AUTOMATIC SKIN AND HAIR MASKING USING FULLY CONVOLUTIONAL NETWORKS

Siyang Qin, Seongdo Kim, Roberto Manduchi

University of California Santa Cruz, Santa Cruz, CA
{siqin, seongdo, manduchi}@soe.ucsc.edu

ABSTRACT

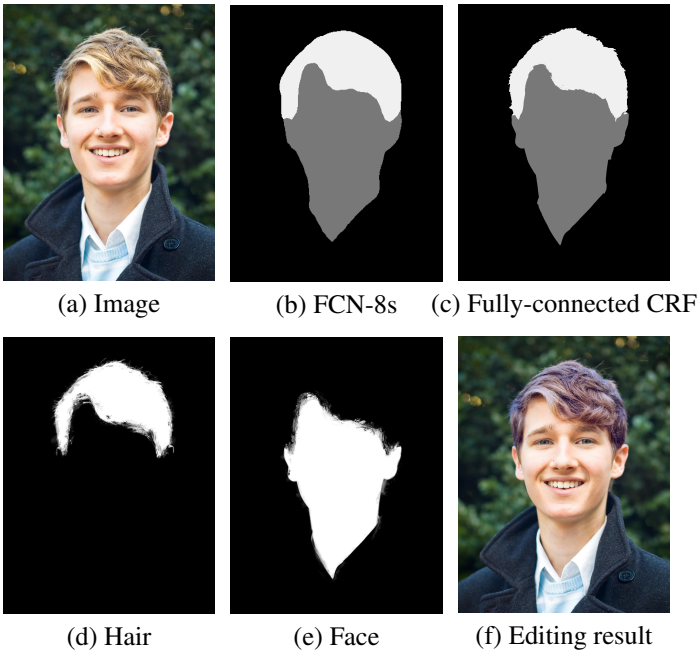
Selfies have become commonplace. More and more people take pictures of themselves, and enjoy enhancing these pictures using a variety of image processing techniques. One specific functionality of interest is automatic skin and hair segmentation, as this allows for processing one's skin and hair separately. Traditional approaches require user input in the form of fully specified trimaps, or at least of "scribbles" indicating foreground and background areas, with high-quality masks then generated via matting. Manual input, however, can be difficult or tedious, especially on a smartphone's small screen. In this paper, we propose the use of fully convolutional networks (FCN) and fully-connected CRF to perform pixel-level semantic segmentation into skin, hair and background. The trimap thus generated is given as input to a standard matting algorithm, resulting in accurate skin and hair alpha masks. Our method achieves state-of-the-art performance on the LFW Parts dataset [1]. The effectiveness of our method is also demonstrated with a specific application case.

Index Terms— Skin and hair segmentation, Convolutional Neural Networks, Image Matting

1. INTRODUCTION

High quality skin and hair segmentation plays an important role in portrait editing, face beatification, human identification, hairstyle classification and many other computer vision problems. To obtain accurate segmentation mask, there exist many interaction-based methods which require users to label some areas as foreground or background. Graph cut [2], fully-connected CRF [3], and matting algorithms [4] are some of the techniques used to produce accurate masks. Some authors [5, 6, 7, 8, 9] have attempted to segment skin and hair automatically; however, without user input, the quality of the result is usually poor. Automatic robust and accurate skin/hair segmentation is still an open problem, even state-of-the-art algorithms have difficulties with different face and hair color, hairstyle, head poses, and confounding background color.

In this work, we take advantage of the power of fully convolutional networks (FCN). These architectures have been used successfully in a wide range of computer vision problems, including semantic segmentation [10, 11, 12] and boundary detection [13, 14]. We fine-tune the FCN-8s model [10] using LFW Parts dataset [1] with three labels: skin, hair, and background. The output of the network is a pixel-wise prediction into one of these labels. The FCN is trained in an end-to-end manner, and it can capture both local and global context information. In FCN-8s (with two skip layers) the final prediction map is downsampled by 8, which leads to poor localization. This is a particularly serious problem for hair segmentation, as the hair mask often contains very thin details. To overcome this issue, we follow the approach of Chen *et al.* [11], and feed the FCN output to a fully-connected CRF [3] as a unary energy term, while spatial and color information are used as pairwise terms. Our method achieves state-of-the-art performance on LFW Parts dataset with 97.32% accuracy.



(a) Image (b) FCN-8s (c) Fully-connected CRF

(d) Hair (e) Face (f) Editing result

Fig. 1. (a) original image, (b) output of FCN-8s model (section 3.1), (c) output of fully-connected CRF (section 3.2), (d) and (e) are outputs of matting step (section 3.3), (f) edited image with modified hair color.

Using FCN and fully-connected CRF we have been able to obtain good segmentation masks, with state-of-the-art performance on the LFW Parts dataset (97.32% accuracy). Yet, these results are still not accurate enough for portrait editing. This is due to two main reasons. First, the segmentation mask fails to capture some fine details of the hair area. Second, the mask produced by the algorithm is binary, and does not give a natural, soft transition between the skin and hair region. To solve both problems, we automatically generate trimaps based on our binary segmentation masks using morphological operations, then employ standard image matting algorithm to obtain an appropriate alpha masks for the hair and skin areas. The accurate binary segmentation masks produced by FCN and fully-connected CRF enable us to generate high-quality trimaps, and the high-quality alpha matting.

In summary, our contributions are three-fold: (1) We apply FCN followed by fully-connected CRF to automatically segment skin and hair, achieving state-of-the-art accuracy on the LFW Parts dataset. (2) We produce soft and high quality alpha matte for hair and skin regions by combining image matting and binary masks. (3) We demonstrate how high-quality skin/hair masks can be employed for image enhancement applications such as modifying the color of skin or hair and smoothing one's skin.

2. RELATED WORKS

In this section we briefly review previous work in automatic skin and hair segmentation, as well as the main algorithms used in our system.

The work of Yacoob *et al.* [5] relies on a simple color model to recognize hair pixels. However, this method cannot deal with large hair color variation and confounding background color. Lee's algorithm [6] adds location information for increased robustness, and reformulates the segmentation task in terms of a Markov network inference problem, which is solved via Loopy Belief Propagation. Huang *et al.* [15] trained a standard CRF on images from the LFW dataset to build a skin, hair and background classifier. Each node in the network is a superpixel; information from color, texture and location is used. For adjacent superpixels, the sum of PB [16] values along the boundary, color and texture histogram residual are used to compute pairwise potential.

Wang *et al.* [17, 18] incorporated a part-based model for hair and face segmentation. A measurable statistic, called Subspace Clustering Dependency (SC-Dependency), was used to capture the co-occurrence probabilities between local shapes. This algorithm produces reasonable results but suffers from poor localization, which makes it unsuitable for face and hair editing tasks. Recently, Kae *et al.* [8] proposed a model using CRF to capture local appearance features and restricted Botlzmann machines to model global shapes. The result is evaluated on the LFW Parts dataset [1], which contains 2927 images (2000 for training, 927 for testing) with ground truth labels for each superpixel. The work of Liu *et al.* [9] introduced a multi-objective learning method for deep convolutional networks that jointly models pixel-wise likelihoods and label dependencies. A nonparametric prior was used for additional regularization, resulting in better performance.

In recent years, fully convolutional networks (FCN) [10] have been used in a variety of contexts in computer vision. In our work, we adopt the FCN-8s model [10] as it performs reasonably well and it is easy to train. Although FCN-8s produces robust results, it suffers from poor localization. To overcome this limitation, fully-connected CRF [3] model are often used. Chen *et al.* [11] feed the label assignment probabilities produced by FCN to a fully-connected CRF, where the two modules are trained separately. Zheng *et al.* [12] reformulate mean-field approximate inference for the fully-connected CRF as a Recurrent Neural Network (RNN), which enables an end-to-end training process.

3. OUR APPROACH

Our proposed framework is shown in Figure 2. The input image is fed to the FCN-8s [10] network to produce a pixel-wise prediction map. This is followed by a fully-connected CRF, used for improved localization. Finally, an image matting algorithm is applied to obtain accurate and soft skin and hair alpha mattes, from trimaps automatically generated using morphological operations on the fully-connected CRF output. The different system components are described in detail in the following.

3.1. Fully Convolutional Network

In order to robustly segment skin and hair from cluttered background, both local and global context information need to be taken into consideration. For this purpose, we use the widely successful FCN algorithm, which has the ability to leverage different scales for feature computation. Specifically, we use the FCN-8s [10] model, which achieves excellent performance in semantic segmentation tasks, and has also been used to solve other dense prediction problems. The FCN-8s model is derived from the VGG 16-layer network [19] by discarding the final classifier layer and converting all fully connected layers to convolution layers. We attach an additional final 1×1 convolution layer with channel dimension 3 to obtain prediction scores for skin, hair and background. We use *score32* to represent the prediction score produced by the last convolution layer since the stride is of 32 pixels. In order to increase spatial accuracy, two skip layers are added to combine high level semantic information (*score32*) with shallow, fine appearance features (*pool3*, *pool4*). A 1×1 convolution layer is added on top of *pool4* to produce a fine-scale prediction *score-pool4*. Since *score-pool4* has twice the size of *score32*, to add two prediction scores, *score32* is upsampled by factor 2 using a deconvolu-

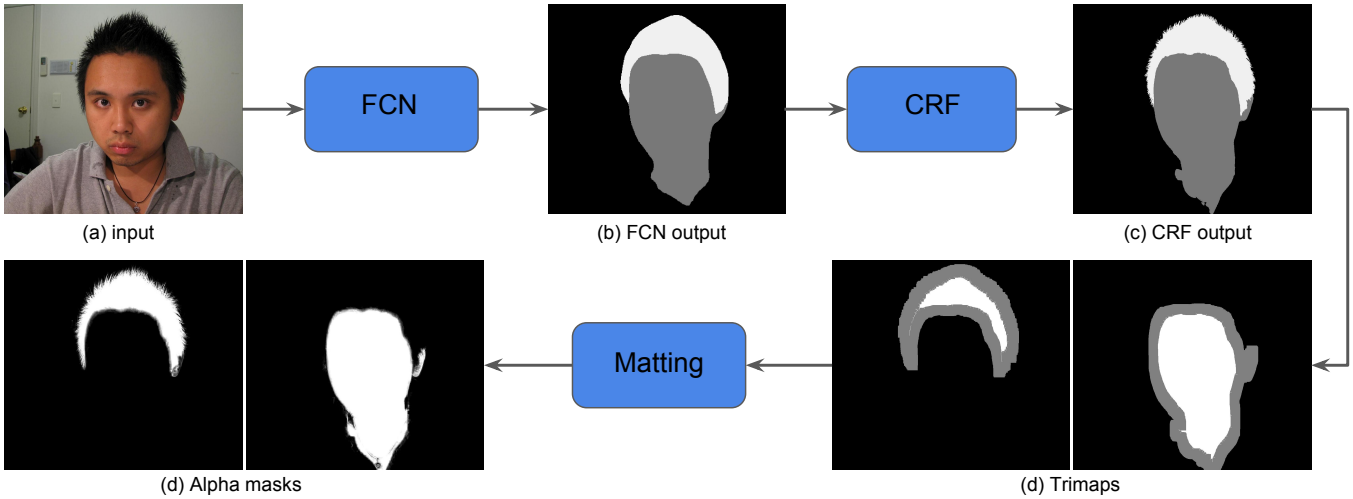


Fig. 2. Our framework. In the trimaps from (d), white pixels represent foreground (skin/hair), black pixels are background constrains, gray stands for unknown region.

tion layer. The parameters of the deconvolution layer is initialized by a bilinear interpolation kernel, but allowed to be learnt during training. The prediction score of *pool3* is fused in similar fashion. The final output prediction score of FCN-8s has stride 8; this is then upsampled using bilinear interpolation to ensure that the prediction has the same size as the input image. Softmax loss layer is used in training.

3.2. Fully-Connected Conditional Random Field

The output segmentation of FCN-8s network is robust in the face of variable skin and hair colors, hairstyles, and head poses. However, the achievable localization is relatively poor, see figure 2 (b). There are two main reasons for this. First , the output segmentation map from FCN-8s has stride 8, and thus the prediction resolution is much lower than that of the input image. Second, FCN takes a large range of context information into consideration, which produce homogeneous predictions. This is particularly vexing problem for hair segmentation, since hair usually contains very thin structures.

To overcome this drawback, we feed the label assignment probability map to a fully-connected CRF with the purpose of obtaining a finer segmentation result, as suggested by Chen *et al.* [11]. The model use the following energy function:

$$E(x) = \sum_i -\log P(x_i) + \sum_{ij} w_1 \exp(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}) + w_2 \exp(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}). \quad (1)$$

where $P(x_i)$ is the label assignment probability for the i-th pixel; the first kernel of pairwise energy depends on both pixel position p and pixel color value I; and the second kernel only depends on pixel positions. The hyper parameters θ_α, θ_β, θ_γ, w_1 and w_2 control the scale and weights of the Gaussian kernels, respectively. Pixels i and j are connected regardless of their distance, which makes the graph fully connected.

3.3. Image Matting

Although use of a fully-connected CRF significantly improves the localization of raw FCN output, the result is still not good enough for portrait editing. The first reason is that the fully-connected CRF still fails to recover very thin structures, especially for hair. Secondly, using the binary segmentation mask in editing can result in visible boundary effects. To overcome this issue, we use an image matting algorithm in order to obtain an accurate hair and skin alpha masks, with input trimaps automatically generated from the segmentation label map.

We apply morphological operators on the binary segmentation mask for hair and skin, obtaining a trimap that indicates foreground (hair/skin), background and unknown pixels. In order to deal with segmentation inaccuracies, and to best capture the appearance variance of both foreground and background, we first erode the binary mask with a small kernel, then extract the skeleton pixels as part of foreground constrain pixels. We also erode the binary mask with a larger kernel to get more foreground constrain pixels. The final foreground constrain pixels is the union of the two parts. If we only keep the second part then some thin hair regions will be gone after erosion with a large kernel. If a pixel is outside the dilated mask then we take it as background constrain pixel. All other pixels as marked as unknown, see figure 2 (d).

Finally, the automatically generated trimap is fed to an image matting algorithm to calculate an alpha mask. In our experiment, we use the matting algorithm from [20].

3.4. Application

Our algorithm for automatic generation of high quality skin and hair masks may be used in applications such as face skin manipulation [21, 22, 23], hair manipulation [24], or for creating facial and hairstyle databases for further processing [25].

As an example of application, we implemented a simple tool to manipulate hair, skin, and background using the masks produced by our algorithm. If only the skin or hair is to be processed, then one could just use the output alpha matte as described in Section 3.3. When both regions need to be processed, better rendition is obtained if pixels near skin and hair boundary are included in both masks. We propose to use a simple weighted average as follows:

$$w_p^k = \frac{m_p^k}{\sum_{k\in\{s,h,b\}} m_p^k} \quad (2)$$

where m_p^k is the value of the mask for label k at pixel p, and k could be skin (s), hair (h) or background (b). m^s and m^h are the output alpha masks. The background mask m^b is defined as $m^b = 255 - \min(255, m^s + m^h)$. With the soft masks, we can apply Bilateral Filtering [26] for skin smoothing and color manipulation on hair and skin regions. Two examples of processed images are shown in Figure 3. The results appear natural, without noticeable boundary effects.

Fig. 3. We change the hair color using the soft masks generated by our algorithm.

4. EXPERIMENTS

4.1. Implementation Details

The images used to train our networks come from the LFW Parts dataset training portion, which contains 2000 images with superpixel-level labels. Each image in the training set has size of 250×250 pixels, which is small compared to the receptive field size of FCN-8s. For this reason, we resize each training image to 500×500 pixels, and also add a copy of each image flipped horizontally for data augmentation.

Following the fine-tuning mechanism proposed in [10], we first fine-tune the FCN-32s (no skip layers) model using 4000 training images. The minibatch size is 1, learning rate is 10^{-9}, momentum is 0.99, weight decay is 0.0005. We then add one skip layer a time with reduced learning rate (10^{-12} and 10^{-13} respectively). The training process takes around 20 hours to complete. The inference time is around 90 ms for 500×500 input image.

Our system is implemented with Caffe and Matlab, and runs on a workstation (3.3Ghz 6-core CPU, 32G RAM, Nvidia GTX Titan X GPU and Ubuntu 14.04 64-bit OS).

4.2. LFW Parts Dataset

In this section, we evaluate the quality of the segmentation results on the LFW Parts dataset using fully convolutional networks and fully-connected CRF, as compared with previous methods. The use of the LFW Parts dataset, a subset of the widely used LFW dataset, was originally proposed by Kae *et al.* [8]. The dataset contains 2927 face images with size of 250×250 pixels, with large variance in background, hair and skin color, head poses and hairstyles. The training portion contains 1500 images, the validation portion contains 500 images, and the test portion contains 927 images. All images are manually labelled as skin, hair and background at the superpixel level.

In Table 1, we compare our results with those from previous work, including [8], [9] and other baselines proposed in [8]. The CRF model is implemented by Kae *et al.* [8] based on [15], Spatial CRF and CRBM are two more baselines used in [8]. For more details about these methods, the reader is referred to [8].

The results of [8] and all the baselines are evaluate in superpixel level, while Liu *et al.* [9] report performance at the pixel level. In order to allow for a full comparison, we report performance at both the pixel and the superpixel levels. Since our system produces only pixel-level classification, we label each superpixel with the most frequent label across the pixels in the superpixel.

Table 1 shows that our method outperforms the other considered algorithms by a noticeable margin. Comparing with [9] using the same pixel-level evaluation mechanism, our raw FCN output acieves about 1% accuracy improvement, while the additional CRF processing gives a 0.42% boost. At superpixel level, our method produces results with accuracy that is more than 2.3% than the algorithm by Kae *et al.* [8]. The absolute accuracy value increase may not seems so significant when the baseline accuracy is already pretty high, the error reduction rate is much more convincing. It is interest-

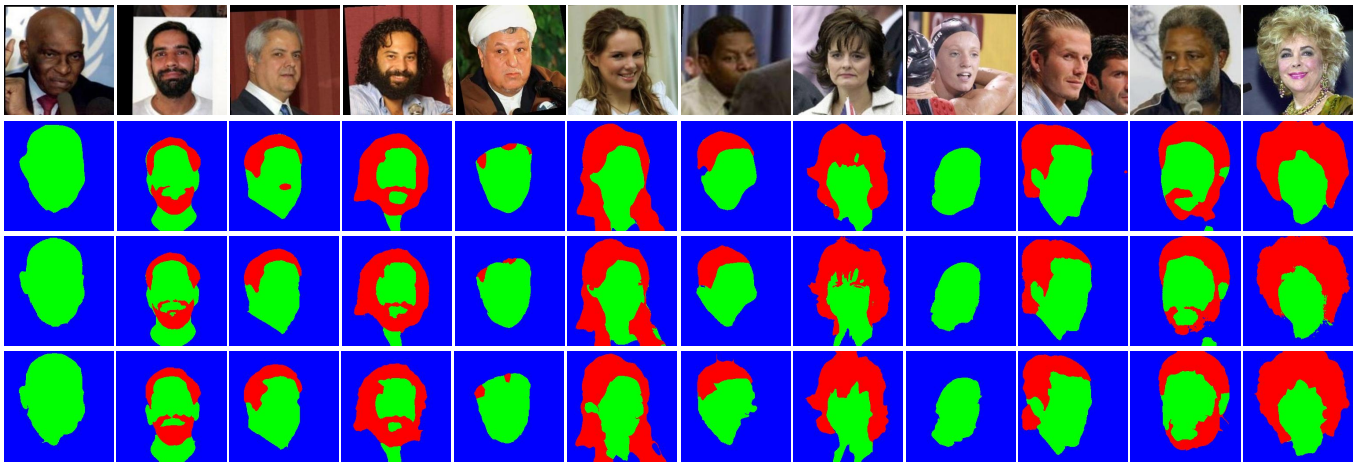


Fig. 4. Sample segmentation results on the LFW Parts dataset. From top to bottom: a) Input image, b) Raw FCN output, c) FCN+CRF, d) Superpixel level ground truth labeling. The results shows that our proposed method is robust against challenging situations such as occlusion, complex hairstyle, no hair, beard as well as variance head poses. For several samples like the 5th, 7th, 8th and 11th image, our result appears to be more accurate than the ground truth. This is due to the fact that the ground truth is provided at the superpixel level (which may not accurately represent all fine details), while our algorithm works at the pixel level.

ing to note, that, while the fully-connected CRF allowed for a small boost in accuracy at pixel level, no significant difference could be registered at the superpixel level. This should not come as a surprise, as the increase in localization accuracy at pixel level may simply be lost when considering whole superpixels. Some examples of results are shown in Figure 4.

Table 1. Overall accuracy on LFW Parts dataset compared to [8], [9] and other baselines. Note that [8] and [9] are evaluated at the superpixel (SP) and pixel level, respectively. We provide results of our system at both the superpixel and pixel level.

Method	Accuracy	Error Reduction
CRF	93.23%	-
Spatial CRF	93.95%	10.64%
CRBM	94.10%	12.85%
GLOC [8]	94.95%	25.41%
MO-GC [9]	95.24%	29.69%
Ours Pixel (FCN)	96.34%	45.94%
Ours Pixel (FCN+CRF)	**96.76**%	**52.14**%
Ours SP (FCN)	97.30%	60.12%
Ours SP (FCN+CRF)	**97.32**%	**60.41**%

The image matting step, producing a soft and more accurate mask, is crucial for skin and hair editing. We show this in Figure 5, which compares editing results with and without image matting step. Use of image matting allows for more natural results, thanks to the soft transition between different areas and more accurate masks, especially for hair area.

Fig. 5. Comparison of editing results with and without image matting step. Note how, by softening the transition between skin and hair, image matting produces more natural results.

5. CONCLUSION

We presented a system for accurate pixel-level segmentation of hair and skin for photo editing applications. We achieved state-of-the-art results on the LFW Parts data set using a fully convolutional networks, followed by a fully-connected conditional random field. In addition, we showed how this segmentation can be used to generate trimaps automatically. These trimaps can then be used as input to an image matting algo-

rithm for "soft" image rendering. Our system can find application in image enhancement and beautification and in portrait editing.

6. REFERENCES

[1] http://vis-www.cs.umass.edu/lfw/part_labels/.

[2] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM transactions on graphics (TOG)*. ACM, 2004, vol. 23, pp. 309–314.

[3] Vladlen Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[4] Anat Levin, Dani Lischinski, and Yair Weiss, "A closed-form solution to natural image matting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 228–242, 2008.

[5] Yaser Yacoob and Larry S Davis, "Detection and analysis of hair," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 7, pp. 1164–1169, 2006.

[6] Kuang-chih Lee, Dragomir Anguelov, Baris Sumengen, and Salih Burak Gokturk, "Markov random field models for hair and face segmentation," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–6.

[7] Cedric Rousset and Pierre-Yves Coulon, "Frequential and color analysis for hair mask segmentation," in *2008 15th IEEE International Conference on Image Processing*. IEEE, 2008, pp. 2276–2279.

[8] Andrew Kae, Kihyuk Sohn, Honglak Lee, and Erik Learned-Miller, "Augmenting crfs with boltzmann machine shape priors for image labeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2019–2026.

[9] Sifei Liu, Jimei Yang, Chang Huang, and Ming-Hsuan Yang, "Multi-objective convolutional learning for face labeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3451–3459.

[10] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[12] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[13] Saining Xie and Zhuowen Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1395–1403.

[14] Iasonas Kokkinos, "Pushing the boundaries of boundary detection using deep learning," *arXiv preprint arXiv:1511.07386*, 2015.

[15] Gary B Huang, Manjunath Narayana, and Erik Learned-Miller, "Towards unconstrained face recognition," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. IEEE, 2008, pp. 1–8.

[16] David R Martin, Charless C Fowlkes, and Jitendra Malik, "Learning to detect natural image boundaries using brightness and texture," in *Advances in Neural Information Processing Systems*, 2002, pp. 1255–1262.

[17] Nan Wang, Haizhou Ai, and Shihong Lao, "A compositional exemplar-based model for hair segmentation," in *Asian Conference on Computer Vision*. Springer, 2010, pp. 171–184.

[18] Nan Wang, Haizhou Ai, and Feng Tang, "What are good parts for hair shape modeling?," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 662–669.

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

[20] Ehsan Shahrian, Deepu Rajan, Brian Price, and Scott Cohen, "Improving image matting using comprehensive sampling sets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 636–643.

[21] Xiaowu Chen, Hongyu Wu, Xin Jin, and Qinping Zhao, "Face illumination manipulation using a single reference image by adaptive layer decomposition," *IEEE Transactions on Image Processing*, vol. 22, no. 11, pp. 4249–4259, 2013.

[22] Ai Mizokawa, Hiroki Nakai, Akinobu Maejima, and Shigeo Morishima, "Photorealistic aged face image synthesis by wrinkles manipulation," in *ACM SIGGRAPH 2013 Posters*. ACM, 2013, p. 64.

[23] Lingyun Wen, GD Guo, et al., "Dual attributes for face verification robust to facial cosmetics," *J. of Comput. Vision and Image Process*, vol. 3, no. 1, pp. 63–73, 2013.

[24] Menglei Chai, Lvdi Wang, Yanlin Weng, Yizhou Yu, Baining Guo, and Kun Zhou, "Single-view hair modeling for portrait manipulation," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, pp. 116, 2012.

[25] Liwen Hu, Chongyang Ma, Linjie Luo, and Hao Li, "Single-view hair modeling using a hairstyle database," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, pp. 125, 2015.

[26] Carlo Tomasi and Roberto Manduchi, "Bilateral filtering for gray and color images," in *Computer Vision, 1998. Sixth International Conference on*. IEEE, 1998, pp. 839–846.