HLNet: A Unified Framework for Real-Time Segmentation and Facial Skin Tones Evaluation

Xinglong Feng †, Xianwen Gao * and Ling Luo †

College of Information Science and Engineering, Northeastern University, Shenyang 110819, China; 1610238@stu.neu.edu.cn (X.F.); lingluo@stumail.neu.edu.cn (L.L.)
*
Correspondence: gaoxianwen@ise.neu.edu.cn
† These authors contributed equally to this work.

Received: 21 October 2020; Accepted: 30 October 2020; Published: 1 November 2020

Abstract: Real-time semantic segmentation plays a crucial role in industrial applications, such as autonomous driving, the beauty industry, and so on. It is a challenging problem to balance the relationship between speed and segmentation performance. To address such a complex task, this paper introduces an efficient convolutional neural network (CNN) architecture named HLNet for devices with limited resources. Based on high-quality design modules, HLNet better integrates high-dimensional and low-dimensional information while obtaining sufficient receptive fields, which achieves remarkable results on three benchmark datasets. To our knowledge, the accuracy of skin tone classification is usually unsatisfactory due to the influence of external environmental factors such as illumination and background impurities. Therefore, we use HLNet to obtain accurate face regions, and further use color moment algorithm to extract its color features. Specifically, for a $224 \times 224$ input, using our HLNet, we achieve 78.39% mean IoU on Figaro1k dataset at over 17 FPS in the case of the CPU environment. We further use the masked color moment for skin tone grade evaluation and approximate 80% classification accuracy demonstrate the feasibility of the proposed method.

Keywords: semantic segmentation; deep convolutional neural network; skin tone classification; color moment

1. Introduction

Augmented Reality (AR) technology has been widely used in various fields as a hot spot in recent years. Among them, automatic hair-dyeing based on 2D color imaging, as shown in Figure 1, attracts the most attention, the prerequisite for which is the precise segmentation of the hair area. Early studies on hair segmentation focused primarily on hand-crafted features [1–3], which required professional skills and was labor-intensive. In the meanwhile, the generalization of the model was generally poor.

In recent years, the advent of deep convolutional neural networks (DCNNs) improved the performance of many tasks, the most significant of which is semantic segmentation. Semantic segmentation is an advanced visual task, whose goal is to assign dense labels to each image pixel. As its subtask, hair segmentation has also received widespread attention in recent years. For example, Borza et al. [4] performed hair segmentation with the aid of symmetrical UNet, which was subsequently refined using morphological knowledge. Wen et al. [5] proposed an end-to-end detection-segmentation system to implement detailed face labeling, including hair. Using a pyramid FCN encoded multi-level feature maps, this method effectively alleviates the imbalance of semantic categories. More recently, Luo et al. [6] designed a lightweight segmentation network that...
combines the advantages of multiple modules to effectively solve the ambiguity of edge semantics. At the same time, the method is suitable for mobile devices.

Even so, its application is limited by the following factors. First of all, due to the diverse appearance of hair and its complex structural information, it is extremely difficult to accurately process the edges [7]. Although existing semantic segmentation methods [8–10] have relatively high segmentation performance for simple objects, only coarse masks can be obtained for hair segmentation. Secondly, the vast majority of networks require graphics processing units (GPUs) with powerful computing capabilities that most mobile devices do not have, which greatly limits their usage scenarios; Thirdly, taking into account runtime limitations, Conditional Markov random fields (CRFs) [11] is not suitable for processing the edge (e.g., shredded hair), so it is necessary to find an alternative solution. Taking all these factors into consideration, real-time hair dyeing faces enormous challenges. On the other hand, e-commerce and digital interaction with clients allows people to buy their favorite products without leaving home. Among them, the robust product recommendation function plays an important role. Automatic assessment of skin tone levels which makes it possible to personalize recommendation for beauty products. However, taking into account complex external factors, such as illumination, shadows, and background impurities can affect the determination, in which case even an experienced skin therapist can hardly judge it with the naked eye. This paper is dedicated to solving the aforementioned problems using machine learning and promising deep learning algorithms.

![Figure 1](image)

**Figure 1.** Automatic hair dyeing exemplar. (a) Input RGB image. (b) The guided filter output of our proposed algorithm. (c) Final dyed rendering.

In this paper, we strive to balance the relationship between performance and efficiency, and provide a much simpler and more compact alternative for our segmentation task. To get accurate segmentation results, local and global context information should be considered simultaneously. Based on this observation, we propose a spatial and context information fusion framework called HLNet, which integrates high-dimensional and low-dimensional feature maps in parallel. While increasing the receptive field, it effectively alleviates the insufficient extraction of shallow features. Moreover, inspired by BiSeNet [12], the Feature Fusion Module (FFM) module is used to re-encode feature channels using context to improve feature representation in a particular category. Extensive experiments confirm that our HLNet achieves significant trade-off between efficiency and accuracy. Considering that background illumination is not conducive to identifying skin, we extract features (a.k.a. masked color moments) based on the segmented face and color moment algorithm [13]. The mask color moments are then thrown into a powerful Random Forest Classifier [14] to evaluate a person’s skin tone level. Furthermore, we verify the feasibility of the method on a manually labeled dataset.

In summary, our main contributions are as follows:

1. We propose an efficient hair and face segmentation network that uses newly proposed modules to achieve real-time inference while guaranteeing performance.
(2) A module called InteractionModule is given, which exploits multi-dimensional feature interactions to mitigate the weakening of spatial information as the network becomes deeper and deeper.

(3) A novel skin color level evaluation algorithm is proposed and obtains accurate results on a manually labeled dataset.

(4) Our method achieves superior results on multiple benchmark datasets.

The rest of the paper is organized as follows. In Section 2, we review previous work done on lightweight model design and edge post-processing. In Section 3, we describe the proposed method in detail. Section 4 provides experimental data and parameter configuration as well as a manually annotated dataset. In Section 5, we report the experimental results. Conclusion marks and future work are drawn in Section 6.

2. Related Works

Real-time semantic segmentation. Since pioneering work [8] based on deep learning, many high-quality backbones [15–18] have been derived. However, due to the requirements of computationally limited platforms (e.g., drones, autonomous driving and smartphone), researchers pay more attention to the efficiency of the networks than just the performance. ENet [19] is the first lightweight network for real-time scene segmentation which does not apply any post-processing steps in an end-to-end manner. Zhao et al. [20] introduced a cascade feature fusion unit to quickly achieve high-quality segmentation. Howard et al. [21] proposed a compact encoder module based on a streamlined architecture that uses depthwise separable convolutions to build light-weight deep neural networks. Poudel et al. [22] combined spatial detail at high resolution with deep features extracted at lower resolution yielding beyond real-time effects. DFANet [23] starts from a single lightweight backbone and aggregates discriminative features through sub-network and sub-stage cascade respectively. Recently, LEDNet [24] has been proposed which channel split and shuffle are used in each residual block to greatly reduce computation cost while maintaining higher segmentation accuracy.

Contextual information. Some details cannot be recovered during conventional up-sampling of the feature maps to restore the original image size. The design of skip connections [25] can alleviate this deficiency to some extent. Besides, Zhao et al. [17] proposed a pyramid pooling module that can aggregate context information from different regions to improve the ability to capture multi-scale information. Zhang et al. [26] designed a context encoding module to introduce global contextual information, which is used to capture the context semantics of the scene and selectively highlight the feature map associated with a particular category. Fu et al. [27] addressed the scene parsing task by capturing rich contextual dependencies based on spatial and channel attention mechanisms, which significantly improved the performance on numerous challenging datasets.

Post processing. Generally, the quality of the above segmentation methods is obviously rough and requires additional post-processing operations. Post-processing mechanisms are usually able to improve image edge detail and texture fidelity, while maintaining a high degree of consistency with global information. Chen et al. [28] proposed a CRF post processing method that overcomes poor localization in a non-end-to-end way. CRFasRNN [11] considers the CRF iterative reasoning process as an RNN operation in an end-to-end manner. To eliminate the excessive execution time of CRF, Levinhtein et al. [29] presented a hair matting method with real-time performance on mobile devices.

Our approach draws upon these strengths. Furthermore, for the upstream skin tone grading task, we employ masked color moment to handle it, which will be discussed in Section 3.2.
3. Methodology

3.1. High-To-Low Dimension Fusion Network

The proposed HLNet network is inspired by HRNet [30] which maintains high-resolution representation through the whole process by connecting high-to-low resolution convolutions in parallel. Figure 2 illustrates the overall framework of our model. We experimentally prune the model parameters to increase the speed without excessive performance degradation. Furthermore, the existing SOTA modules [12,22,31,32] are reasonably combined to further improve the performance of the network. Table 1 gives an overall description of the modules involved in the designed network. The model consists of different kinds of convolution modules, bilinear up-sampling, bottlenecks, and other feature maps communication modules. In the following part, we will expand the above modules in detail.

![Figure 2: An overview of our asymmetric encoder-decoder network. Blue, red and green represents background, hair mask recolored and face mask recolored, respectively. In the dotted rectangle (also called InteractionModule), arrows in different directions represent different operations. “C” and “+” represent Add and Concatenate (abbreviated as Concat) operations, respectively.](image)

<table>
<thead>
<tr>
<th>Stage</th>
<th>Type</th>
<th>Output Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoder</td>
<td>Conv2D</td>
<td>224 × 224 × 3</td>
</tr>
<tr>
<td></td>
<td>DwConv2D</td>
<td>112 × 112 × 32</td>
</tr>
<tr>
<td></td>
<td>DwConv2D</td>
<td>56 × 56 × 64</td>
</tr>
<tr>
<td></td>
<td>InteractionModule</td>
<td>28 × 28 × 64</td>
</tr>
<tr>
<td></td>
<td>FFM</td>
<td>28 × 28 × 128</td>
</tr>
<tr>
<td></td>
<td>DilatedGroup</td>
<td>28 × 28 × 32</td>
</tr>
<tr>
<td>Decoder</td>
<td>UpSample2D</td>
<td>224 × 224 × 32</td>
</tr>
<tr>
<td></td>
<td>Conv2D</td>
<td>224 × 224 × 3</td>
</tr>
<tr>
<td></td>
<td>SoftMax</td>
<td>224 × 224 × 3</td>
</tr>
</tbody>
</table>

Table 1. HLNet consists of an asymmetric encoder and decoder. The whole network is mainly composed of standard convolution (Conv2D), deep separable convolution (DwConv2D), inverted residual bottleneck blocks, bilinear upsampling (UpSample2D) module and several custom modules.
To preserve details as much as possible, the downsampling rate of the entire network is set to 1/8. Specifically, in the first three layers, we refer to Fast-SCNN [22] to employ vanilla convolution and depth separable convolution for fast down-sampling in order to ensure low-level feature sharing. Depth separable convolution reduces the amount of model parameters effectively while achieving a comparable representation ability. The above-mentioned convolution is uniform with a step size of 2 and a kernel size of 3 × 3, followed by BN [33] and a ReLU activation function.

According to FCOS [34], the low-dimensional detail information of the feature map promotes the segmentation of small objects, so we strengthen the model’s ability to represent details by stacking low-dimensional layers. Moreover, the interaction of high-resolution and low-resolution information facilitates learning of multi-scale information representation. We draw the above advantages and propose an information interaction module (InteractionModule) with feature maps of different resolutions to obtain elegant output results. Conceptually, for the backbone \( \phi^i_n(x) \), a stage process can be defined as \( \phi^i_n \), where \( n \) and \( i \) represent the index and the width of the stage, respectively. The calculation process in the dotted rectangle can be formulated as:

\[
\phi^i_n = \begin{cases} 
\text{Conv}(\phi^i_{n-1}), & n = 4, \\
\sum_{i=1}^{M} \text{Conv}(\phi^i_n), & n = 5, \\
\text{Concat}(\phi^i_{n-1}, \ldots, \phi^M_{n-1}), & \text{otherwise}
\end{cases}
\]

where \( M \) is 3. \text{Conv} and \text{Concat} represent convolution operator and feature maps are stacked in the channel dimension, respectively. MobileNet v2 [31] takes advantage of residual block and deep separable convolution, which greatly reduces the calculation parameters while effectively avoiding gradient dispersion. The inverted residual block proposed by MobileNet v2 is utilized to improve the sparse parameter space by proper pruning. In particular, for \( \phi^i_n(i = 1, \ldots M) \), the corresponding parameters \( \{k = 3, c = 64, t = 6, s = 1, n = 3\} \rightarrow \{k = 3, c = 96, t = 6, s = 2, n = 3\} \rightarrow \{k = 3, c = 128, t = 6, s = 4, n = 3\} \) are given in order, where \( k, c, t, s \) and \( n \) denote the size of convolution kernel, the number of feature map channels, the channel multiplication factor, stride and the number of module repetitions, respectively. Next, feature maps of different scales are combined and exchanged by using a 1 × 1 convolution, strided convolution or upsampling. 1 × 1 convolution can well perform the dimensional increase and decrease of the feature map without significantly increasing the amount of parameters. In addition, the ReLU behind it can increase the overall nonlinear fitting ability of the network. The last part of the InteractionModule is implemented by using \text{Concat} in order to aggregate multi-scale context features. Subsequently, following the FFM Attention [12], the model focuses more on channels that contain important features and suppress those that are not important. It is composed of the following: the FFM performs an element-level multiplication operation with the input after passing through a global pooling layer and two convolutional layers with ReLU and Sigmoid, respectively. In order to mitigate the gradient disappearance in the back propagation of parameters, skip connections are added between the input and output. Then, to capture multi-scale context information, we also introduce a multi-receptive field fusion block (DilatedGroup), whose dilation rates are set to 2, 4, and 8.

For simplicity, the decoder performs bilinear upsampling (transposed convolution layer can cause gridding artifacts [29]) directly on the 28 × 28 feature map followed by a 3 × 3 convolution to maintain that the number of channels and the number of categories are consistent. Finally, a SoftMax layer is connected for dense classification.

In terms of loss function, we apply generalized dice loss (GDL) [35] to compensate for the segmentation performance of small objects, which is formulated as:

\[
\text{GDL}_{\text{loss}} = 1 - \frac{2 \cdot \sum_{l=1}^{L} \omega_l \cdot \sum_{n=1}^{N} r_{ln} p_{ln}}{L \cdot \sum_{l=1}^{L} \omega_l \cdot \sum_{n=1}^{N} r_{ln} + p_{ln}}
\]
\[ \omega_l = \frac{1}{(\sum_{n=1}^{N} r_{ln})^{2}} \]  

(3)

where \( p \) denotes the SoftMax output and \( r \) denotes the one-hot encoding of the ground truth. \( N \) and \( L \) represent the total number of pixels and categories, respectively. Equation (3) gives the expression of \( \omega_l \), which is the category balance coefficient.

To pursue perceptual consistency and reduce the time complexity of running, we advocate the idea of Guided Filter \([36,37]\) to achieve edge-preserving and denoising. Guided Filter can effectively suppress gradient-reversal artifacts and produce visually pleasing edge profiles. Given a guidance image \( I \) and filtering input image \( P \), our goal is to learn a local linear model to describe the relationship between the former and the output image \( Q \) while seeking consistency between \( P \) and \( Q \) just like the role of Image Matting \([38]\). During the experiment, \( s, r, \zeta \) are empirically set to 4, 4, and 50, respectively.

### 3.2. Facial Skin Tone Classification

The purpose of the second stage is to classify the facial skin tone. Usually for Asians, we divide it into porcelain white, ivory white, medium, yellowish and black. For skin tone features, due to the small feature space, it is not suitable to use DCNNs-based methods for feature extraction. Therefore, after repeated thinking and experimental trial and error, the scheme is selected to extract the color moment of the image as the features to be learned and put it into the classic machine learning algorithm for learning. Considering facial skin tone in complex scenes, background lighting has a incurable impact on the results. So we employ image morphology algorithms and pixel-level operations to get rid of background interference. Algorithm 1 summarizes the pseudo code of the extraction process. The pre-processed face image is used to extract the color moment features, which are then put into a powerful Random Forest Classifier \([14]\) for learning. Color moment can be expressed as:

\[ \mu_i = \frac{1}{N} \sum_{j=1}^{N} p_{i,j} \]  

(4)

\[ \sigma_i = \left( \frac{1}{N} \sum_{j=1}^{N} (p_{i,j} - \mu_i)^2 \right)^{\frac{1}{2}} \]  

(5)

\[ s_i = \left( \frac{1}{N} \sum_{j=1}^{N} |p_{i,j} - \mu_i|^3 \right)^{\frac{1}{3}} \]  

(6)

where \( p_{i,j} \) denotes the probability of a pixel in the \( i \) channel with a value of \( j \), and \( N \) denotes the total number of pixels. Color feature \( F_{\text{color}} = [\mu_U, \sigma_U, s_U, \mu_V, \sigma_V, s_V, \mu_Y, \sigma_Y, s_Y] \), \( U, V, Y \) denote each channel of the image.
**Algorithm 1:** Segmentation-based inference algorithm for smoothed facial region extraction

**Input:** Given a dataset \( D = \{X_1, X_2, \ldots, X_n\}, \) where \( m \) denotes the total number of the image \( X_i; \) \( y = f(X, \theta) \) represents the output of HLNet, and \( \theta \) is the weight of our fine-tuned network. For image processing, we use the Python bindings of the OpenCV library, named cv2.

**Output:** Smoothed facial region \( F_i. \)

1. **Step 1:**
   1. \( M_i^j = 0; \)
   2. **for** \( i = 1, 2, \ldots, n \) **do**
   3. **for** \( j = 1, 2, \ldots, m \) **do**
   4. if \( y_i^j == \) index of the face **then**
   5. \( M_i^j = 255; \)
   6. **end**
   7. **end**
   8. **end**

2. **Step 2:**
   1. **for** \( i = 1, 2, \ldots, n \) **do**
   2. \( P_i(w, h) \leftarrow \text{cv2.ero} \leftarrow \text{min} M_i(w + w', h + h'); \)
   3. \( Q_i(w, h) \leftarrow \text{cv2.bilateralFilter} \leftarrow P_i(w, h); \)
   4. \( F_i(w, h) \leftarrow \text{cv2.bitwise}_\text{and} \leftarrow (I_i(w, h), Q_i(w, h)) \)
   5. **end**

**4. Experiments**

**4.1. Implementation Details**

Our experiments are conducted using Keras framework with Tensorflow backend. Standard mini-batch gradient descent (SGD) is employed as the optimizer with a momentum of 0.98, a weight decay of \( 2 \times 10^{-5} \), and a batch size of 64. We adopt the widely equipped “poly” learning rate policy in configuration where the initial rate is multiplied by \( (1 - \frac{\text{iter}}{\text{total_iter}})^{\text{power}} \) with \( \text{power} 0.9 \) and initial learning rate is set as \( 2.5 \times 10^{-3} \). Data augmentation includes normalization, random rotation \( \theta_{\text{rotation}} \in [-20, 20] \), random scale \( \theta_{\text{scale}} \in [-20, 20] \), random horizontal flip and random shift \( \theta_{\text{shift}} \in [-10, 10] \). For fair comparison, all the methods are conducted on a server equipped with a single NVIDIA GeForce GTX1080 Ti GPU. Code is available at: https://github.com/JACKYLUO1991/Face-skin-hair-segmentaiton-and-skin-color-evaluation.

**4.2. Datasets**

Data is the soul of deep learning, because it determines the upper limit of an algorithm to some extent. To ensure the robustness of the algorithm, it is necessary to construct a dataset with human faces in extreme situation such as large angles, strong occlusions, complex lighting changes, etc.

**4.2.1. Face and Hair Segmentation Datasets**

Labeled Faces in the wild (LFW). Ref. [39] dataset consists of more than 13,000 images on the Internet. We use its extension version (Part Labels) during the experiment which automatically labeled via a super-pixel segmentation algorithm. We adopt the same data division method in [4] as 1500 images in the training, 500 used in validation, and 927 used for testing.

Large-scale CelebFaces Attributes dataset (CelebA). Ref. [40] consisting of more than 200 k celebrity images, each with multiple attributes. The main advantage of this dataset is that it combines large pose variations and background clutter making the knowledge learned from this dataset easier to
satisfy demand of actual products. In the experiment, we adopt the CelebHair version (http://www.cs.ubbcluj.ro/~dadi/face-hair-segm-database.html) of CelebA in [4] which includes 3556 images. We use the same configuration as the original paper, i.e., 20% for validation.

**Figaro1k.** For the last dataset, we employ Figaro1k [41], which is dedicated to hair segmentation. It needs to be considered that the dataset is developed for general hair detection, many of which do not include faces, which is not conducive to subsequent experiments. In this case, we follow the pre-processing in [7], leaving 171 images for experiments. To better take advantage of batch training, offline data augmentation is adopted to expand the images (×10).

### 4.2.2. Manually Annotated Dataset

An outstanding contribution of this work is a manually labeled facial skin tone rating dataset. In the process of labeling, three professionally trained makeup artists rated the face tone color using a voting mechanism. Once all three markers have judged the results differently, the label will be decided by a makeup artist with 5 or more years of experience. Our face data is collected from the web without conflicts of interest. The obtained image is filtered by an off-the-shelf face detection library (i.e., MTCNN [42]) to remove images without detected faces, and the remaining ones are used for feature extraction and further machine learning. The number of each category is 95, 95, 96, 93 and 94, samples are shown in Figure 3. Besides, their statistical distribution are plotted in Figure 4.

![Figure 3](image-url)

**Figure 3.** Manually marked face skin hue level sample after voting mechanism. From (a–e), it represents porcelain white, ivory white, medium, yellowish and black races. In order to have a consistent understanding of image classification criteria, only Chinese actresses are used here. It is worth noting that the images are crawled from the online web with no conflict of interest.

![Figure 4](image-url)

**Figure 4.** Pie chart visualization. (a) viewpoint graph. According to the yaw angle, it is divided into small (|θ| < 15°), moderate (15° ≤ |θ| ≤ 45°) and large (|θ| > 45°). (b) Occlusion graph. Including ‘mild occlusion’ (<20%), ‘severe occlusion’ (>50%) and ‘moderate occlusion’ between the two items.
4.3. Evaluation Metrics

All segmentation experiments are applied to mean-intersection-over-union (mIoU) criterion. The definition of mIoU is as follows:

\[
mIoU = \frac{1}{1 + k} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}
\]  

(7)

where \( k + 1 \) is the number of classes (including background), \( p_{ij} \) indicates the number of pixels that belong to category \( i \) but have been misjudged as category \( j \). For more metrics, please refer to [8].

5. Results and Discussion

5.1. Segmentation Results

In this section, we carry on the experiments to demonstrate the potential of our segmentation architecture in terms of accuracy and efficiency trade-off.

5.1.1. Overall Comparison

We use four FCN [8] introduced metrics to evaluate the performance of our algorithm. Subsequently, comparative experiments across different datasets with the outstanding UNet variant [4] are constructed. Unless otherwise stated, the input resolution is \( 224 \times 224 \). The training continues for 200 epochs, after which the model will become saturated. Table 2 reports the qualitative results.

<table>
<thead>
<tr>
<th>Metric</th>
<th>LFW (OC = 3)</th>
<th>CelebHair (OC = 3)</th>
<th>Figaro1k (OC = 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mIoU</td>
<td>83.46</td>
<td>83.81</td>
<td>77.75</td>
</tr>
<tr>
<td>fwIoU</td>
<td>92.75</td>
<td>90.28</td>
<td>83.01</td>
</tr>
<tr>
<td>pixelAcc</td>
<td>95.83</td>
<td>94.69</td>
<td>90.28</td>
</tr>
<tr>
<td>mPixelAcc</td>
<td>88.84</td>
<td>90.35</td>
<td>84.72</td>
</tr>
</tbody>
</table>

Experimental results show that our HLNet outperforms the trimmed U-Net (tU-Net) [4] by a large margin, except for LFW dataset. Nevertheless, one drawback of fast down-sampling is that the feature extraction for the shallow layers is not sufficient. As we know, shallow features contribute to extracting texture and edge details, so our HLNet is slightly worse than the tU-Net in LFW dataset (LFW facial details are blurrier than others).

From another perspective, considering the latency time, we reach 60 ms per image on an Intel Core i5-7500U CPU without any tricks. We can further reach no more than 10 ms under GPU. Comparing tU-Net with HLNet (8 ms vs. \( 7.2 \pm 0.3 \) ms) shows that the latter is more efficient, while performance is more remarkable. This conclusion suggests that we can further apply this framework to the edge and embedded devices with small memory and battery budget. The qualitative analysis results are shown in Figure 5. Post-processing employs Guided Filter to achieve more realistic edge results.
Figure 5. Samples of hair and face segmentations on different datasets.

5.1.2. Comparison with SOTA Lightweight Networks

In this subsection, we compare our algorithm with several state-of-the-art (SOTA) lightweight networks including ENet [19], LEDNet [24], Fast-SCNN [22], MobileNet [21] and DFANet [23] on CelebHair test set. For fair comparison, we re-implement the above networks under the same hardware configuration without any fine-tuning or fancy tuning techniques. It should be noted that the framework implementation is slightly different from the original, so the results may be slightly different, but the overall performance deviation is within the acceptable range. Since ENet has a downsampling rate of 32, we resize all the input to $256 \times 256$. In addition, we measure frames per second (FPS) in our CPU environment without any running loads, which takes an average of 200 forward propagations.

From Figure 6 and Table 3, we can observe that our proposed method is more accurate than other methods. Compared with the sub-optimal ENet, our method improves mIoU by 0.35%, while the FPS is half higher. Although DFANet has $2 \times$ less parameters, as well as $11 \times$ less FLOPs than our HLNet, it delivers poor segmentation accuracy of 7.44% in terms of mIOU. We conjecture that this is due to DFANet’s overdependence on pre-trained lightweight backbones. From Figure 6c, it can be clearly observed that the DFANet has a serious misclassification on pixels. MobileNet’s situation is consistent with DFANet. In particular, our HLNet is 3.18% higher than Fast-SCNN in terms of accuracy, and the parameters are reduced by 0.4 M. Excessive deep separable convolutions affect its performance, and even if this reduces time delay and computational complexity (FLOPs), it gets insufficient generalization capabilities. Compare the last line of Figure 6g,h, which contains the second person (the latter one), even if the Ground Truth does not mark it. In contrast, benefiting from the rich context captured by the DilatedGroup, our method can roughly segment it. Moreover, compared with other methods, with the help of the introduced InteractionModule, HLNet has an advantage in detail processing of multi-scale objects (e.g., hairline). A more intuitive reference diagram for comparison of different methods is shown in Figure 7. The whole experiment demonstrate that our HLNet achieve the best trade-off between accuracy and efficiency.
Figure 6. Qualitative comparison results with other SOTA methods. From (a) to (h) are input images, ground truth, segmentation outputs from DFANet [23], ENet [19], MobileNet [21], LEDNet [24], Fast-SCNN [22] and our HLNet. From top to bottom, the difficulty of segmentation increases in turn.

Table 3. Comparison with SOTA approaches on CelebHair test set in terms of segmentation accuracy and execution efficiency. “†” indicates fine-tuning from LFW. 0.5 represents the contraction factor. “#Param” represents the number of model parameters. Bold means better.

<table>
<thead>
<tr>
<th>Model</th>
<th>#Param (M)</th>
<th>FPS</th>
<th>FLOPs (G)</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENet [19]</td>
<td>0.36</td>
<td>8.24</td>
<td>0.94</td>
<td>89.97</td>
</tr>
<tr>
<td>LEDNet [24]</td>
<td>2.3</td>
<td>6.44</td>
<td>3.28</td>
<td>88.63</td>
</tr>
<tr>
<td>Fast-SCNN [22]</td>
<td>1.6</td>
<td>20.35</td>
<td>0.41</td>
<td>87.14</td>
</tr>
<tr>
<td>MobileNet(0.5) + UNet [21]</td>
<td>0.37</td>
<td>5.80</td>
<td>0.75</td>
<td>86.08</td>
</tr>
<tr>
<td>DFANet [23]</td>
<td>0.42</td>
<td>17.72</td>
<td>0.08</td>
<td>82.88</td>
</tr>
<tr>
<td>HLNet (ours)</td>
<td>1.2</td>
<td>12.23</td>
<td>0.94</td>
<td>90.32</td>
</tr>
<tr>
<td>HLNet (ours) †</td>
<td>1.2</td>
<td>12.23</td>
<td>0.94</td>
<td>90.98</td>
</tr>
</tbody>
</table>

Figure 7. Results for running speed vs. performance of different methods.

5.1.3. Ablation Study

We further conduct the ablation experiments on the Figaro1k test set, and follow the same training strategy for the fairness of the experiments. In addition, we mainly evaluate the impact of
InteractionModule (IM) and DilatedGroup (DG) components on the results, as illustrated in Figure 8. IM without information exchange (connected using Upsampling and Concat) and a $3 \times 3$ convolution with a rate of 1 are used to replace the corresponding components as baseline.

![Figure 8. Ablation Experiments on the Figaro1k test set.](image)

On the one hand, IM module can capture multi-resolution patterns. On the other hand, DG module fuses multi-scale features while increasing the receptive field. When we append DG and IM modules respectively, mIoU increases by 1.54% and 3.19% relative to the baseline. When we apply two modules at the same time, mIoU increases dramatically by 4.26%. The obvious performance gains reflects the efficiency of our proposed modules.

5.2. Facial Skin Tone Classification Results

In the second phase of the experiment, we construct comparative studies to compare the influence of different color spaces and different experimental protocols over the results.

As shown in Table 4, we report the accuracy of facial skin color classification. The best results are obtained using the YCrCb color space with color moment backend, with an accuracy of 80%. It should be noted that before putting into the Classifier, the data first needs to be oversampled to ensure that the number of samples is consistent across the different categories. We simply split the dataset into 8:2 for training and testing, and then use the powerful Random forest Classifier for training. Figure 9 provides the confusion matrix for this configuration. As can be seen from it, the main errors are between adjacent categories, a situation that also plagues a trained professional makeup artist when he/she is labeling data.

<table>
<thead>
<tr>
<th>Method</th>
<th>RGB</th>
<th>HSV</th>
<th>YCrCb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram (8 bins)</td>
<td>75%</td>
<td>78%</td>
<td>73%</td>
</tr>
<tr>
<td>Histogram with PCA (256 bins)</td>
<td>77%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Color Moment</td>
<td>73%</td>
<td>77%</td>
<td>80%</td>
</tr>
</tbody>
</table>

The shortcoming of the experiment is the paucity of data. There are reasons to believe that with sufficient data, the accuracy will be further improved.
6. Conclusions

In this paper, we propose a fully convolutional network that leverages lightweight components such as InteractionModule, depth separable convolution and DilatedGroup to solve the real-time semantic segmentation problem in order to achieve a balance between speed and performance. We further apply it to hair and skin segmentation tasks, and extensive experiments confirm the effectiveness of the proposed method. Moreover, based on the segmented skin regions, we introduce color moments to extract color features and then classify the skin tones. 80% classification accuracy demonstrate the effectiveness of the proposed solution.

The aim of this work is to apply our algorithms to real-time coloring, face swapping, skin tone rating systems, and skin care product recommendations based on skin tone ratings in real-life scenarios. In our future work, we will investigate semi-supervised methods to address the lack of data volume.

Author Contributions: Conceptualization, X.F.; methodology, X.F. and L.L.; software, X.F. and L.L.; validation, X.F. and L.L.; formal analysis, X.F.; investigation, X.F.; resources, X.G.; data curation, L.L.; writing—original draft preparation, X.F.; writing—review and editing, X.G.; visualization, L.L.; supervision, X.G.; project administration, X.G.; funding acquisition, X.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China grant number 61573087 and 61573088.

Acknowledgments: This work is done when L.L. was an intern at Meidaojia Research, Beijing, China. Thanks their support for this work.

Conflicts of Interest: The authors declare no conflict of interest.

References


Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.