment and the various opacities created by different garment materials.

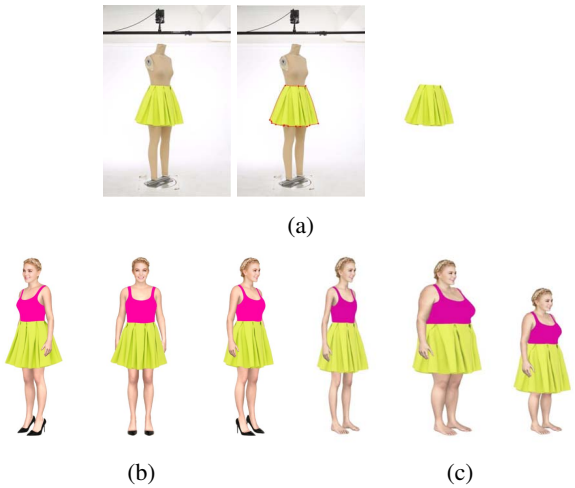


Figure 1: Example of virtual try-on using garment images; a) a process of extracting an initial cutout from a few 2D control points, which will be used to deliver a trimap for our matting solution; b) different views of a garment rendered by a 3D body model and garment cutouts; and c) warped garment images on personalised body models generated from different body measurements.

This image separation problem is closely related to an image matting problem, where the main objective is estimating the unknown alpha values involved in the image compositing process [6]. For example, if we denote the alpha values as a single channel float matrix, A, the image composition result I_c from a foreground (FG) image I_f and a background (BG) image I_b can be described as

$$I_c = A \odot I_f + (\mathbf{1} - A) \odot I_b, \tag{1}$$

where $\odot$ denotes an elementwise multiplication operator and each element of A is between 0 and 1.

Although most general garments (e.g. simple trousers or a T-shirt) can be safely extracted by latest semantic segmentation techniques, the fuzzy separation shown in (1) is much

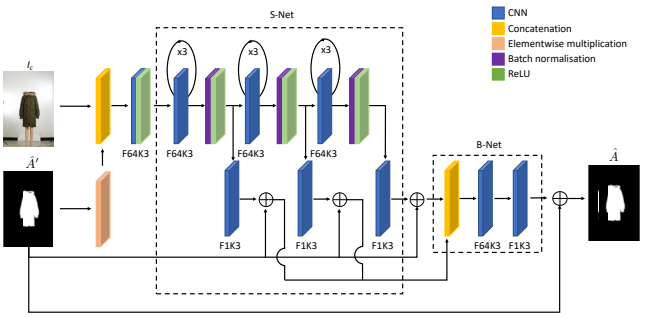


Figure 2: Network topology of the proposed recursive refinement network (RRN), where S-Net is designed to extract multiple frequency details from different scales whilst B-Net blends them to produce a final result. F and K values shown under each CNN layer diagram (i.e. the blue block) denote the number of feature maps and the size of a convolutional kernel, respectively.

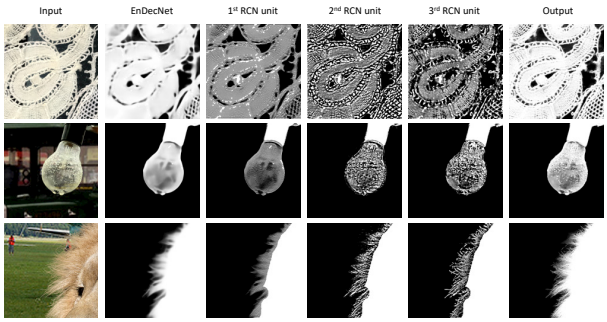


Figure 3: Example of S-Net results: the proposed RRN produces the final results (the last column) by blending the results from intermediate scales (the column 2-5).

suitable for capturing the high frequency image details observed in many fashion garments. The difficulty is that recovering A, I_f, and I_b from given I_c in (1) is a highly ill-posed problem [3]. Thus, it is normally incorporated with additional constraints provided by users, such as a trimap (which is an initial image segmentation defining true FG, true BG, and unknown pixels) or simple scribbles on an image.

The latest approaches argue that this manual annotation could be an issue for developing a fully automated matting system, and propose to learn the initial segmentation (i.e. a trimap) as well as an alpha map using deep learning networks [9, 12, 1]. However, we have noticed that many fashion retailers prefer to have some interaction points (such as a set of control points defining a trimap) to adjust the matting quality. In addition, conventional photograph processes have already implemented a pipeline performing initial garment segmentation to remove the background scene (see Fig. 1(a)). Thus, our question in this project is more about developing a semi-automated system, which can utilise the existing initial cutouts as the matting constraints.

2. The Proposed Model

Deep CNN matting (DCNN) is one of the early deep matting networks developed for a natural image matting problem [2]. However, this model can be seen as refining the alpha maps from the initial estimations from conventional approaches rather than giving direct estimation from an input image. More practical matting solution can be found in Deep Image Matting (DIM) [10]. In this approach, the matting problem is posed as image translation, where a RGB image is augmented with a trimap to produce a single channel alpha map. At the beginning of this model, the VGG network is used to encode a natural input image to a set of smaller feature maps, which turn into a scaled-up alpha map in a follow-up decoding network. It is also worth noting that DIM employs an additional refinement network on top of the encoder-decoder network (EnDecNet) structure, and the authors mentioned that the additional refinement is beneficial to enhance the oversmoothed alpha maps produced from the initial result.

An intriguing observation from the recent development is that the refinement network originally proposed in DIM is getting redundant [8]. This is mainly because of the emergence of accurate networking techniques and adversarial training. However, the advances in deep Single Image Super Resolution (SISR) [11] support the argument that a simple convolutional network is actually effective to learn the non-linear mapping from a low resolution image to a high resolution output. We believe this incremental refinement is better than having a single EnDecNet in our working scenarios; *e.g.* we can only retrain a small refinement network whenever we have more challenging garments.

Based on this observation, we have developed a new refinement network. Since a deeper network generally performs better in SISR, we need to implement multiple convolutional layers. However, we also need to minimise the number of unknown parameters because preparing high resolution training samples (particularly for a complex garment images) is very expensive and there is no publicly available garment image dataset with high resolution alpha maps at the moment to the best of our knowledge. To address this, we adopt a recursive convolutional network (RCN) [7]. Since a RCN reuses fixed weights over time, it does not increase the number of unknown parameters too much but simulates a similar effect that a deeper network can create.

Figure 2 shows the overall network architecture for the proposed Recursive Refinement Network (RRN). We use the same EnDecNet proposed in DIM to generate the initial alpha map (i.e. $\hat{A}'$ in Fig. 2), which will be fed to the RRN with a 320×320 composite RGB image, I_c. The RRN

consists of two subnets called a scale analysis network (S-Net) and a blending network (B-Net). The main motivation of S-Net is capturing the fuzziness of furry garment more accurately. For example, a garment with fine hair strands or voile materials may be considered as high frequency image features only but in a larger scale these fine details are better to be represented with a smooth transition. In other words, to represent the garment details more effectively, we need a special image sharpener that can blend the features from multiple different scales. The recursive connection in a RCN can create an effect of increasing the size of filter kernel [5] which allows us to perform this scale analysis.

To understand what is trained in the S-Net, we store the results from each RCN unit[1]. Since they are a single channel feature map, we can visualise them as an image, which is shown in Fig. 3. As shown in the figure, we have found that each RCN unit works as an image sharpener and employing multiple RCN units can help to enhance image features from different scales. As the RRN goes deeper, we can add more features from higher scales and this scale analysis is good for depicting the different size of the gaps in a voile pattern and it makes fine hairs look more highlighted by producing smoother alpha values in its background.

3. Experimental Results

To compare the performance of the proposed refinement network, we adopt an incremental training strategy. For example, the first subnet of DIM (i.e. the EnDecNet) is trained initially and the same model is attached to different refinement networks for additional training.

As a loss function for the EnDecNet, we use the same metrics suggested in DIM, i.e. a weighted sum of the alpha prediction loss and the composition prediction loss defined on the unknown area of a trimap. To facilitate the training process, a pre-trained VGG model is imported to initialise the encoder network and the zero-mean normalisation is applied to the RGB values of an input image. This can help us to achieve 0.064 loss for the EnDecNet after 94 epochs. On the other hand, when training the refinement network, we only use the alpha prediction loss, and a standard Adam optimisation with a fixed learning rate is used for all training sessions.

The training and testing samples are obtained from the Adobe dataset [10]. Since the number of the images in the dataset is not sufficient for training a large network (e.g. in our test we use 429 foreground images for training and 50 foreground images for testing), each sample FG image is subdivided and composited with a random background image from MS-COCO for training and ImageNet dataset for testing. Random data augmentations, such as random scaling and image flipping, are also applied dynamically during each batch construction, and this can help to improve the generality of a trained model.

[1] A single RCN unit in our paper denotes a sequence of layers, such that RCN-BN-ReLU-CNN.

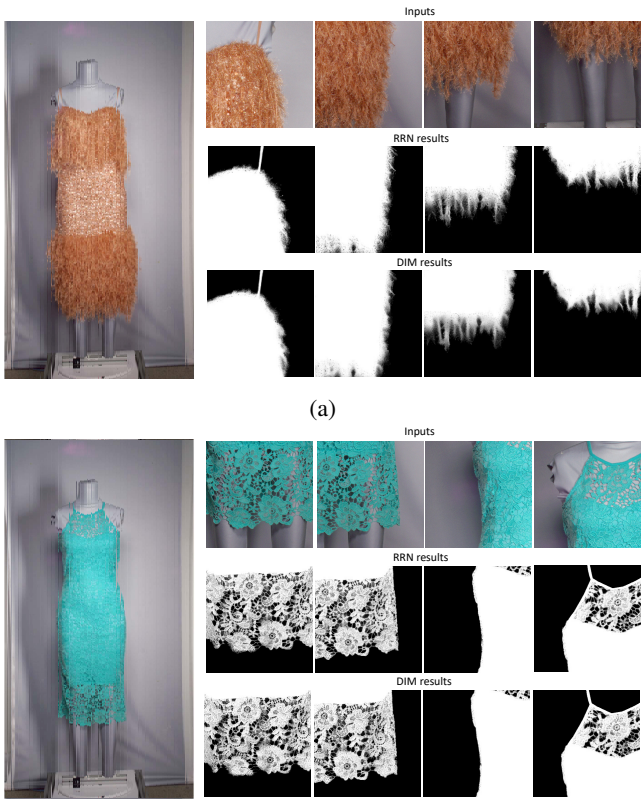
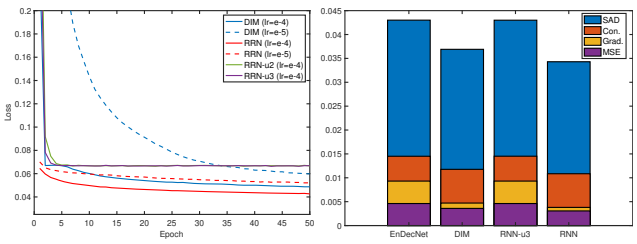


Figure 4: Examples of garment prediction: the estimate results from different parts of a garment are presented in the second row (RRN) and third row (DIM).

The matting performance is generally dependent on an initial trimap. In order to make the resulting model more robust against different tirmaps, we populate multiple trimaps for the same alpha map during the training stage. Various morphological operators (e.g. eroding, dilating, opening, closing with different kernel sizes and processing iterations) are applied to a reference alpha map for the automatic creation of such a trimap.

Figure 4 shows the prediction results from the actual garment images captured from a studio environment. This studio image normally has less cluttered background but the colour values can be affected by different lighting conditions. Although both RRN and DIM are trained with the same dataset, the proposed method generally performs better than DIM and it also demonstrates that the proposed RRN can be used for extracting the fine garment details which look more like refined binary segmentation (see Fig. 4(b)) as well as the fuzziness of hair strands shown in Fig. 4(a).

Fig. 5 (a) summarises the training performance of the



(a) (b)

Figure 5: (a) The training performance of 4 different matting methods at a fixed learning rate (lr), where the dotted lines represent the results of a different lr value with the same RRN and DIM model; (b) Evaluation results of different matting methods.

different models. Four refinement networks (i.e. DIM, RRN, RRN-u3, and RRN-u2) are trained up to 50 epochs using the same optimisation parameters, i.e. 8 samples for each training batch and a fixed learning rate at 10^{-4}. Our baseline refinement network (i.e. DIM) is quickly converged around 0.049 (see the solid blue line in Fig. 5 (a)). However, the proposed RRN can start from lower loss (i.e. 0.065) and produce better converged values around 0.043 (see the solid red line in the figure). A lower learning rate 10^{-5} is also tested to see the performance change and it shows that DIM (i.e. the dotted blue line) is more sensitive to the learning rate than the proposed RRN (i.e. dotted red line).

RRN-u3 and RRN-u2 are created to test the performance of a simpler RRN configuration. These networks are exactly same as the refinement network of DIM but some convolutional layers are modified to include additional recursive loops, e.g. RRN-u3 has 3 recursive convolutional layers. Thus, the number of trainable parameters of RRN-u2 is the same as that of DIM and RRN-u3 has 36,928 more parameters than DIM. Although there is no change in terms of the number of the trainable parameters, RRN-u2 performs worse than DIM and converges around 0.067 and the RRN-u3 produces a similar result (i.e. 0.066).

4. Conclusions and Future Work

In this paper, we propose a new refinement network that can improve the performance of garment image matting for virtual try-on solutions. Motivated by the recent development in the SISR research, the proposed method adopts a recursive convolutional network (RCN) to configure a new refinement model. The use of RCN proves that we can maintain the number of unknown parameters low, whilst producing a similar effect created by a deeper network.

The proposed method is intended to provide a semi-automatic image matting solution, so that a human creator can adjust results by tweaking an initial trimap, if needed. Therefore, the best training result is achieved when we can simulate the trimap patterns that human operators would create. These patterns can be trained with an additional sub-network or using recent GAN approaches, but we have found collecting such training data is expensive. In the future, it is worth investigating new subjective metrics for garment matting. Also, hardware-based image segmentation should be reviewed as a means of populating a new garment image dataset for matting more efficiently.

References

[1] Q. Chen, T. Ge, Y. Xu, Z. Zhang, X. Yang, and K. Gai. Semantic human matting. *2018 ACM Multimedia Conference on Multimedia Conference - MM '18*, 2018. 2

[2] D. Cho, Y.-W. Tai, and I. Kweon. Natural image matting using deep convolutional neural networks. *Lecture Notes in Computer Science*, pages 626–643, 2016. 2

[3] Y.-Y. Chuang, B. Curless, D. H. Salesin, and R. 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, Dec. 2001. 2

[4] S. Hauswiesner, M. Straka, and G. Reitmayr. Virtual try-on through image-based rendering. *IEEE Transactions on Visualization and Computer Graphics*, 19(9):1552–1565, 2013. 1

[5] J. Kim, J. K. Lee, and K. M. Lee. Deeply-recursive convolutional network for image super-resolution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1645, 2016. 3

[6] A. Levin, D. Lischinski, and Y. Weiss. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):228–242, 2008. 1

[7] M. Liang and X. Hu. Recurrent convolutional neural network for object recognition. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015)*, pages 3367–3375, June 2015. 2

[8] S. Lutz, K. Amplianitis, and A. Smolic. Alphagan: Generative adversarial networks for natural image matting. *CoRR*, abs/1807.10088, 2018. 2

[9] X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia. Deep automatic portrait matting. In *14th European Conference on Computer Vision (ECCV 2016)*, volume I, pages 92–107, October 11-14 2016. 2

[10] N. Xu, B. Price, S. Cohen, and T. Huang. Deep image matting. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR2017)*, pages 311–320, 2017. 1, 2, 3

[11] W. Yang, X. Zhang, Y. Tian, W. Wang, and J. Xue. Deep learning for single image super-resolution: A brief review. *CoRR*, abs/1808.03344, 2018. 2

[12] B. Zhu, Y. Chen, J. Wang, S. Liu, B. Zhang, and M. Tang. Fast deep matting for portrait animation on mobile phone. *Proceedings of the 2017 ACM on Multimedia Conference - MM '17*, 2017. 2