Face Parsing with RoI Tanh-Warping

Jinpeng Lin[1*]   Hao Yang[2*]   Dong Chen[2]   Ming Zeng[1†]   Fang Wen[2]   Lu Yuan[2]

[1]Software School of Xiamen University

[2]Microsoft Research

jplinforever@gmail.com zengming@xmu.edu.cn

{haya, doch, fangwen, luyuan}@microsoft.com

Abstract

Face parsing computes pixel-wise label maps for different semantic components (e.g., hair, mouth, eyes) from face images. Existing face parsing literature have illustrated significant advantages by focusing on individual regions of interest (RoIs) for faces and facial components. However, the traditional crop-and-resize focusing mechanism ignores all contextual area outside the RoIs, and thus is not suitable when the component area is unpredictable, e.g. hair. Inspired by the physiological vision system of human, we propose a novel RoI Tanh-warping operator that combines the central vision and the peripheral vision together. It addresses the dilemma between a limited sized RoI for focusing and an unpredictable area of surrounding context for peripheral information. To this end, we propose a novel hybrid convolutional neural network for face parsing. It uses hierarchical local based method for inner facial components and global methods for outer facial components. The whole framework is simple and principled, and can be trained end-to-end. To facilitate future research of face parsing, we also manually relabel the training data of the HELEN dataset and will make it public. Experiments on both HELEN and LFW-PL benchmarks demonstrate that our method surpasses state-of-the-art methods.

1. Introduction

Given an input face image, face parsing assigns a pixel-wise label for each semantic component, *e.g.*, hair, facial skins, eyes, nose, mouth and *etc.*. Compared with face alignment, face parsing can provide more precise areas, and more importantly, face parsing can output the hair area, which is necessary for a variety of high level applications, such as face understanding, editing and animation.

---

*Equal contribution. This work is done when Jinpeng Lin is an intern at Microsoft Research Asia.

†Corresponding author.

Figure 1. Our face parsing results. The proposed method is able to segment facial components as well as whole hair regions. It can also distinguish people who are close. Different brightness represent different individuals. (Best viewed in color)

Previous face parsing methods cannot handle hair segmentation very well. Some previous works [28, 14, 6] crop a fixed-size area around the face and ignore the surrounding area. These methods often fail to obtain the whole hair area. [17] propose using Fully Convolutional Networks (FCN) and Recurrent Propagation for face parsing, but it cannot distinguish the hair area of different people nearby. There are also some face parsing algorithms [31, 10, 20, 35] that just ignore the hair area due to the limitations of the algorithm or the data.

The difficulty of the segmentation of both the inner facial components and the hair at one time is as follows. Generally, it is possible to improve accuracy by predicting each face region first, and then predict the per-pixel mask in each region individually. The estimated region yields good hints, including spatial constraint and instance information, for predicting high accuracy masks. However, if only focusing on the center of the face, the hair in the surrounding area will be ignored. This is a problem because we need an accurate and tight location of each face for inner facial components while requiring the surrounding context for the hair region.

Our method is inspired by the human vision. When looking at something, human combine information of central vi-

sion and peripheral vision [13]. While central vision covers narrow degrees of the visual field straight ahead of us, peripheral vision covers the rest. Peripheral vision is not as sharp as central vision, but has wider range and helps us detect events to the side, even when we're not looking in that direction. Based on this characteristic, we propose a novel RoI Tanh-warping operator which non-linearly maps the *whole image* into a fixed-size. It addresses the dilemma between fixed input size and the unpredictable area of hair while reserving the amplified resolution on important regions.

Then, we feed the warped face image into a neural network. We use different strategies to process the inner and outer components of the face. For the inner components, *e.g.* brows, eyes, nose and mouth, the network first predicts the bounding-box of each component, then maps the features of each component to a fixed size through RoI align [9], a subnetwork is adopted to get the segmentation mask for each component. For the outer components, *e.g.* hair, face region and background, we append a FCN [29] to predict their masks. Compared with existing local-based face parsing methods [20, 35, 17, 6], the proposed architecture reduces the computation cost greatly through sharing features and can be trained in an end-to-end way.

To the best of our knowledge, there are only a few public face parsing datasets, such as the HELEN [27] and LFW part label (LFW-PL) [11]. However, LFW-PL only labels 3 classes and the labeling of the HELEN's training data is not very precise, especially for hair. To facilitate future research of face parsing, we manually relabel the training data of HELEN. New labels are more accurate. We will publicize the new labellings, and hope it will attract more researcher to the topic of face parsing.

Without bells and whistles, our network surpasses previous state-of-the-art results on HELEN [27] (trained on both new and old labellings) and LFW part label (LFW-PL) [11] datasets. We summarize our contributions as follows:

1. We propose a novel method to address the problem of face parsing, for inner facial parts and hair. To the best of our knowledge, it is the first attempt to jointly solve the strongly related and severely imbalanced parts together, efficiently leveraging the spatial relationships between different parts.

2. We introduce a new *RoI Tanh-warping* operation, which emphasizes the central face while retaining the peripheral parts (*i.e.*, surrounding context) of the face, addressing the dilemma between fixed input size and unpredictable area of hair.

3. We devise a neural network which integrates a Mask R-CNN-fashion [9] branch and a FCN-fashion [19] branch to treat inner and outer facial parts, respec-

tively. The hybrid method is simple, principled and can be trained end-to-end.

4. The results of our method surpasses all previous methods significantly on HELEN and LFW-PL dataset. We also relabel the original HELEN dataset (mislabeled much) and plan to publicize it to facilitate future research.

2. Related Work

Semantic Segmentation Semantic segmentation for generic images has become a fundamental topic in computer vision, and achieved significant progress, e.g.[7, 19, 22, 3, 4, 33, 5, 21, 12, 9]. FCN [19] is a well-known baseline for generic images which employs full convolution on the entire image to extract per-pixel feature. Following this work, CRFasRNN [33] and DeepLab [4] adopt dense CRF optimization to refine the predicted label map. Hayder et al. [8] represent the segmentation mask as a truncated distance transform to alleviate the information loss caused by erroneous box cropping. Recently, Mask R-CNN [9] further advances the cutting edge of semantic segmentation through extending Faster R-CNN [23] and integrating a novel RoIAlign. However, directly applying these generic methods for face parsing may fail to model the complex-yet-varying spatial layout across face parts, especially hair, leading to unsatisfactory results.

Face Parsing Most existing approaches for face parsing can be categorized into two groups: global-based and local-based methods.

Global-based methods directly predict per-pixel semantic label over the whole face image. Early works represent spatial correlation between facial parts by various designed models, such as the epitome model [30] and the exemplar-based method [27]. With the advance of deep learning techniques, a variety of CNN structures and loss functions are proposed to encode the underlying layouts of the whole face. Liu et al. [18] integrate the CNN into the CRF framework, and jointly model pixel-wise likelihoods and label dependencies through a multi-objective learning method. Jackson et al. [10] use facial landmarks as the guidance, and integrate boundary cue into CNN to implicitly confine facial regions. Zhou et al. [34] design an architecture which employs fully-convolutional network, super-pixel information, and CRF model jointly. Wei et al. [31] propose automatically regulating receptive fields in a deep image parsing network, therein obtaining better receptive fields for facial parsing. Besides these works, Saito et al. [24] try to reduce computation to achieve real-time performance.

These kinds of methods treat face parts globally and inherently integrate the prior of the face layout. Nevertheless, accuracy is limited due to the lack of focusing on each individual part.
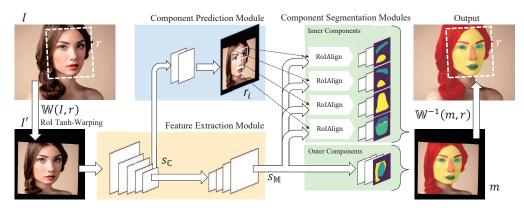
Figure 2. Proposed network structure. Given an input image I, we retreive a face-aligned rectangle r, and apply RoI Tanh-warping $\mathbb{W}$ to retrieve a distorted and face-aligned image I'. We propose the *feature extraction* module to extract feature maps from I'. For inner facial components (*e.g.*, eye, nose, mouth and *etc*.), we use a *component prediction* module to regress their local bounding boxes and use RoI align to extract their local features. For outer facial components (*i.e.*, face, hair, background), we directly use the global feature. Then, for each component, we use an individual *component segmentation module* to predict its segmentation scores though convolutions. The results are gathered as m and warped back to the input domain as $\mathbb{W}^{-1}(m, r)$.

Local-based methods train separated models for various facial components (*e.g.* eyes, nose *etc*.) to predict masks for each part individually. Luo et al. [20] propose a hierarchical method which segments each detected facial part separately. Zhou et al. [35] design an interlinked CNN-based pipeline which predicts pixel labels after facial localization. Benefiting from the complicated design, the interlinked CNN structure is able to pass information between coarse and fine levels bidirectionally, thus achieving good performance at the expense of large memory and computation consumption. Liu et al. [17] achieve state-of-the-art accuracy with very fast running speed by combining a shallow CNN and a spatially variant RNN in two successive stages.

These local-based methods almost adopt the coarse-to-fine strategy, taking into account both global consistency and local precision. However, different from our proposed method, the previous methods model the two stages separately, without pursuing the gains of accuracy and efficiency from backbone sharing and joint optimization.

Portrait/Hair Segmentation Portrait segmentation and hair segmentation, such as the works of Shen et al. [25, 26] and Chai et al. [1], to name a few, are closely related to the literature of face parsing. Recent approaches for these two tasks adopt knowledge of specific domains into DCNN and achieve practical results for following up applications. Nevertheless, they only tackle a sub-problem of face parsing, without addressing the task of segmenting all parts on the face, while the latter is more general and challenging.

3. Method

We introduce the network with *RoI Tanh-warping* for face parsing. Given a face image I of varied dimensions,

we first utilize a face and landmarks detector [2] to retrieve five landmark points and construct a face rectangle r based on the points. Then, we warp the whole image into fixed-sized $\mathbb{W}(I, r)$ with RoI Tanh-warping, where the face is aligned and centered within the output. The proposed RoI Tanh-warping operator $\mathbb{W}$ focuses on the aligned target face without loosing peripheral information.

We use a hybrid solution to separately estimate masks for inner and outer facial components. For each inner facial component $\{P_i\}_{i=1}^N$ (*e.g.*, eye, nose, mouth and *etc*.), where N is the number of individual component, we predict the local rectangle $\{r_i\}_{i=1}^N$ bounding each component P_i in the warped image. Then we extract local features within each rectangle r_i through RoI align operators [9]. The pixel-wise segmentation scores $\{m_i\}_{i=1}^N$ for each component P_i are predicted individually. For the outer facial components (*i.e.*, hair, facial skin and background), we directly use the global feature $s_{\mathbb{M}}$ to predict the pixel-wise segmentation scores m_{outer}. Finally, we gather all scores denoted as m, and warp them back to the original input domain as $\mathbb{W}^{-1}(m, r)$ to get the final labels.

The whole pipeline of our method is shown in Figure 2. Our framework consists of four major components: 1) the *RoI Tanh-warping* operator $\mathbb{W}$, which warps input face image with varied dimensions into the face aligned and uniformly shaped domain; 2) the backbone *feature extraction* module $\mathbb{F}$, which extracts informative features from the warped image for subsequent operations; 3) the *component prediction* module $\mathbb{R}$, which predicts the bounding rectangles of inner facial components; 4) the multiple *component segmentation* modules $\{\mathbb{M}_i\}$, which extract features from the predicted rectangles using RoI align operators for inner facial components or directly through up-samplings for outer facial components, and predict the pixel-wise segmen-
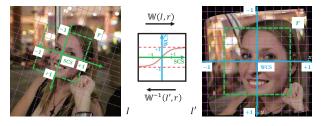
Figure 3. Essentially, Tanh-warping performs a $\tanh$ shaped distortion on x and y axes between two coordinate systems: the *source coordinate system* (SCS, in green color) and the *warped coordinate system* (WCS, in blue color) so that point located at (x, y) in SCS in I is warped to $(\tanh(x), \tanh(y))$ in WCS in I'.

tation scores.

Unlike [20, 35], which separate the component prediction and segmentation into two individual steps, we share the common feature computation for component prediction and segmentation, which are jointly optimized and yield efficient computation. Next, we introduce each part in detail.

3.1. RoI Tanh-Warping

Previous face parsing works usually ignore the hair area, *e.g.* [31, 20, 10, 35], or only consider hair within a certain area, *e.g.* [27, 18, 15, 14]. This is because the area of the hair varies greatly, and the input to the network is usually a fixed-size picture of the face in the center. Fully Convolutional Networks (FCN) [19] can handle various input size. Unfortunately, the original FCN performing on the whole image treats the face globally. Its accuracy is limited due to lack of focusing on each individual part, for example, producing fragmentary masks for face parts.

To solve this problem, we propose RoI Tanh-warping, which maps the *whole image of any size* into a limited view using a guiding rectangle. For convenience, we define two new coordinate systems, a *source coordinate system* and a *warped coordinate system*.

Source Coordinate System. Given an image I, we first apply a face and landmarks detector [2] to get five landmarks representing the eyes, the nose and two corners of the mouth. We estimate a 2D similar transformation T that transforms five detected landmarks to match five predefined template points[1]. The face rectangle r is then determined by the four corner points computed by $T^{-1}(\pm 1, \pm 1)$. With the rectangle r in I, a local coordinate system can be determined: its origin is the center of r, its $x = \pm 1$ and $y = \pm 1$ lines coincide with r's borders. The source coordinate system is shown in the left image of Figure 3.

Warped Coordinate System. Let I' be the warped image, the warped coordinate system is defined as: its origin is the center of I', while its $x = \pm 1$ and $y = \pm 1$ lines determine the borders of I', as shown in right image of Figure 3.

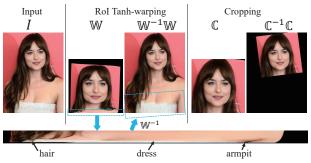[1]Predefined as $\{(-0.25, -0.1), (0.25, -0.1), (0, 0.1), (-0.15, 0.4), (0.15, 0.4)\}$.

Figure 4. Tanh-warping vs. cropping. The Tanh-warping $\mathbb{W}$ is better invertible than cropping $\mathbb{C}$ on peripheral area.

Suppose a pixel p in I is warped to p' in I' using RoI Tanh-warping, then its original coordinates (x, y) under the source coordinate system and the new coordinates (x', y') under the warped coordinate system satisfy:

$$x' = \tanh(x), \quad y' = \tanh(y). \tag{1}$$

We denote the RoI Tanh-warping as $\mathbb{W}(I, r)$. It is implemented as a bilinear sampling according to Equation 1. In theory, $\mathbb{W}$ is invertible. Let $\mathbb{W}^{-1}$ be the inverse operator of $\mathbb{W}$. We show results of $\mathbb{W}$ and $\mathbb{W}^{-1}\mathbb{W}$ on a face image I in Figure 4. The proposed $\mathbb{W}(I, r)$ contains almost all information of the original image I, including the hair, dress and armpit in the bottom. It is observed that: i) although some fine details may be lost during the warping, the whole shape of hair can be basically recovered through $\mathbb{W}^{-1}$; ii) the RoI Tanh-warping operator preserves more linearity near the center of the rectangle r, but imposes more distortion on the peripheral pixels (*i.e.* surrounding context).

Most previous face parsing methods [17, 34, 6] apply face alignment via cropping. It is hence infeasible for these approaches to predict the labels (*e.g.*, hair) on peripheral pixels distant from the face, as shown in Figure 4. The proposed RoI Tanh-warping addresses the dilemma between fixed input size and unpredictable area of hair while reserving the focusing on the RoI.

3.2. Feature Extraction

Given the warped face image $\mathbb{W}(I, r)$, which contains only a single face in the center of the image, the *feature extraction* module $\mathbb{F}$ is deployed to capture implicit features shared by multiple tasks. $\mathbb{F}$ uses the ResNet-18 and FPN [16] structure as the backbones. FPN helps achieve a higher spatial resolution through layer-by-layer feature upsampling since the output feature map from C4 (output of the conv4 block of the ResNet-18) has very low resolution, which is too coarse for accurate segmentation. We can get the feature maps $s_\mathbb{R}$ and $s_\mathbb{M}$ for component prediction and segmentation respectively by

$$(s_\mathbb{R}, s_\mathbb{M}) = \mathbb{F}(x, \theta_\mathbb{F}), \tag{2}$$

where $\theta_\mathbb{F}$ denotes the parameters of $\mathbb{F}$, including the parameters of the ResNet and the FPN. $s_\mathbb{R}$ represents the C4 layer

Figure 5. Directly applying Mask R-CNN in face parsing causes problems: the left brows and left eyes are recognized as the right brows/eyes. The region proposal network that Mask R-CNN relies on misclassified instances that share similar appearance but have different semantic labels. The box regression we use is more straight-forward but effective for parsing facial components.

of ResNet-18. $s_\mathbb{M}$ represents the last layer of FPN.

3.3. Component Prediction

The *component prediction* module $\mathbb{R}$ locates bounding rectangles of the N inner facial components: $\{r_i\}_{i=1}^N$. It directly outputs the coordinates of each rectangle, namely

$$\{r_i\}_{i=1}^N = \mathbb{R}(s_\mathbb{R}, \theta_\mathbb{R}), \tag{3}$$

where N is the number of inner facial components, and $\theta_\mathbb{R}$ is the parameters of $\mathbb{R}$. The *component prediction* module $\mathbb{R}$ consists of two convolutional layers followed by a global average pooling and a fully connected layer. We consider a tight rectangle surrounding the annotated component mask as the ground-truth bounding-box, denoted as $\{r_i^g\}_{i=1}^N$. We adopt the L_1 loss for the bounding-box regression:

$$\mathcal{L}_{comp} = \frac{1}{N} \sum_{i=1}^N \|r_i - r_i^g\|_1. \tag{4}$$

Here we explicitly regress the area of each component instead of detecting them individually like in Mask R-CNN. The semantic label of every predicted facial component is explicitly defined. It avoids ambiguities in components and reduces computation cost, as shown in Figure 5.

3.4. Component Segmentation

We proposed a hybrid structure for component segmentation. For each inner facial component (eyes, brows, nose, mouth), we use RoI align to extract the local feature and predict its masks individually. The estimated region in previous step yields good hints for predicting high accuracy masks. For the outer facial components (face, hair), we use FCN to directly predict the segmentation mask in the warped face. It can handle unpredictable area of hair.

Inner Facial components. For the obtained N bounding boxes, N light and parallel component segmentation modules $\{\mathbb{M}_i\}_{i=1}^N$ are applied to predict the pixel-wise masks for each inner facial component. All component segmentation modules share the same network architecture, but their weights are independent. Without loss of generality, we only discuss the ith segmentation network as follows.

Given a bounding-box r_i regressed from $\mathbb{R}$, we apply RoI-align operator to sample a local feature patch out from the feature map $s_\mathbb{M}$ and resize it to a squared size. We observe that sometimes the regressed bounding-box is not very

accurate: some areas may fall outside the box, especially for the wide open mouth. This may be caused by insufficient training data, but if it happens, the area outside the box will not be correctly segmented. So we add a padding outside the box to solve this problem. The RoI align feature l_i for the ith component is defined as

$$l_i = \texttt{RoIAlign}(s_\mathbb{M}, \texttt{Padding}(r_i)). \tag{5}$$

Then we perform several convolutions and up-sampling operations to generate a segmentation m_i of the ith component.

$$m_i = \mathbb{M}_i(l_i, \theta_{\mathbb{M}_i}) \tag{6}$$

where $\theta_{\mathbb{M}_i}$ represents the parameters of the ith segmentation module. We use the pixel-wise cross-entropy to measure the component segmentation accuracy. The segmentation loss $\mathcal{L}_{inner}$ is defined as the averaged cross-entropy among all the segmentation networks:

$$\mathcal{L}_{inner} = \frac{1}{N} \sum_{i=1}^N \texttt{CrossEntropy}(m_i, m_i^g), \tag{7}$$

where m_i^g is the ground-truth segmentation of the ith component.

Outer Facial Components. For the outer facial components (*i.e.*, hair, facial skin and background), we apply FCN to get the segmentation mask, as shown in Figure 2.

$$m_{outer} = \mathbb{M}_{outer}(s_\mathbb{M}, \theta_\mathbb{M}) \tag{8}$$

We also use the cross-entropy loss to constrain the segmentation accuracy:

$$\mathcal{L}_{outer} = \texttt{CrossEntropy}(m_{outer}, m_{outer}^g), \tag{9}$$

where m_{outer}^g is the ground-truth segmentation of the outer facial component.

Finally, all the resulting segmentation scores are gathered, denoted as m. We de-warp the scores to the original image domain as $\mathbb{W}^{-1}(m, r)$ using the same rectangle r from the input stage and form the final face parsing result.

3.5. Implementation Details

Since the component segmentation relies on a good component region estimation, we divide the training process into two stages. In the first stage, we only train the *feature extraction* module and the *component prediction* module for a good component regressing accuracy. Here, only the component loss $\mathcal{L}_{comp}$ (in Equation 4) is used for training. In the second stage, we perform joint training by updating all parameters, including $\theta_\mathbb{F}$, $\theta_\mathbb{R}$, and $\theta_\mathbb{M}$, with the component loss $\mathcal{L}_{comp}$ (in Equation 4) and the mask loss $\mathcal{L}_{inner}$ and $\mathcal{L}_{outer}$ (in Equation 7 and 9) together.

By default, the size of warped images I' is set to 512×512. We use ResNet-18 in *feature extraction*. We select the feature from C_4 of ResNet as $s_\mathbb{R}$ for component prediction, and the feature from P_2 of FPN as $s_\mathbb{M}$ for component segmentation. The spatial resolution of $s_\mathbb{R}$ is 32×32, the spatial resolution of $s_\mathbb{M}$ is 128×128. The *component prediction*

module consists of two $3 \times 3 \times 320$ convolutional layers, one $1 \times 1 \times 1280$ convolutional layer, one global average pooling and one fully connected layer. Its output is a tensor of $N \times 4$ that encodes the location of N bounding boxes, where N is the number of inner facial components. N varies for different databases. Before RoI align, regressed boxes are padded by 10% the feature map size for mouth and 5% otherwise. The output size of RoI align defined by Equation 5 is 32×32 for all inner components. Each *component segmentation* module is built with two $3 \times 3 \times 256$ convolutions each followed by one bilinear up-sampling if for inner component. A 1×1 convolution followed by softmax is appended to each *component segmentation* module to output the masks. The sizes of masks $\{m_i\}$ are all 128×128.

4. Experiments

We use HELEN [27] and LFW-PL [11] for the experiments. The HELEN dataset contains 2,330 face images. Each image is annotated with 11 labels: "background", "facial skin", "left/right brow", "left/right eye", "nose", "upper/lower lip", "inner mouth" and "hair". We adopt the same dataset division setting as in [18, 32, 31] that uses 2,000 images for the training, 230 images for the validation and 100 images for the testing. The LFW-PL dataset contains 2,972 face images. All of them are manually annotated with 3 labels: "skin", "hair" and "background". Following [11, 34], we use 1,500 images for the training, 500 images for the validation, and 927 images for the testing.

Due to the inaccurate annotations in HELEN, we re-annotate its "hair" and "facial skin" labels manually. We also reprocess the facial component labels according to the modified facial skin label. For fair comparison with the previous methods, we only re-annotate both training and validation sets, and the test set remains unchanged. We refer to this relabeled dataset as HELEN*. Figure 6 compares the annotations between HELEN and HELEN*.

We augment data during training: 1) randomly replace the background with non-face images or pure colors, 2) random rotation (within $[-18°, 18°]$), scaling (within $[0.9, 1.1]$) around the face center, 3) random horizontal flipping and 4) random gamma adjustment with $\gamma \in [0.5, 2]$.

Figure 6. HELEN vs. HELEN*. Obvious annotation errors of face and hair in HELEN are all rectified in HELEN*.

4.1. Comparison with State-of-the-art

We perform a thorough comparison between our model and existing state-of-the-art methods on HELEN and LFW-PL datasets. Results are measured by F-measure which is commonly used by existing face parsing literature. Our results are calculated using the original image sized annotations without any transformation or cropping.

We show the comparison results on HELEN in Table 1. Each column shows the F-measure percentage corresponding to a specific face label. *I-mouth* is short for inner mouth, *U/L-lip* is short for *upper/lower lip*, and *overall* represents a union of all inner facial component (eyes/brows/nose/mouth) labels. We report the results of our model trained on both HELEN and HELEN* in Table 1. The testing results on HELEN show that our method outperforms all state of the art methods. We also observe an improvement brought by training on the relabeled HELEN* dataset, especially on the skin and hair labels.

Methods	eyes	brows	nose	I-mouth	U-lip	L-lip	mouth	skin	hair	overall
Smith *et al.* [27]	78.5	72.2	92.2	71.3	65.1	70.0	85.7	88.2	-	80.4
Zhou *et al.* [35]	87.4	81.3	95.0	83.6	75.4	80.9	92.6	-	-	87.3
Liu *et al.* [18]	76.8	71.3	90.9	80.8	62.3	69.4	84.1	91.0	-	84.7
Liu *et al.* [17]	86.8	77.0	93.0	79.2	74.3	81.7	89.1	92.1	-	88.6
Wei *et al.* [31]	84.7	78.6	93.7	-	-	-	91.5	91.5	-	90.2
Ours (trained on HELEN)	89.6	83.1	95.6	86.7	79.6	**89.8**	95.0	94.5	83.5	92.4
Ours (trained on HELEN*)	**89.7**	**85.9**	**95.6**	**86.7**	**80.8**	89.7	**95.2**	**95.3**	**88.7**	**93.1**

Table 1. Comparison with state-of-the-art methods on HELEN.

We also show the comparison between our model and existing face parsing methods on LFW-PL in Table 2. We report the F-measure percentages corresponding to skin, hair and background. We compare the overall accuracies as well. Our method also surpasses state of the art on the LFW-PL dataset. The improvement of our method is relatively small, since all images in LFW-PL have been aligned and cropped, only one face is in the middle of the image.

Methods	skin	hair	bg	accuracy
Liu *et al.* [18]	93.93	80.70	97.10	95.12
Long *et al.* [19]	92.91	82.69	96.32	94.13
Chen *et al.* [4]	92.54	80.14	95.65	93.44
Chen *et al.* [3]	91.17	78.85	94.95	92.49
Zhou *et al.* [34]	94.10	85.16	96.46	95.28
Liu *et al.* [17]	**97.55**	83.43	94.37	95.46
Ours	95.77	**88.31**	**98.26**	**96.71**

Table 2. Comparison with state-of-the-art methods on LFW-PL.

4.2. Ablation Study

To understand the role of the hybrid network structure and the proposed RoI Tanh-warping, we conduct several baseline methods for comparisons. We substitute certain submodules of our proposed framework with alternatives and construct four network structures, which are:

FCN: An FCN structure without the component prediction module $\mathbb{R}$. All segmentation modules $\mathbb{M}_i$ directly take $s_{\mathbb{M}}$ as the input feature without using RoI-align.

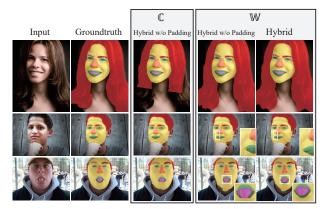| input	network structure	eyes	brows	nose	I-mouth	U-lip	L-lip	mouth	skin	hair	overall
rescale	FCN	77.5	66.0	69.7	71.4	62.7	68.4	79.9	80.3	82.7	73.0
ℂ	FCN	82.7	79.6	93.7	86.3	78.1	85.2	92.5	94.4	85.0	89.4
ℂ	Hybrid w/o Padding	88.6	83.9	94.5	**87.6**	79.4	89.7	94.5	95.1	84.6	92.0
ℂ	Hybrid	89.7	84.7	95.5	86.9	80.3	90.7	**95.4**	95.2	84.5	92.9
ℂ2	Hybrid	85.9	83.3	94.4	84.1	73.6	85.5	92.5	91.7	83.8	90.9
𝕎	FCN	82.6	79.3	93.8	85.7	77.0	84.4	92.4	94.0	**88.8**	89.3
𝕎	Hybrid w/o Padding	88.1	84.7	94.6	87.4	76.7	89.6	94.1	95.0	88.4	91.9
𝕎	Hybrid w/ WeightSharing	89.6	85.1	95.6	85.4	75.8	89.2	94.3	94.9	88.1	92.7
𝕎	Hybrid w/ EyeBrowSymmetry	89.7	85.8	95.5	86.8	80.5	89.6	95.1	95.3	88.5	93.0
𝕎	Hybrid (Ours)	**89.7**	**85.9**	**95.6**	86.7	**80.8**	**89.7**	95.2	**95.3**	88.7	**93.1**

Table 3. Comparison of ablation models trained on HELEN*.

Figure 7. Qualitative comparisons on HELEN dataset. The proposed Tanh-warping 𝕎 addresses the hair cropping issues; the padding in our hybrid structure is also necessary when box regression is not accurate.

Hybrid w/o Padding: A hybrid structure with component prediction module applied to regress bounding boxes of inner facial components. Segmentation modules for inner components all take the RoI-aligned features as $\text{RoIAlign}(s_\mathbb{M}, r_i)$, but without any padding on the bounding boxes r_i.

Hybrid: The proposed hybrid structure. Segmentation modules for inner components take RoI-aligned features as inputs. The bounding boxes for RoI-align are padded to better tolerate box regression errors following Equation 5.

Hybrid w/ WeightSharing: A hybrid structure similar to our proposed model, except that weight sharing is applied to all its segmentation modules for inner components (except for the last 1×1 convolution before soft-max). We use this baseline model to simulate the mask heads of Mask R-CNN.

Hybrid w/ EyeBrowSymmetry: The parameters $\theta_{\mathbb{M}_i}$ are shared between branches of left/right eyes and branches of left/right brows. Specifically, features l_i cropped from Equation 5 for the *left eye* and *left brow* labels are horizontally flipped before forwarding to the component segmentation modules $\mathbb{M}_i$. In return, the output segmentations m_i from Equation 6 for the *left eye* and *left brow* labels are flipped back to compose the final prediction.

We combine these network structures with different techniques that help focus the network on the input: 1) cropping the image within the face rectangle r, denoted by ℂ;

2) Tanh-warping the image guided by r, denoted by 𝕎; 3) rescaling all the input images to 512×512 (with padding to preserve the aspect ratios), denoted by rescale.

Eight baseline models are listed by Table 3, including the proposed model 𝕎+*Hybrid*. All the models are trained on HELEN*. We compare the proposed model with other baseline methods. Some visual results are compared in Figure 7. It shows the improvement brought by the Tanh-warping operator in parsing hairs, and the advantage of our hybrid network structure in parsing inner facial components. From these results, we can draw some conclusions:

Importance of Tanh-warping: Models that apply cropping ℂ or Tanh-wrapping 𝕎 using face rectangles consistently outperform the baseline rescale+*FCN* model (Table 3 Row 1) on all labels. The key factor is that both ℂ and 𝕎 operators can allow the network to focus on a smaller but more important region by amplifying its resolution. It evidences the effectiveness of the *central vision*. It is not guaranteed, however, that the central vision always covers all true regions. For example, using ℂ on an input image may crop off hairs distant from the face, as shown by Figure 4. By combining the peripheral vision with the central vision together, Tanh-warping 𝕎 brings significant improvements in F-measure on hairs over ℂ (88.8 vs. 85.0).

Cropping a larger region seems to be a reasonable alternative to Tanh-warpping since it increases the hair coverage of the network input. However, it conversely decreases the resolution of the face, causing the segmentation performance to drop. We add a result called ℂ2 with 2 times larger cropping area in Table 3. ℂ2 is even worse than ℂ for either outer components (*hair* 83.8 vs. 84.5) or inner components

Figure 8. Our method can handle hairs with various lengths.

Figure 9. Qualitative results on challenging images. Adjacent individuals are distinguished by different brightnesses. (Best viewed in color)

(*overall* 90.9 vs. 92.9). Therefore, a larger crop region is not a good solution to the limited region problem caused by cropping.

Importance of Hybrid Structure: The proposed hybrid structure combines a FCN with a Mask R-CNN-like structure. The Mask R-CNN part adopts the central vision for inner facial components by applying RoI-align in the feature level. Table 3 illustrates the advantages of *Hybrid* structures over pure *FCN* structures on the overall scores for inner facial components (92.0 vs. 89.4, 91.9 vs. 89.3). Padding the regressed bounding boxes for RoI-align further improves the overall scores (92.9 vs. 92.0, 93.1 vs. 91.9). Meanwhile, we observe that our *Hybrid* structure does not affect the hair scores much compared with *FCN* (88.7 vs. 88.8).

Importance of Separated Segmentation Modules: Our hybrid network consists of multiple segmentation modules for different inner facial components. Different from the mask heads of Mask R-CNN, our segmentation modules do not share weights. The importance of separated weights is verified by the results from Table 3, which shows that sharing weights across all component segmentation modules for inner facial components reduces the overall accuracy (92.7 vs. 93.1). The eye/brow symmetric weight sharing is comparable to separated weights (*eyes* 89.7 vs. 89.7; *brows* 85.8 vs. 85.9; *overall* 93.0 vs. 93.1). We also train an official Mask R-CNN on HELEN* where facial components are detected through a region proposal network. As explained in Section 3.3, it does not perform well for face parsing, which is verified by the scores (*hair* 78.5, *overall* 84.9).

4.3. More Results

Varying hair lengths: Due to the nonlinear rescaling ability of Tanh-warping, as shown in Figure 8, our method is suitable for segmenting hairs with various lengths.

In-the-wild and multi-face conditions: Figure 9 shows vi-

sual results on challenging images from HELEN datasets and Internet. Although our model is trained on HELEN* dataset, it shows the ability to handle large pose, expression variations, occlusion and multiple closely-snuggled faces in the wild condition. More concretely, for the multi-face scenario, we first detect 5-point landmarks for each face appearing in the image, then apply the RoI Tanh-warping and hybrid network for each face independently, and de-warp its softmax activation maps onto the original image through bilinear mapping. Finally, on each pixel, its instance is determined by selecting the maximal foreground activation score among all faces. These results in Figure 9 show that our method is capable of distinguishing different face instances.

Efficiency: The proposed network is efficient. It runs at 40ms per face on Nvidia Titan Xp GPU.

5. Conclusion

We propose a novel hybrid network combined with RoI Tanh-warping for face parsing with hairs. We use RoI Tanh-warping to align the face in the middle while preserving the peripheral context for parsing hairs. Our hybrid network applies a Mask R-CNN-fashion branch for inner components (eyes, brows, nose, mouth), while applying a FCN-fashion branch for outer components (face, hair). Ablation studies show the effectiveness of RoI Tanh-warping and our hybrid structure. The superior performances on public datasets HELEN/LFW-PL and in-the-wild images show the ability of our method to handle the problem of face parsing with hairs under various environments.

Acknowledgments

References

[1] Menglei Chai, Tianjia Shao, Hongzhi Wu, Yanlin Weng, and Kun Zhou. Autohair: Fully automatic hair modeling from a single image. *ACM Trans. Graph.*, 35(4):116:1–116:12, July 2016. 3

[2] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision*, pages 109–122. Springer, 2014. 3, 4

[3] Liang-Chieh Chen, Jonathan T Barron, George Papandreou, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with task-specific edge detection using CNNs and a discriminatively trained domain transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4545–4554, 2016. 2, 6

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *arXiv preprint arXiv:1606.00915*, 2016. 2, 6

[5] Lianli Gao, Jingkuan Song, Feiping Nie, Fuhao Zou, Nicu Sebe, and Heng Tao Shen. Graph-without-cut: An ideal graph learning for image segmentation. In *AAAI*, volume 2, page 6, 2016. 2

[6] Umut Güçlü, Yağmur Güçlütürk, Meysam Madadi, Sergio Escalera, Xavier Baró, Jordi González, Rob van Lier, and Marcel AJ van Gerven. End-to-end semantic face segmentation with conditional random fields as convolutional, recurrent and adversarial networks. *arXiv preprint arXiv:1703.03305*, 2017. 1, 2, 4

[7] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision (ECCV)*, 2014. 2

[8] Zeeshan Hayder, Xuming He, and Mathieu Salzmann. Boundary-aware instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 587–595, 2017. 2

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 2, 3

[10] Aaron S Jackson, Michel Valstar, and Georgios Tzimiropoulos. A CNN cascade for landmark guided semantic part segmentation. In *European Conference on Computer Vision (ECCV)*, pages 143–155. Springer, 2016. 1, 2, 4

[11] 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*, pages 2019–2026, 2013. 2, 6

[12] Suha Kwak, Seunghoon Hong, Bohyung Han, et al. Weakly supervised semantic segmentation using superpixel pooling network. In *AAAI*, pages 4111–4117, 2017. 2

[13] Jerome Y Lettvin. On seeing sidelong. *The Sciences*, 16(4):10–20, 1976. 2

[14] Jianshu Li, Shengtao Xiao, Fang Zhao, Jian Zhao, Jianan Li, Jiashi Feng, Shuicheng Yan, and Terence Sim. Integrated face analytics networks through cross-dataset hybrid training. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1531–1539. ACM, 2017. 1, 4

[15] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 6, 2017. 4

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 4, 2017. 4

[17] Sifei Liu, Jianping Shi, Ji Liang, and Ming-Hsuan Yang. Face parsing via recurrent propagation. *arXiv preprint arXiv:1708.01936*, 2017. 1, 2, 3, 4, 6

[18] 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*, pages 3451–3459. IEEE, 2015. 2, 4, 6

[19] 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*, pages 3431–3440, 2015. 2, 4, 6

[20] Ping Luo, Xiaogang Wang, and Xiaoou Tang. Hierarchical face parsing via deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2480–2487. IEEE, 2012. 1, 2, 3, 4

[21] Yulei Niu, Zhiwu Lu, Songfang Huang, Xin Gao, and Ji-Rong Wen. Feaboost: Joint feature and label refinement for semantic segmentation. In *AAAI*, pages 1474–1480, 2017. 2

[22] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015. 2

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2

[24] Shunsuke Saito, Tianye Li, and Hao Li. Real-time facial segmentation and performance capture from rgb input. In *European Conference on Computer Vision*, pages 244–261. Springer, 2016. 2

[25] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian Price, Eli Shechtman, and Ian Sachs. Automatic portrait segmentation for image stylization. In *Proceedings of the 37th Annual Conference of the European Association for Computer Graphics*, EG '16, pages 93–102, Goslar Germany, Germany, 2016. Eurographics Association. 3

[26] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *European conference on computer vision*. Springer, 2016. 3

[27] Brandon M Smith, Li Zhang, Jonathan Brandt, Zhe Lin, and Jianchao Yang. Exemplar-based face parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3484–3491. IEEE, 2013. 2, 4, 6

[28] Stavros Tsogkas, Iasonas Kokkinos, George Papandreou, and Andrea Vedaldi. Deep learning for semantic part

segmentation with high-level guidance. *arXiv preprint arXiv:1505.02438*, 2015. 1

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 2

[30] Jonathan Warrell and Simon JD Prince. Labelfaces: Parsing facial features by multiclass labeling with an epitome prior. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 2481–2484. IEEE, 2009. 2

[31] Zhen Wei, Yao Sun, Jinqiao Wang, Hanjiang Lai, and Si Liu. Learning adaptive receptive fields for deep image parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2434–2442, 2017. 1, 2, 4, 6

[32] Takayoshi Yamashita, Takaya Nakamura, Hiroshi Fukui, Yuji Yamauchi, and Hironobu Fujiyoshi. Cost-alleviative learning for deep convolutional neural network-based facial part labeling. *Transactions on Computer Vision and Applications*, 7:99–103, 2015. 6

[33] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P.H.S. Torr. Conditional random fields as recurrent neural networks. In *In Computer Vision (ICCV), 2015 IEEE International Conference on*, 2015. 2

[34] Lei Zhou, Zhi Liu, and Xiangjian He. Face parsing via a fully-convolutional continuous CRF neural network. *arXiv preprint arXiv:1708.03736*, 2017. 2, 4, 6

[35] Yisu Zhou, Xiaolin Hu, and Bo Zhang. Interlinked convolutional neural networks for face parsing. In *International Symposium on Neural Networks*, pages 222–231. Springer, 2015. 1, 2, 3, 4, 6