

Semi-supervised Skin Detection by Network with Mutual Guidance

Yi He^{1*} Jiayuan Shi^{2*} Chuan Wang^{2†} Haibin Huang² Jiaming Liu²
Guanbin Li³ Risheng Liu¹ Jue Wang²

¹Dalian University of Technology, {heyi@mail., rslu@}dlut.edu.cn

²Megvii Technology, {shijiayuan, wangchuan, huanghaibin, liujiaming, wangjue}@megvii.com

³Sun Yat-sen University, liguanbin@mail.sysu.edu.cn

Abstract

In this paper we present a new data-driven method for robust skin detection from a single human portrait image. Unlike previous methods, we incorporate human body as a weak semantic guidance into this task, considering acquiring large-scale of human labeled skin data is commonly expensive and time-consuming. To be specific, we propose a dual-task neural network for joint detection of skin and body via a semi-supervised learning strategy. The dual-task network contains a shared encoder but two decoders for skin and body separately. For each decoder, its output also serves as a guidance for its counterpart, making both decoders mutually guided. Extensive experiments were conducted to demonstrate the effectiveness of our network with mutual guidance, and experimental results show our network outperforms the state-of-the-art in skin detection.

1. Introduction

Skin detection is the process of finding skin-colored pixels and regions from images and videos. It is a very interesting problem and typically serves as a pre-processing step for further applications like face detection, gestures detection, semantic filtering of web contents and so on [5, 26, 2, 15, 10].

Skin detection has been proven quite challenging with large variation in skin appearance, depending on its own color proprieties and illumination conditions. Previous methods like [9, 42] tried to model skin color in different color spaces and train skin classifiers in those spaces. However, these methods heavily rely on the distribution of skin color, and with no semantic information involved, they suffer from a limited performance. In recent years, with the development of deep neural networks, skin detection methods have been proposed by adaption of networks used for other



Figure 1. Skin detection results by our approach vs. solutions of a UNet and a tradition Gaussian Mixture Model (GMM). The intersection-over-union (IoU) rates demonstrate our approach has a better performance.

detection tasks [13, 4, 28]. Although these DNN based skin detection methods reveal promising accuracy improvements, they are still limited by annotated skin data which is expensive and time-consuming to collect.

To this end, we propose to improve skin detection by introducing body detection as a guidance. If a body mask is available, it could potentially facilitate the skin detection in two-folds. First, it provides a prior information for a skin detector where higher probability of skin is located. Second, after a skin mask is detected, it can filtered out the false positive pixels in the background. Meanwhile, with skin mask as a guidance, a body detector is also provided with more information. To enable the mutual guidance scheme, we designed a dual-task neural network for jointly detection of skin and body. The entire network contains a shared encoder but two decoders for skin and body detection sep-

*Equal Contributors

†Corresponding Author

arately. The output from each decoder would be fed to the other one so as to form a recurrent loop as shown in Figure 2(a). The shared encoder of the two detectors would extract common feature maps from the input image, considering the similarity of the two tasks and the compactness of the network. This structure enables us to train the skin detection network without increasing the annotated training data but simply adding a human body mask dataset, which is rather easier to obtain. Since the two datasets contain two types of ground truth separately, i.e. a data sample has either a target skin mask or body mask, we train the network in a semi-supervised manner with a newly designed loss and a customized training strategy. Experimental results demonstrate the effectiveness of all the newly involved techniques for our network, and qualitative, quantitative evaluations also show that our network outperforms state-of-the-art methods as shown in Figure 1, 5 and Table 1 for skin detection task. We also build a new dataset composed of 5,000 annotated skin masks and 5,711 annotated body masks which can be released for future research upon the acceptance of this paper.

To summarize, our main contributions are:

- A novel uniform dual-task neural network with mutual guidance, for joint detection of skin and body which can boost the performance of both tasks, especially for skin.
- A newly designed loss and customized training strategy within a semi-supervised fashion, which performs well in the case of missing ground truth information.
- A new dataset containing skin and body annotated masks to demonstrate the effectiveness of our network, and facilitate the future research in community.

2. Related Work

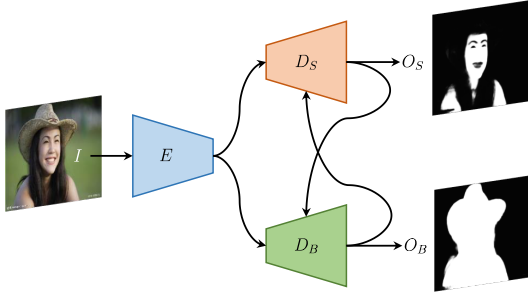
Skin detection and segmentation. Skin detection has been studied in the past two decades. Existing methods can be grouped into three categories, i.e. defining boundary models explicitly on color spaces [18, 9, 23, 24, 30] (thresholding), applying traditional machine learning techniques to learn a skin color model [20, 40, 42], and using a deep neural network to learn an end-to-end model for skin segmentation [1, 38, 18, 29, 3]. The thresholding methods focus on defining a specified region in color spaces like RGB, YCbCr, HSV so that a pixel falls in the regions is considered to be skin. However, there is a significant overlap between the skin and non-skin pixels in color space, for example numerous objects in the background such as wall, cloth could also have similar color. Traditional machine learning techniques further involve generative and discriminative models to predict the probability of a pixel belonging to skin, which may also take local features like texture into

consideration. Even though, these models commonly suffer from low accuracy due to their limited learning capabilities. Early neural network based approaches usually applied Multi-Layer Perceptrons (MLP) whose classification accuracy are still limited. In recent years, fully convolutional neural network (FCN) is widely applied in image segmentation tasks [21], hence skin detection naturally becomes an application of it [43]. However, the FCN based segmentation usually require large-scale of strong supervision in training stage, which restricts that a high-quality model can be easily trained. In [32], a conditional random field is involved as a loss for the end-to-end image segmentation task, which enables the use of weakly supervised data. Unlike these methods, our approach can take advantage of an extra dataset of body segmentation, which is commonly easier to acquire, to boost the performance of a CNN based skin detector.

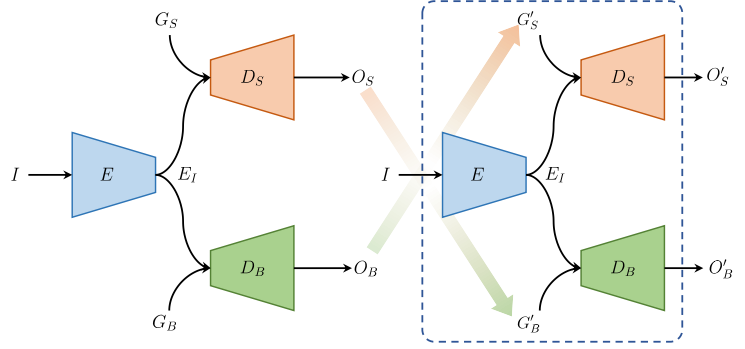
Multi-task joint learning. Multi-task learning (MTL) has been used successfully across all applications of machine learning, from natural language processing [6] and speech recognition [7] to computer vision [11]. It is generally applied by sharing the hidden layers between all tasks, while keeping several task-specific output layers as branches. Some multi-task networks generally learn common feature maps via shared encoders, so as to potentially improve the performance of all tasks simultaneously. For example, [17] utilized a three-branch network to solve semantic segmentation, instance segmentation and depth prediction in a unified framework. There are more multi-task networks which exist for solving a complicated task, where all the outputs of the task-specified networks are fused for further processing. For example, [35] proposed a network containing two sub-networks that jointly learning spatial details and temporal coherence for a video inpainting task. In [12], Han et al. decompose a shape completion task into two sub-tasks, to reconstruct global and local structure respectively and then fuse together. These methods commonly involve guidance from one branch to another to reduce the learning difficulty. Our approach follows a similar idea, while the two branches of the network can mutually guide the other, so as to boost the performance of skin detection via the recurrent loop in the network.

3. Algorithm

Our method is built upon dual-task fully convolutional neural networks. It takes a single RGB image I as input, and produces probability maps of skin O_S and body O_B as outputs. The network contains two decoders D_S, D_B in separate branches, for the detection tasks of skin and body respectively. The two decoders share a common encoder E , which extracts the feature map of I as E_I . The output O_S , together with E_I , are fed to the decoder of body D_B



(a) Our dual-task network with mutual guidance



(b) Our network decoupled in two stages, for ease of analyzing

Figure 2. Structure of our dual-task network with mutual guidance. (a) The original network structure with mutual guidance loop. (b) The decoupled network into two stages for ease of analyzing.

in the other branch, and vice versa. For either decoder, the output from the other branch serves as a guidance for the task, making the dual tasks mutually guided. The network structure is illustrated at Figure 2(a).

3.1. Network with Mutual Guidance

Our network is a dual-task network with mutual guidance, which can be viewed as a recurrent network due to the structure containing signal loop. For ease of analysis, we decoupled the original network into two stages with no loop as shown in Figure 2(b). To differentiate the symbols in the two stages, we use X for Stage 1 and X' for Stage 2 accordingly. And for brevity, we use $\kappa \in \{S, B\}$ to represent a module or variable of *skin* or *body*. Here *skin* refers to pixels of entire body skin area, and *body* is a super-set of *skin* that also includes pixels of hair, cloth, etc. A set of $\{X_\kappa\}$ represents both X_S and X_B , so that $\{D_\kappa\}$ means D_S and D_B as an example.

In Stage 1, we feed decoders $\{D_\kappa\}$ with guidances $\{G_\kappa\}$ and produce outputs $\{O_\kappa\}$ as intermediate results. Then we feed the decoders with $\{G'_\kappa\}$ in Stage 2 and produce the final outputs $\{O'_\kappa\}$. For two stages, the input I and the weights in E and $\{D_\kappa\}$ are identical, while guidances in two stages are commonly various, i.e. $G_\kappa \neq G'_\kappa$ for $\kappa \in \{S, B\}$. That is because in Stage 1, commonly we have very limited or even no information to provide, while in Stage 2 we have the initial results $\{O_\kappa\}$ detected to serve as guidances. Moreover, here we design a shared encoder E instead of two independent ones, not only for reducing redundancy, but also based on the following two considerations. First, even though the training data for the two tasks have different ground truth, the input RGB images share very similar statistics. Second, there also exist some common properties for the extracted feature map that are desirable for the two tasks, such as robustness to distinguish human foreground and non-human background. Experimental results demonstrate this shared encoder could improve

the performance of skin detection by seeing more data and learning the common features, as shown in Table 1 and Figure 4. In summary, the entire network can be written as follows.

- Stage 1 $\begin{cases} G_S = e_B, & G_B = e_S \\ O_S = D_S(E_I, G_S), & O_B = D_B(E_I, G_B) \end{cases}$
- Stage 2 $\begin{cases} G'_S = O_B, & G'_B = O_S \\ O'_S = D_S(E_I, G'_S), & O'_B = D_B(E_I, G'_B) \end{cases}$

where e_S and e_B are the signals provided as guidances in Stage 1, which are commonly set to 0 in most cases in this paper. For the structures of E and $\{D_\kappa\}$, we adapted the standard UNet [28] architecture including 4 downsampling blocks in E and 4 upsampling blocks in D_κ . The size of input I is $512^2 \times 3$ so that the feature maps between E and D_κ , i.e. E_I is of size $32^2 \times 1024$. We also applied an encoder of the same structure as E but of half number of channels for each layer to the guidance $\{G_\kappa\}$, to ensure its extracted feature can be well concatenated to E_I , after they are fed to D_κ . For each fully convolutional layer, the kernel size is set to 3×3 , and is followed by a BatchNorm and a ReLU layer.

With the initial results $\{O_\kappa\}$ detected, the decoders are provided with more informative guidances, that are helpful for the second stage detection.

3.2. Learning Algorithm

The goal of our learning algorithm is to train a dual-task CNN which can detect skin and body end-to-end, which is far from straightforward. On one hand, for skin detection task, lacking enough training data is a common issue, and human labelling is usually very expensive and time-consuming. On the other hand, for body detection, due to the extensive research in recent years, its data is relatively easier to obtain. So in our problem settings, for each data pair, it contains ground truth mask of skin or body only,

noted as M_S or M_B . Since there is few such training data triple (I, M_S, M_B) provided, it naturally makes training our network a semi-supervised task, which is achieved by a semi-supervised loss we design and several training details we adopt.

3.2.1 Semi-supervised loss

Our newly designed semi-supervised loss consists of three parts, including strongly-supervised and weakly-supervised ones. The former one is the cross-entropy loss between the output and the ground truth; and the latter ones include CRF loss and a weighted cross-entropy (WCE) loss between skin output and body output.

Cross-entropy loss. As aforementioned, the training data provided to our problem, is a data pair with either skin or body ground truth. For a data sample with M_S , we compute the cross entropy losses between M_S and its outputs O_S, O'_S respectively, making them strong supervision to the skin detection task. Similarly, it also applies to data sample with M_B , so that we produce a sum of four terms of cross-entropy losses:

$$\mathbf{L}_{ce} = \sum_{\kappa \in \{S, B\}} \sum_{x \in \{O_\kappa, O'_\kappa\}} l_\kappa \cdot L_{ce}(x, M_\kappa) \quad (1)$$

where $L_{ce}(x, y) = x \cdot \log(y) + (1 - x) \cdot \log(1 - y)$. Here we use a label notation l_κ to present whether the current data sample has a ground truth M_κ . For example, if a data sample has M_S only, then $l_S = 1, l_B = 0$, and vice versa. l_κ works as a switch for enabling the contribution of a loss or not. This notation also applies for the rest of this paper.

CRF loss. For a data sample with a single type of ground truth, one of its outputs can contribute to the cross-entropy loss yet the other one cannot. For this case, we involve a CRF loss as in [33]. By computing a CRF given image I and a mask O_κ , CRF loss can constrain neighboring pixels in I with similar color tend to have a consistent label in O_κ . In most cases when strong supervision is unavailable, this property could potentially refine the output mask. Similarly, the total CRF loss can be written as

$$\mathbf{L}_{crf} = \sum_{\kappa \in \{S, B\}} \sum_{x \in \{O_\kappa, O'_\kappa\}} (1 - l_\kappa) \cdot L_{crf}(x, I) \quad (2)$$

where $L_{crf} = S^T W S$ where W is an affinity matrix of I and S is a column vector of flattened O_κ . We refer the readers to [33] for more details about CRF loss.

WCE loss. It is also a prior knowledge that the skin mask should be covered by its body mask for the same image. The consistency should be preserved in the outputs O_S, O'_S

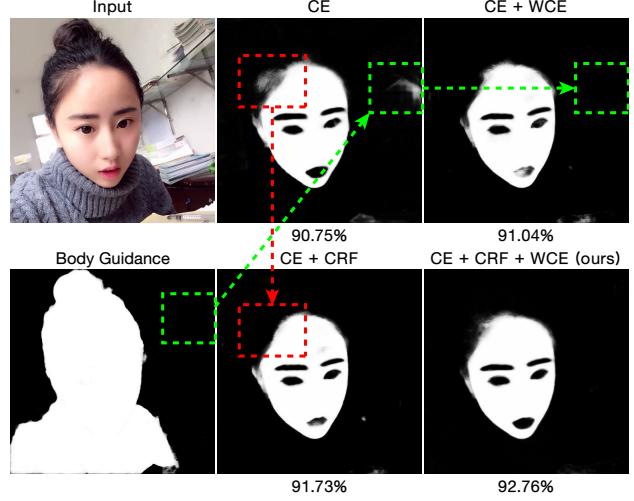


Figure 3. Weakly-supervised losses improve the detection result. Red box and arrow: with CRF loss involved, the region between hair and head tends to be classified with the same label, causing the hair region removed. Green boxes and arrows: with body guidance and WCE loss involved, the false alarm region, i.e. the background is removed. The IoUs are listed under the corresponding results. (Best view in color)

and O_B, O'_B . For a pixel classified with high probability of skin, it should also have a high probability of body. This does not hold if a lower probability of skin is detected, because the pixel may belong to non-skin regions as cloth or hair, where body probability is still high. To characterize the above relationship, we compute a cross-entropy loss between skin and body probability, then weight it by the skin probability itself, i.e. $L_{wce}(x, y) = x \cdot L_{ce}(x, y)$, where $x \in \{O_S, O'_S\}, y \in \{O_B, O'_B\}$. As a result, the total WCE loss is calculated as

$$\mathbf{L}_{wce} = \sum_{x \in \{O_S, O'_S\}, y \in \{O_B, O'_B\}} L_{wce}(x, y) \quad (3)$$

The CRF and WCE are two weakly-supervised losses. Compared with cross-entropy loss as strong supervision, they weakly take effect for the tasks of skin and body detection, which finally improve the performance. To sum up, our semi-supervised loss is

$$\mathbf{L} = \mathbf{L}_{ce} + \lambda_1 \cdot \mathbf{L}_{crf} + \lambda_2 \cdot \mathbf{L}_{wce} \quad (4)$$

where λ_1 and λ_2 are the balancing hyper-parameters. We set λ_1 to 0.0001 and λ_2 to 0.001 in our experiments. Figure 3 illustrates an example to reveal the effectiveness of CRF and WCE losses, and more discussion is involved in Section 4.3.2.

3.2.2 Training details

Dual-task joint learning. Our network is trained by the Adam Optimizer, where each branch is handled exclusively

	IoU (%)	IoU Top-1 (%)	Precision (%)	Recall (%)
Thresholding [18]	50.84 / 60.20	1.06 / 0.00	59.30 / 65.31	81.75 / 89.58
GMM [16]	50.06 / 60.46	2.34 / 0.00	53.45 / 62.36	89.31 / 91.50
Chen et. al's [3]	55.77 / 62.05	0.43 / 3.12	74.31 / 72.50	70.94 / 79.18
Zuo et. al's [43]	69.94 / 79.81	0.21 / 0.00	84.38 / 88.97	80.31 / 88.03
UNet [28]	75.59 / 85.50	15.53 / 28.13	89.38 / 93.42	83.14 / 90.91
ResNet50 [13]	75.44 / 84.33	11.49 / 12.50	88.77 / 92.19	82.97 / 90.72
Deeplab-v3-ResNet50 [4]	75.97 / 85.88	10.64 / 6.25	86.98 / 92.51	85.58 / 92.48
Deeplab-v3-MobileNet [4]	73.66 / 83.96	7.02 / 9.38	87.16 / 91.91	82.48 / 90.48
Ours	81.18 / 87.90	51.27 / 40.63	90.01 / 95.23	89.01 / 92.08

	IoU (%)	IoU Top-1 (%)	Precision (%)	Recall (%)
Thresholding [18]	50.84 / 60.20	3.19 / 0.00	59.30 / 65.31	81.75 / 89.58
GMM [16]	50.06 / 60.46	5.74 / 6.25	53.45 / 62.36	89.31 / 91.50
Chen et. al's [3]	51.44 / 62.43	1.28 / 6.25	76.11 / 76.36	63.18 / 77.89
Zuo et. al's [43]	63.98 / 73.91	0.85 / 0.00	81.28 / 85.19	74.99 / 82.88
UNet [28]	69.62 / 79.62	16.81 / 18.75	83.96 / 89.55	80.61 / 87.87
ResNet50 [13]	66.03 / 77.97	7.66 / 3.12	84.73 / 88.30	74.82 / 86.87
Deeplab-v3-ResNet50 [4]	69.04 / 76.63	12.34 / 12.50	81.81 / 86.19	81.34 / 87.39
Deeplab-v3-MobileNet [4]	67.95 / 77.63	6.60 / 9.38	81.92 / 86.93	79.90 / 87.59
Ours	75.29 / 81.89	45.53 / 43.75	87.34 / 92.58	84.64 / 87.51

Table 1. Evaluated IoU, IoU Top-1 rate, precision and recall on our validation dataset (black) and Pratheepan Face dataset (blue), trained by balanced dataset ($\#skin, \#body = 5k$) (top) and unbalanced dataset ($\#skin = 1k, \#body = 5k$) (bottom).

in each iteration, while they are jointly learned for the dual-task. For even and odd iteration, we feed data samples with M_S and M_B respectively, i.e. $(I, M_S, M_B = 0, l_S = 1, l_B = 0)$ or $(I, M_S = 0, M_B, l_S = 0, l_B = 1)$. Given each data sample, thanks to the existence of label l_κ ($\kappa \in \{S, B\}$), its cross-entropy loss is computed in one branch and CRF loss is done in the other. With the training going on, the outputs from Stage 1 $\{O_\kappa\}$, gradually provide guidance for Stage 2. Meanwhile, with the increasingly informative guidances from $\{O_\kappa\}$, the detection difficulty for decoders in Stage 2 is reduced, so that the final outputs $\{O'_\kappa\}$ are expected to become increasingly accurate.

Finetune. To develop the potential of the dual-task network with mutual guidance, care must be taken during training. In practice, we first train the Stage 1 network by involving losses on $\{O_\kappa\}$ only. Due to the lack of guidance at present stage, we feed the network with $G_\kappa = E_\kappa = 0, \kappa \in \{S, B\}$ instead. With the convergence of the network, the outputs $\{O_\kappa\}$ tend to become informative but still of limited accuracy.

We further involve training in the 2nd stage. We feed the decoders $\{D_\kappa\}$ with $\{G'_\kappa\}$, where G'_κ is obtained with the following manner. For a data sample with $M_S \neq 0$ and $l_S = 1, G'_B$ is set to M_S ; otherwise, $G'_B = O_S$. Similar rules also apply to G'_S , i.e.

$$\begin{cases} G'_B = l_S \cdot M_S + (1 - l_S) \cdot O_S \\ G'_S = l_B \cdot M_B + (1 - l_B) \cdot O_B \end{cases}$$

This strategy ensures that, we feed the most trusted data as guidance to the decoders, to avoid misleading with incorrect guidance data, especially if O_κ is of low quality. Furthermore, due to the large variation between guidances in the two stages, i.e. $G_\kappa \neq G'_\kappa$, the decoders of sharing weights in the two stages have to own the power to

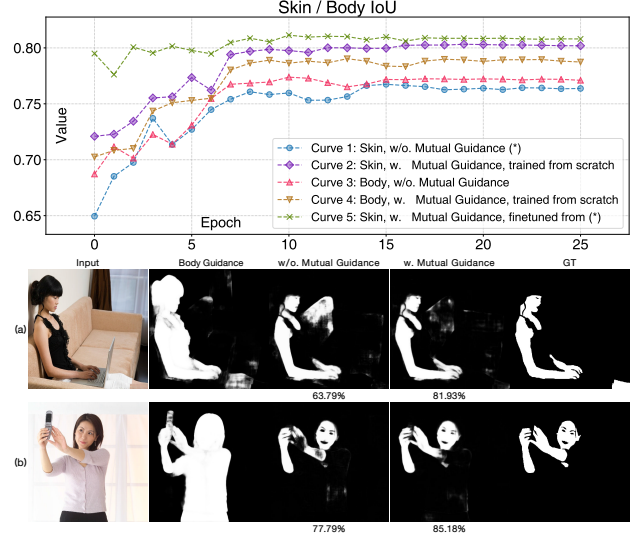


Figure 4. Mutual guidance. Top: Curve 1 \sim 4: IoU for detected skin and body mask by our dual-task network, trained with or without mutual guidance, with respect to the number of epoches. Curve 5: detected skin IoU by our network in finetuned version. Bottom: two examples showing the detected masks by our network without (column 3) or with (column 4) mutual guidance. The body masks serving as guidances are shown in column 2.

regress the same data sample with various guidance, i.e. (I, G_S, G_B) and (I, G'_S, G'_B) , to the unique ground truth M_S (if $l_S = 1$) or M_B (if $l_B = 1$). To achieve it, we apply a gradient stopping scheme to disable the back-propagation from G'_B, G'_S to their corresponding decoders in Stage 1, so as to avoid the outputs $\{O_\kappa\}$ tending to trivially regress to values like $\{E_\kappa\}$. Meanwhile, the semi-supervised loss additionally involves the ones computed with $\{O'_\kappa\}$. With the training keeping on, the decoders gradually obtain the tolerance to handle various guidances in the two stages, while with informative guidance they can perform better. We demonstrate the effectiveness of the two-stage training strategy, mutual guidance and gradient stopping scheme in Section 4.3.

4. Experimental Results

4.1. Dataset and Implementation Details

Our dataset is composed of 10,711 RGB images, 5,000 of which have human-annotated skin masks M_S ($l_B = 0, l_S = 1$) and the rest have body masks M_B ($l_S = 0, l_B = 1$), noted as \mathcal{D}_S and \mathcal{D}_B . The original RGB images are collected from the Internet, and we resized them into 512^2 resolution. We randomly selected 470 samples from \mathcal{D}_S and 475 ones from \mathcal{D}_B , to establish two validation datasets. During training, we augmented the training data by randomly flipping, resizing and cropping the original data samples to ensure data diversity. Our code was developed with



Figure 5. Typical skin detection results on our validation dataset, by various methods including thresholding, GMM, Chen et al’s [3], Zuo et al’s [43], UNet, ResNet50, DeepLab-v3 with ResNet50 as backbone, DeepLab-v3 with MobileNet as backbone, ours (Column 2 to 10). Input and ground truth are shown in column 1 and 11.

TensorFlow, and the whole training was completed in about 12 hours by one NVIDIA GeForce GTX 1080Ti GPU. We will release our dataset to the public upon the acceptance of this paper.

4.2. Comparison with Existing Methods

We compared our method with some state-of-the-art ones, including two traditional algorithms and six NN based methods. [18] is a pixel value thresholding method which establishes some rules on pixel RGB and HSV color to classify a pixel to skin or not, rather than a soft probability map. Gaussian Mixture Model (GMM) [16] based method improves the mechanism, where a skin color GMM is learned, given an initial skin mask. The learned GMM then predicts the skin probability for each pixel. The problems behind the

two traditional methods are that they lack high level features involved in the detection task, and they are far from robustness to light change or complex background. The other six NN based methods are end-to-end, producing a skin probability map given an RGB image, where the differences lie in the structures of networks only, i.e. Chen et al’s [3], Zuo et al’s [43], U-Net [28], ResNet50 [13] and DeepLab-v3 [4] with ResNet50 or MobileNet as backbones.

We trained the six networks to convergence with multiple trials with dataset \mathcal{D}_S , and selected their best results. To quantitatively compare our method with them, we evaluated precision, recall, and intersection-over-union (IoU) of all the results, and list them at Table 1. The data shows that in terms of IoU and precision, our approach outperforms the state-of-the-art for skin detection. For recall, our method

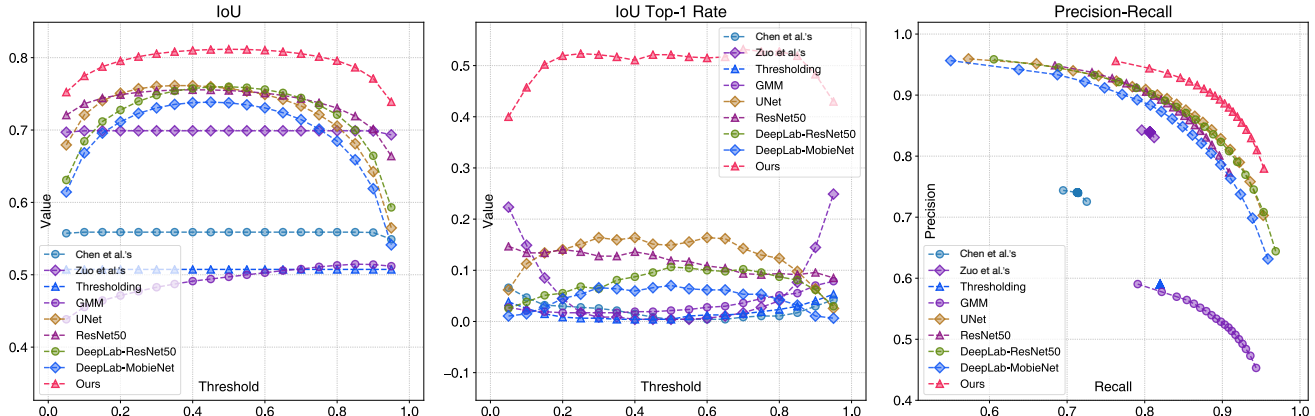


Figure 6. Curves of IoU, IoU Top-1 rate with respect to the probability threshold, and Precision-Recall on our dataset.

ranks only below the GMM method, which has more false alarms so as to suffer from a poor precision. For the mean IoU of all the validation data samples, our method is about 4% higher than the 2nd competitor in average. Even though, we further evaluates the robustness of our method by calculating an IoU Top-1 rate, i.e. what percentage of data each method can win the competition in terms of IoU. We found our method wins for nearly 51% validation data and none of the others has a comparable performance. We illustrate the curves of IoU, IoU Top-1 rate and precision-recall at Figure 6. We also compared our network with four CNNs on the public dataset of Pratheepan Face [30], and results also show that our method outperforms the others in Table 1 (blue values).

We list several typical detected skin masks in Figure 5 and 1 for qualitative comparison, where the examples cover various skin colors, complex illuminance, white balance, similar color in background especially cloth etc. They are captured in various conditions, by casual cameras or in studio. Figure 5(a) is a man of black skin wearing a navy suit, and (h) captures an asian girl wearing a camouflage suit with spots in skin color, making their skins so hard to distinguish; (b) contains white background light around the woman’s naked back and arm, and (g) is in a warm color style; (d)(e)(f) contain multiple people in various poses, especially in (d) three people (2 close, 1 far away, with various scales) exist in a yellow lighting condition. (c) shows a woman holding a phone which has reflectance and occludes part of her arm, making the visible skin spatially discontinuous. These challenging conditions make other methods fail or perform poorly, for example traditional thresholding or GMM method totally fail in (g)(h) and the end-to-end CNN methods work unstable in (a)~(f). In contrast, our approach overcame the difficulties as stated above and produced accurate and robust results, especially in Figure 5(d), where the man in the distance looks too tiny to be visible for human eyes.

4.3. Ablation Studies

4.3.1 Mutual guidance

We further reveal the effectiveness of mutual guidance scheme by experiments with or without it, both trained from scratch for fair comparison. By disabling the mutual guidance, i.e. training the proposed dual-task network in Stage 1 only, we plot the IoU of skin and body in the validation dataset for every epoch until convergence, as illustrated in Figure 4. From it, we can see that with mutual guidance involved, the IoU for both skin and body can be raised to a higher value at the same epoch, compared with the case of mutual guidance excluded. Note that, even for the case without mutual guidance, our network achieves 76.74% IoU, still higher than state-of-the-art single-task CNN solutions as shown in Table 1 (top, black values). It is due to the structure of our network with shared encoder E , which enables the learning from the extra body data. We also show two skin detection results by the two methods in Figure 4(a)(b) for a visual comparison. In both examples, the network without mutual guidance produced results with lower IoU due to false alarm (sofa in (a)) or mis-detection (hand in (b)). With body guidance involved, performance is improved with the false positive pixels and mis-detected pixels being corrected.

4.3.2 Weakly supervised losses

We also demonstrate the effectiveness of the weakly supervised losses we introduce, by disabling either one of them or both of them. We found that although these two losses contribute insignificantly compared with the strong supervised cross-entropy loss, they indeed take effects proven by the fact that each one raises up IoU by approximate 0.25%, and both can raise up to 1.9%, as shown in Table 2 Top. Figure 3 illustrates an example where if neither of the CRF and WCE losses is involved, there exists some misclassified

CE (Strong)	CRF	WCE	IoU (%)
✓			79.28
✓	✓		79.48
✓		✓	79.52
✓	✓	✓	81.18

	Ours (DeepLab-v3 -MobileNet) (%)	Ours (UNet) (%)	IoU Gain by Method (%)
w/o. M.G.	75.15	76.56	↑ 1.41
w. M.G.	79.02	80.11	↑ 1.09
IoU Gain by M.G.	↑ 3.87	↑ 3.55	–

Table 2. Performance for various compositions of losses (top) and different backbone networks (bottom). M.G. is the abbreviation of Mutual Guidance.

background pixels. In this case, WCE takes effects because the detected body mask supervises the region to be classified as background. Meanwhile, CRF loss weakly supervises the region between the hair and head to have a consistent labeling, causing the hair pixels filtered out. With both losses enabled, the final IoU outperforms the CE-loss-only version by 2% .

4.3.3 Unbalanced dataset

We also conducted a comparison on unbalanced dataset. In this experiment, we extracted only $1k$ skin samples from \mathcal{D}_S , together with the $5k$ body samples \mathcal{D}_B for training. We also list the IoU, IoU Top-1, Precision and Recall in Table 1 (bottom). Compared with the results trained by balanced dataset, the IoU value drops about 6% for our method but is still obviously higher than the others. We also applied this experiment to Pratheepan Face dataset [31] and similar conclusion was drawn.

4.3.4 Backbone networks

We also explore the influence by the backbone network embedded in our network structure, by replacing the existing U-Net structure with a DeepLab-v3 with MobileNet backbone structure, whose number of parameters is about 60% of UNet. Experimental results show that, in this smaller network, lower IoU is obtained but more capability of mutual guidance is released. See Table 2 bottom for more comparison details.

4.3.5 Training strategy

Gradient stopping. We also conducted an experiment to check the necessity of gradient stopping. Figure 7 shows two examples. From them, we see that with gradient stopping disabled, the detected skin masks tend to have a high precision but low recall, which is more likely to be trivial results like $e_\kappa = 0$. This is a local minimum of our network, caused by the setting $G_\kappa = e_\kappa = 0$ in Stage 1. When gradient stopping is enabled, we keep the gradients from being back-propagated to $\{O_\kappa\}$, so that the trivial local minimum cannot be easily reached.



Figure 7. Skin detection results trained without (Column 2) or with (Column 3) gradient stopping. The three values under the results are IoU (black), Precision (red) and Recall (blue).

Initial guidance e_S, e_B . We also conducted an experiment by providing the guidance $\{G_\kappa\}$ with $\{e_\kappa \neq 0\}$. Specifically, e_B is a body bounding box mask and e_S is skin detection results by GMM. We also trained our network with mutual guidance from scratch, and achieved 80.74% IoU. This value is higher than 80.11% which was produced by the version of $\{e_\kappa = 0\}$, meaning by providing more informative guidance in Stage 1, our network could be more easily trained.

Finetune. We also compared the performance of our network between train-from-scratch and finetune versions, illustrated in Curve 2 and 5 in Figure 4 Top. With finetune involved, our network obtained a higher average IoU in validation dataset.

5. Conclusion

We have presented a new data-driven method for robust skin detection from a single human portrait image. To achieve this goal, we designed a dual-task neural network for joint detection of skin and body. Our dual-task network contains a shared encoder but two decoders, for the two tasks separately. The two decoders work in a mutually guided manner, i.e. either output of the skin or body decoder also serves as a guidance to boost the detection performance for its counterpart. Furthermore, our network can be trained in a semi-supervised manner, i.e. we do not require both types of ground truth exist in one training data sample. It is achieved by a newly designed semi-supervised loss as proposed. We conducted extensive experiments to demonstrate the effectiveness of mutual guidance, semi-supervised losses and various training strategies. Results also show that our method outperforms the state-of-the-art in skin detection. We also hope that the idea of mutual guidance could inspire more works in related problems like image/video denoising [19], detection [27], completion [8, 39, 14], segmentation [34], generation or compression [37, 25, 22, 36, 41] etc. in the future.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China under Grant No. 61672125, No. 61702565 and No. U1811463. We also thank the anonymous reviewers' comments to facilitate the improvement of this paper.

References

- [1] H. K. Al-Mohair, J. Saleh, and S. Saundi. Impact of color space on human skin color detection using an intelligent system. In *1st WSEAS International Conference on Image Processing and Pattern Recognition (IPPR'13)*, 2013. 2
- [2] D. Chai, S. L. Phung, and A. Bouzerdoum. Skin color detection for face localization in human-machine communications. In *Proceedings of the Sixth International Symposium on Signal Processing and its Applications (Cat. No. 01EX467)*, volume 1, pages 343–346. IEEE, 2001. 1
- [3] L. Chen, J. Zhou, Z. Liu, W. Chen, and G. Xiong. A skin detector based on neural network. In *IEEE 2002 International Conference on Communications, Circuits and Systems and West Sino Expositions*, volume 1, pages 615–619. IEEE, 2002. 2, 5, 6
- [4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018. 1, 5, 6
- [5] W. Chen, K. Wang, H. Jiang, and M. Li. Skin color modeling for face detection and segmentation: a review and a new approach. *Multimedia Tools and Applications*, 75(2):839–862, 2016. 1
- [6] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008. 2
- [7] L. Deng, G. Hinton, and B. Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8599–8603. IEEE, 2013. 2
- [8] Y. Ding, C. Wang, H. Huang, J. Liu, J. Wang, and L. Wang. Frame-recurrent video inpainting by robust optical flow inference. *arXiv preprint arXiv:1905.02882*, 2019. 8
- [9] C. Erdem, S. Ulukaya, A. Karaali, and A. T. Erdem. Combining haar feature and skin color based classifiers for face detection. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1497–1500. IEEE, 2011. 1, 2
- [10] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, and S. Iyengar. Computational health informatics in the big data age: a survey. *ACM Computing Surveys (CSUR)*, 49(1):12, 2016. 1
- [11] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2
- [12] X. Han, Z. Li, H. Huang, E. Kalogerakis, and Y. Yu. High-resolution shape completion using deep neural networks for global structure and local geometry inference. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 85–93, 2017. 2
- [13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 5, 6
- [14] L. Huynh, W. Chen, S. Saito, J. Xing, K. Nagano, A. Jones, P. Debevec, and H. Li. Mesoscopic facial geometry inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8407–8416, 2018. 8
- [15] A. Jaimes and N. Sebe. Multimodal human-computer interaction: A survey. *Computer vision and image understanding*, 108(1-2):116–134, 2007. 1
- [16] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002. 5, 6
- [17] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018. 2
- [18] J. Kovac, P. Peer, and F. Solina. *Human skin color clustering for face detection*, volume 2. IEEE, 2003. 2, 5, 6
- [19] J. Liu, C.-H. Wu, Y. Wang, Q. Xu, Y. Zhou, H. Huang, C. Wang, S. Cai, Y. Ding, H. Fan, et al. Learning raw image denoising with bayer pattern unification and bayer preserving augmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 8
- [20] Q. Liu and G.-z. Peng. A robust skin color based face detection algorithm. In *2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR 2010)*, volume 2, pages 525–528. IEEE, 2010. 2
- [21] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2
- [22] X. Meng, X. Deng, S. Zhu, S. Liu, C. Wang, C. Chen, and B. Zeng. Mganet: A robust model for quality enhancement of compressed video. *arXiv preprint arXiv:1811.09150*, 2018. 8
- [23] V. Powar, A. Kulkarni, R. Lokare, and A. Lonkar. Skin detection for forensic investigation. In *2013 International Conference on Computer Communication and Informatics*, pages 1–4. IEEE, 2013. 2
- [24] J. Qiang-rong and L. Hua-lan. Robust human face detection in complicated color images. In *2010 2nd IEEE International Conference on Information Management and Engineering*, pages 218–221. IEEE, 2010. 2
- [25] H. Qiu, C. Wang, H. Zhu, X. Zhu, J. Gu, and X. Han. Two-phase hair image synthesis by self-enhancing generative model. *arXiv preprint arXiv:1902.11203*, 2019. 8
- [26] S. S. Rautaray and A. Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54, 2015. 1

- [27] J. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan. Look, listen and learn—A multimodal lstm for speaker identification. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. 8
- [28] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 3, 5, 6
- [29] M.-J. Seow, D. Valaparla, and V. K. Asari. Neural network based skin color model for face detection. In *32nd Applied Imagery Pattern Recognition Workshop, 2003. Proceedings.*, pages 141–145. IEEE, 2003. 2
- [30] W. Tan, G. Dai, H. Su, and Z. Feng. Gesture segmentation based on ycb'cr' color space ellipse fitting skin color modeling. In *2012 24th Chinese Control and Decision Conference (CCDC)*, pages 1905–1908. IEEE, 2012. 2, 7
- [31] W. R. Tan, C. S. Chan, P. Yogarajah, and J. Condell. A fusion approach for efficient human skin detection. *IEEE Transactions on Industrial Informatics*, 8(1):138–147, 2012. 8
- [32] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1818–1827, 2018. 2
- [33] M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, and Y. Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 507–522, 2018. 4
- [34] C. Wang, Y. Guo, J. Zhu, L. Wang, and W. Wang. Video object co-segmentation via subspace clustering and quadratic pseudo-boolean optimization in an mrf framework. *IEEE Transactions on Multimedia*, 16(4):903–916, 2014. 8
- [35] C. Wang, H. Huang, X. Han, and J. Wang. Video inpainting by jointly learning temporal structure and spatial details. *arXiv preprint arXiv:1806.08482*, 2018. 2
- [36] C. Wang, J. Zhu, Y. Guo, and W. Wang. Video vectorization via tetrahedral remeshing. *IEEE Transactions on Image Processing*, 26(4):1833–1844, 2017. 8
- [37] Y. Wang, H. Huang, C. Wang, T. He, J. Wang, and M. Hoai. Gif2video: Color dequantization and temporal interpolation of gif images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1419–1428, 2019. 8
- [38] Q. Wu, R. Cai, L. Fan, C. Ruan, and G. Leng. Skin detection using color processing mechanism inspired by the visual system. 2012. 2
- [39] S. Yamaguchi, S. Saito, K. Nagano, Y. Zhao, W. Chen, K. Olszewski, S. Morishima, and H. Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)*, 37(4):162, 2018. 8
- [40] A. Zaidan, N. N. Ahmad, H. A. Karim, M. Larbani, B. Zaidan, and A. Sali. On the multi-agent learning neural and bayesian methods in skin detector and pornography classifier: An automated anti-pornography system. *Neurocomputing*, 131:397–418, 2014. 2
- [41] Y. Zhou, L. Hu, J. Xing, W. Chen, H.-W. Kung, X. Tong, and H. Li. Hairnet: Single-view hair reconstruction using convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–251, 2018. 8
- [42] Q. Zhu, K.-T. Cheng, C.-T. Wu, and Y.-L. Wu. Adaptive learning of an accurate skin-color model. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 37–42. IEEE, 2004. 1, 2
- [43] H. Zuo, H. Fan, E. Blasch, and H. Ling. Combining convolutional and recurrent neural networks for human skin detection. *IEEE Signal Processing Letters*, 24(3):289–293, 2017. 2, 5, 6