A New Dataset and Boundary-Attention Semantic Segmentation for Face Parsing

Yinglu Liu, Hailin Shi, Hao Shen, Yue Si, Xiaobo Wang, Tao Mei
JD AI, Beijing, China
{liuyinglu1, shihailin, shenhao, siyue, wangxiaobo8, tmei}@jd.com

Abstract

Face parsing has recently attracted increasing interest due to its numerous application potentials, such as facial make up and facial image generation. In this paper, we make contributions on face parsing task from two aspects. First, we develop a high-efficiency framework for pixel-level face parsing annotating and construct a new large-scale **La**ndmark guided face **Pa**rsing dataset (LaPa). It consists of more than 22,000 facial images with abundant variations in expression, pose and occlusion, and each image of LaPa is provided with an 11-category pixel-level label map and 106-point landmarks. The dataset is publicly accessible to the community for boosting the advance of face parsing.[1] Second, a simple yet effective **B**oundary-**A**ttention **S**emantic **S**egmentation (BASS) method is proposed for face parsing, which contains a three-branch network with elaborately developed loss functions to fully exploit the boundary information. Extensive experiments on our LaPa benchmark and the public Helen dataset show the superiority of our proposed method.

Introduction

Face parsing, aiming to assign pixel-level semantic labels for facial images, has attracted more and more attentions due to its wide application potentials, such as facial make up (Ou et al. 2016) and facial image synthesis (Zhang et al. 2018). In recent years, deep learning promotes the development of face related fields, such as face recognition (Sun et al. 2014; Taigman et al. 2014; Schroff, Kalenichenko, and Philbin 2015; Deng et al. 2019; Fu et al. 2019), face detection (Zafeiriou, Zhang, and Zhang 2015; Zhang et al. 2017) and face alignment (Zhou et al. 2013; Wu et al. 2017; Merget, Rock, and Rigoll 2018). However, the development of face parsing remains slow. One of the major obstacle is the lack of training data. As is well known, adequate training data is crucial for achieving good results by deep learning methods. However, there are few public datasets for face parsing due to the difficulty and high cost of pixel-level annotation. By contrast, labeling a small number of predefined landmarks in a facial image is much easier, and it is a feasible way to utilize these landmarks to guide the face parsing annotation. Nevertheless, the public datasets for facial landmark localization are limited either on the number of training samples or the number of landmarks. Specifically, most of public available datasets are annotated with less than 100 points, which are not enough to depict the shape of facial parts with fine details. For example, the widely used 68-point landmarks in 300W (Sagonas et al. 2013) describe the eyebrow with only 5 points on the upper boundary while leaving the lower boundary unmarked. The recent 98-point landmarks in WFLW (Wu et al. 2018) do not cover the regions of nose wing. Needless to say other makeups such as 21-point in AFLW (Koestinger et al. 2011) and 6-point in AFW (Baltrusaitis, Robinson, and Morency 2013), they could be applied for face geometric normalization but are incompetent to represent boundaries of facial parts. The He-

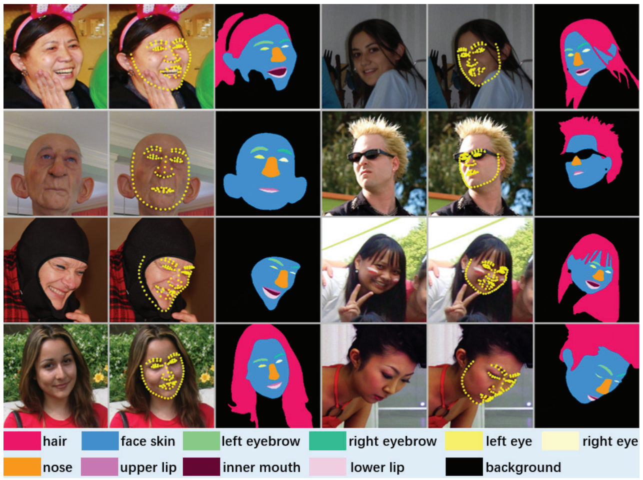


Figure 1: Annotation examples of the proposed LaPa dataset. It consists of more than 22,000 images with large variations in pose, facial expression and occlusion. Each image is provided by 106-point landmarks and an 11-category semantic label map. One can refer to the supplementary material for more examples.

[1] https://github.com/lucia123/lapa-dataset

len dataset (Le et al. 2012) contains 194-point landmarks, but the number of samples is only 2,330 and no landmarks are located on the nose bridge.

To remedy the above problems, in this paper we develop a high-efficiency framework for face parsing annotating. It is composed of two consecutive modules - Dense Landmark Annotation (DLA) module and Pixel-Level Parsing Annotation (PPA) module, both of which simplify the annotation at their respective stages. In the DLA module, we develop a semi-automatic labeling tool for 106-points facial landmark annotation. This tool could greatly reduce the workload for annotators with the help of an iteratively updated auxiliary model. Subsequently, with the color image and the annotated landmarks, the pixel-level semantic labels are produced automatically in the PPA module, no need of manual work. It is accomplished by three steps. First, we propose a category-wise fitting approach, which could draw the contour for each facial part based on the landmarks. Second, a coarse-to-fine segmentation strategy is employed to segment the hair and facial skin regions. Finally, the generated label maps are merged hierarchically as a complete annotation.

Benefiting from the proposed framework, we construct a new facial **La**ndmark guided face **Pa**rsing (LaPa) dataset efficiently. It consists of more than 22,000 images, covering large variations in facial expression, pose and occlusion. Each image is provided with annotations of an 11-category (namely hair, face skin, left/right eyebrow, left/right eye, nose, upper/lower lips, inner mouth and background) pixel-level label map along with the coordinates of 106-point landmarks. Fig. 1 gives some annotation examples of the proposed LaPa dataset.

Beyond the lack of training data, another critical issue of face parsing is that the semantic labels of the boundary pixels are challenging to be predicted. It is caused by the semantic confusion, especially for the facial parts where the boundary pixels cover a nonnegligible proportion. To tackle this problem, we propose an effective Boundary-Attention Semantic Segmentation (BASS) method, which improves the performance by fully utilizing the boundary information from two aspects: 1) the boundary-aware features are integrated into semantic features in the network to preserve more contour details. 2) an additional boundary-attention semantic loss is developed to reinforce the boundary effect in model optimization. Extensive experiments demonstrate the effectiveness of our method.

We summarize the contributions of this paper as follows:

1) We develop a high-efficiency framework for face parsing annotation, which considerably simplifies and speeds up the face parsing annotation, and makes it possible to construct a large-scale face parsing dataset efficiently.

2) By using the proposed framework, we construct a new large dataset for face parsing. It contains more than 22,000 images. Each image is provided with an 11-category pixel-level semantic label map and coordinates of 106-point landmarks. It could be applied to numerous face related applications.

3) We propose an effective boundary-attention semantic segmentation method for face parsing, which provides a baseline for the proposed LaPa dataset and achieves the state-of-the-art performance on the public Helen dataset.

Related Work

Datasets

Due to the high-cost of pixel-level annotation, there are few face parsing datasets published. The most commonly used datasets are LFW-PL (Kae et al. 2013) and Helen (Le et al. 2012; Smith et al. 2013). LFW-PL is a subset of the Labeled Faces in the Wild (LFW) funneled images which is a dataset of face photographs dedicated to the unconstrained face recognition. This dataset contains 2,927 facial images. All the images are first segmented into superpixels, and then each superpixel is manually assigned with one of the hair/skin/background categories. The annotations for facial parts are not provided in this dataset. The original Helen dataset (Le et al. 2012) is composed of 2,330 facial images with densely-sampled, manually-annotated keypoints around the semantic facial parts. Smith *et al.* (Smith et al. 2013) generated segmentation ground truths of eye, eyebrow, nose, inside mouth, upper lip and lower lip automatically by using the contours, together with facial skin and hair categories generated from manually annotated boundaries and an automatic matting algorithm (Levin, Rav-Acha, and Lischinski 2008). Another related dataset is CelebAMask-HQ (Lee et al. 2019). It is a large-scale face image dataset that has 30,000 high-resolution face images selected from the CelebA dataset (Liu et al. 2015b) by following CelebA-HQ (Karras et al. 2018). The masks of CelebAMask-HQ were manually-annotated with 19 classes.

Methods

In recent years, increasing attention has been drawn in face parsing due to its great application potentials. Early works mainly focus on hand-crafted features and probabilistic graphical models. Warrell *et al.* (Warrell and Prince 2009) proposed to use priors to model facial structure and got facial parts labels through a Conditional Random Field (CRF). Smith *et al.* (Smith et al. 2013) adopted SIFT features to select examplers and computed segmentation map of a test image by propagating labels from the aligned exemplar images. Kae *et al.* (Kae et al. 2013) combined CRF with a Restricted Boltzmann Machine (RBM) to model both local and global structures for face labeling. More recently, certain works attempt to tackle the face parsing task with the help of deep learning to break the performance bottleneck of traditional methods. Luo *et al.* (Luo, Wang, and Tang 2012) proposed a hierarchical face parsing framework with Deep Belief Networks (DBNs) as facial parts and components detectors. Liu *et al.* (Liu et al. 2015a) exploited a Convolution Neural Network (CNN) to model both unary likelihoods and pairwise label dependencies. Yamashita *et al.* (Yamashita et al. 2015) proposed a weighted cost function to improve performance for certain classes like eyes. Jackson *et al.*(Jackson, Valstar, and Tzimiropoulos 2016) proposed a two-stage parsing framework with Fully Convolutional Networks (FCNs). Liu *et al.* (Liu et al. 2017) designed a light-weight network which combines a shallow CNN with a spatially variant Recurrent Neural Network

(RNN) and a coarse-to-fine approach for accurate face parsing. Wei *et al*. (Wei et al. 2017) introduced an automatic method for selecting receptive fields. Guo *et al*. (Guo et al. 2018) adopted a prior mechanism to refine the Residual Encoder Decoder Network (RED-Net). Wei (Wei et al. 2019) revisited the structure of traditional FCN and proposed an accurate face parsing method at real-time speed. Lin (Lin et al. 2019) proposed a novel RoI Tanh-warping operator and achieved the state-of-the-art performance on both LFW-LP and Helen datasets.

The New Face Parsing Dataset

In this section, we first describe the proposed high-efficiency framework for face parsing annotating, and then introduce the constructed face parsing dataset LaPa.

High-Efficiency Framework

The framework is composed of two consecutive modules, named Dense Landmark Annotation (DLA) and Pixel-level Parsing Annotation (PPA).

Dense Landmark Annotation module The purpose of the DLA module is to annotate facial images with dense landmarks efficiently. We develop a semi-automatic facial landmark labeling tool with user interface. This tool can give a reference position for each landmark with the help of an auxiliary facial landmark localization model, so that annotators only need to adjust a small number of points for difficult cases rather than annotating all from scratch. In this paper, hourglass network (Newell, Yang, and Deng 2016; Bulat and Tzimiropoulos 2017) is employed, which is trained with mere 2,000 manually annotated images at the beginning and updated once 2,000 additional images are labeled. The Normalized Mean Error (NME) and Area Under Curve (AUC) on the test set (2,000 accurately labeling images by annotators) *w.r.t.* the number of training samples are reported in Fig. 2. We can see that the performance of the auxiliary model keeps improving along with the increasing samples accumulated by our semi-automatic labeling tool. This tool significantly simplifies and speeds up the process of dense landmark annotation. In this work, the DLA module takes the 106-point landmark definition. The outputs of this module will be fed to the PPA module.

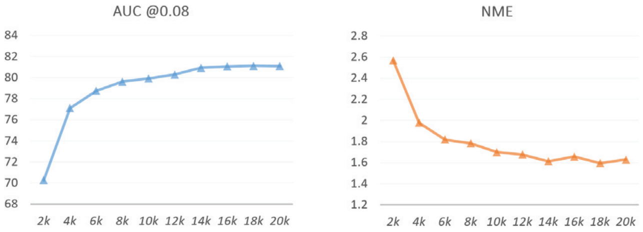


Figure 2: Performance of the auxiliary model for landmark localization *w.r.t.* the number of training samples. The horizontal axis refers to the number of training samples which is accumulated by our facial landmark labeling tool. The vertical axis refers to the corresponding evaluation performance.

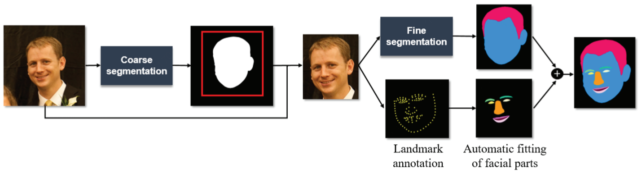


Figure 3: The framework for the proposed PPA module. First, the coarse segmentation is applied on the whole image and then the face region is cropped and finely segmented into three categories including hair, skin and background. Meanwhile, the annotation of facial parts are produced by our category-wise fitting approach according to the manually annotated facial landmarks. Finally, the outputs are merged hierarchically as a complete annotation.

Pixel-Level Parsing Annotation module This module takes a color image and the coordinates of 106-point landmarks as input, and outputs a pixel-level semantic label map including 11 categories. As Fig. 3 shows, it consists of three stages:

1) Coarse-to-fine segmentation for hair and face skin. Parsing hair and skin are important for many facial applications, such as hair coloring, skin whitening, *etc*. However, conventional facial landmarks are not defined on the forehead and hair regions, in part because forehead regions are often occluded by hair. Thus, we solve this problem resorting to the human parsing dataset CIHP[2]. It is the first standard and comprehensive benchmark for instance-level human parsing. Because the test set of CIHP does not supply the ground truth, we just adopt the training and validation set, totally 33,280 images.

In the coarse segmentation stage, we firstly map twenty categories in CIHP into two by taking hair and skin as foreground and others as background. Then we crop the regions of interest from original images according to the mapped labels incorporating with the instance labels. Usually, one image in CIHP could produce several sub-images in which only one major face exists. After filtering the images of which the width or height is less than 80 pixels, we collect about 26,000 images for training. Here we adopt the advanced Pyramid Scene Parsing Network (PSPNet) (Zhao et al. 2017) to segment the foreground (hair and skin) from the background. This stage could be considered as a face detection operation to preserve the hair region while regular face detectors usually focus on the face region and may lose part of hair.

In the fine segmentation stage, we process the data in a similar way as the coarse segmentation stage but retaining hair and skin regions as two separate categories. In order to obtain more accurate segmentation results, the proposed BASS method is adopted in this stage, which will be introduced in the latter section.

2) Category-wise fitting for facial parts. Facial parts include left/right eyebrow, left/right eye, nose, upper/lower lip and inner mouth. In order to obtain more natural and accurate contours, we develop different fitting schemes for dif-

[2] http://sysu-hcp.net/lip/overview.php

ferent facial parts according to their characteristics. For eyebrow, outer contour of mouth and jawline, we adopt polygon fitting to generate approximated contours. The pixels within each polygon are assigned to the corresponding category. In some cases, direct connection of long-distance neighboring landmarks may cause piecewise linear effect. To overcome this problem, prior knowledge is leveraged to make the results smoother by interpolation. For eye and inner mouth, two parabolas are applied to sketch the upper and lower boundary separately. For nose, we separate it into left and right parts to handle the profile case, and piecewise fitting is adopted due to the complex shape of nose. Note that all the partial landmarks are fitted in the transformed space where each part is aligned with a standard pose.

3) Hierarchical fusion. After the above two steps, we could obtain the label maps for hair/skin and facial parts. Then we merge them into a unified label map hierarchically. We emphasize that the order of fusion is important for producing correct results. For example, eye is always beyond the skin but sometimes behind hair or sunglasses, so if the label of eye covers the label of hair, the results will be unreasonable. Therefore, we merge the label maps in the order of skin, facial parts, hair and background.

Dense Landmark Guided Face Parsing (LaPa) dataset

We collect 22,176 images from two popular datasets - the landmark localization dataset 300W-LP (Sagonas et al. 2013; Zhu et al. 2016) and the face recognition dataset Megaface (Kemelmacher-Shlizerman et al. 2016). We randomly select 2,000 images for validation and 2,000 images for test, and the remaining 18,176 images are taken as the training set. All the images are annotated by the DLA module first, and then the color image and the landmarks are fed to the PPA module to obtain an 11-category semantic label for each pixel. The label is determined according to the visible texture. Therefore, some categories may be not presented due to occlusions. For example, eye may be invisible due to large pose or occlusion by other objects such as sunglasses or hair. In this work, we focus on single face parsing, and thus only the major face is annotated even if multiple faces exist in an image. Fig. 1 shows some examples of the proposed dataset.

Comparison with relevant datasets Helen (Smith et al. 2013) is a widely used dataset for face parsing, while it still has several limitations: 1) The labeling is not accurate enough especially for hair and face skin categories produced by matting. As a result, most works based on Helen only focus on facial components, while ignoring hair and skin. 2) The limited number of samples make it difficult to support training large-scale practical models. The LFW-PL dataset (Kae et al. 2013) has the same issue of the lack of training images while only hair and facial skin are annotated without considering facial parts. In contrast, the CelebAMask-HQ dataset (Lee et al. 2019) has plentiful images with more categories. Nevertheless, most of the samples in CelebAMask-HQ are frontal or nearly frontal facial images, lacking the pose diversity. Compared with these existing datasets, the proposed LaPa dataset has obvious advantages on the diversity. Furthermore, resorting to the proposed framework, the LaPa dataset could be scaled up easily with incremental facial images.

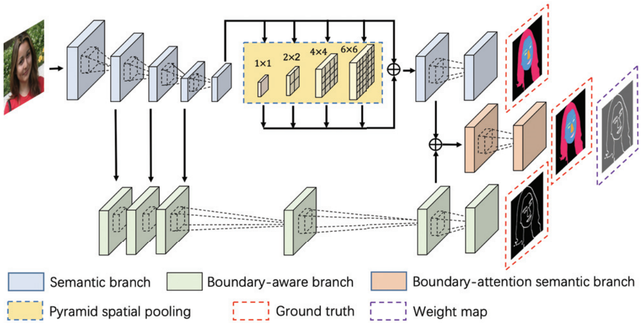


Figure 4: The network structure of BASS. It consists of three branches. The semantic branch runs a multi-category semantic segmentation task. The boundary-aware branch runs a two-category boundary segmentation task. The boundary-attention semantic branch takes the combination of the features from the former two branches as input and employs the boundary map to weight the semantic loss for boundary pixels.

Boundary-Attention Semantic Segmentation

Although face could be approximately considered as a rigid body in which the deformation is limited, face parsing is still a challenging task due to the variations in facial expression and pose. Meanwhile, the facial parts are usually smaller than general objects while the boundary pixels, which are difficult to be distinguished due to the semantic confusion, covers a nonnegligible proportion. To overcome these problems, we propose an effective Boundary-Attention Semantic Segmentation (BASS) method for face parsing, which fully utilizes the boundary information in both the network improvement and loss development.

As Fig. 4 shows, the network structure consists of three branches. The basic branch of the network is called semantic branch. The purpose is to learn semantic features and infer accurate semantic label maps from facial images. Any existing segmentation network structure could be taken in this branch. Here, we employ ResNet-101 (He et al. 2016) as the backbone. In order to reduce the resolution loss caused by pooling or convolution with stride larger than 1, dilation convolution (Chen et al. 2018) is adopted in the fifth residual block, therefore the resolution of the output is 1/16 rather than 1/32 of the input. In order to leverage the global texture information, pyramid spatial pooling with different scales are used before the classifier learning. Then, the output feature maps with different resolutions are concatenated with high-resolution feature maps generated by the last residual block after interpolation to the same size. The integrated feature maps are used to predict the semantic label for each pixel.

As we have mentioned above, boundary pixels are important but challenging to be predicted. In order to improve the segmentation performance for boundary pixels, a boundary-aware branch is added to learn boundary features by a boundary detection task. First, it extracts shared features from different layers of ResNet-101 in the semantic branch, and then projects them into a new space where boundary details are well preserved. The output of this branch is a boundary map in which each value refers to the confidence score that pixel is located on the boundary without considering semantics. Many existing works (Wu et al. 2018; Ruan et al. 2019) and our experiments have proved the effectiveness of exploiting boundary information.

However, the effect is limited due to the independence of these two tasks. Therefore, similar to (Ruan et al. 2019), we employ another branch called boundary-attention semantic branch. The features extracted from the semantic branch and boundary-aware branch are combined in this branch, which are rich in semantics while boundary details are well preserved. We attempt different combination strategies, including element plus, element multiply and channel concatenation. The experimental results show that all the three strategies are useful for the segmentation while concatenation is the best way (Tab. 1).

In addition to the enhancement in feature space, we develop a boundary-attention semantic loss which enlarges the semantic loss of boundary pixels according to the boundary map. This loss could significantly improve the segmentation accuracy, especially for the categories with clear boundaries. The comparison results will be given in the next section. The total loss function are defined as follows:

$$L = \lambda_1 L_s + \lambda_2 L_b + \lambda_3 L_{bs}, \tag{1}$$

$$L_s = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}^{s} \log p_{ij}^{s}, \tag{2}$$

$$L_b = -\frac{1}{N}\sum_{i=1}^{N}(y_i^b \log p_i^b + (1-y_i^b)\log(1-p_i^b)), \tag{3}$$

$$L_{bs} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} w_i \, y_{ij}^{s} \log p_{ij}^{bs}, \tag{4}$$

where L refers to the total loss. L_s, L_b and L_{bs} denote the loss of the semantic, boundary-aware, and boundary-attention semantic branches, respectively. λ_1, λ_2 and λ_3 are hyper-parameters to balance the loss of different branches. N denotes the number of pixels in the whole image while C denotes the number of parsing categories. Here y_{ij}^s equals to 1 if the semantic label of pixel i is j, and $y_{ij}^s = 0$ otherwise. y_i^b is a indicator variable of which the value is 1 if pixel i is located on the boundaries and 0 otherwise. p^s, p^b and p^{bs} are prediction values of the three branches, respectively. To emphasize the effect of boundaries, we introduce a new parameter $w_i = 1 + \alpha$, if $y_i^b = 1$ and $w_i = 1$, otherwise. Note that if $\alpha = 0$, L_{bs} is equivalent to the conventional cross entropy loss. If $\alpha > 0$, the loss of boundary pixels will be enlarged. During test phase, p^{bs} is taken as the prediction of the output.

Experiments

In this section, we first conduct extensive experiments on the LaPa dataset to validate the effectiveness of the proposed method. Then we compare our method with other state-of-the-arts on the public Helen dataset.

Experimental Settings

For both LaPa and Helen datasets, we adopt similar network configurations. For the semantic branch, the network parameters of ResNet-101 are initialized from the model pretrained on the ImageNet dataset (Deng et al. 2009). The input size of the network is 473×473, and dilation convolution is used in the last residual block to retain the feature resolution. The extracted features are then processed by a spatial pyramid pooling module with four different scales of 1×1, 2×2, 3×3 and 6×6 to aggregate global and local contextual information. For the boundary-aware branch, the feature maps from conv2_3, conv3_4 and conv4_23 in ResNet-101 are concatenated as input. As the same in (Ruan et al. 2019), we adopt a positive/negative sample balancing strategy which takes the ratio of pixels belonging to specific class as the weights of the opposite one. For the boundary-attention semantic branch, the last feature maps before predictors of the semantic and boundary-aware branches are combined as the input features.

The network is trained by minimizing the objective function defined in Eq. (1). We use mini-batch gradient descent as the optimizer with the momentum of 0.9, weight decay of 0.0005 and batch size of 64. ”Poly” learning rate policy is used to update parameters and the initial learning rate is set to 0.001. Synchronized Batch Normalization is adopted to accelerate the training procedure. For simplicity, the $\lambda 1$, $\lambda 2$ and $\lambda 3$ are all set to 1. Parameter α are determined on the validation set. Our experiments are implemented by Pytorch framework, and all the models are trained on 4 NVIDIA Tesla P40 with 24GB memory.

Similar to the previous work (Lin et al. 2019), face alignment is implemented as a pre-processing step on the Helen dataset while the results are evaluated on the original annotations without image transformation and cropping. However, face alignment is not employed on the LaPa dataset because it may cause the losses the foreground pixels, *e.g.*, hair. F1-score is taken as the quantitative evaluation metric, which is commonly used by the existing face parsing literature.

Experiments on the LaPa Dataset

Ablation Study To evaluate the effectiveness of the proposed strategies separately, we train the following models on the LaPa dataset. The basic model A is trained only with the semantic branch, which does not utilize any auxiliary boundary information. The model B is trained with both the semantic branch and the boundary-aware branch. The two tasks share low-level features but are independent on the high-level layers. The model C^x is trained with all the three branches (*i.e.* semantic branch, boundary-aware branch and boundary-attention semantic branch) while $\alpha = 0$ in the boundary-attention semantic loss. The Model D^x is trained with the same network structure as Model C^x while α is set to 1. The superscript x denotes different combination strategies for the two kinds of features, where p means element plus, m means element multiplication, and c means concatenation on the channels.

Tab. 1 gives the comparison results. We can see that the basic model A achieves an overall F1-score of 88.62%. While adding the boundary-aware branch, the accuracy is improved to 89.08% by model B. That means the boundary detection task could preserve the auxiliary boundary information in the low-level features, which is useful for the high-level semantic predictions. Models C^p, C^m and C^c are all higher than model B, achieving 89.25%, 89.26% and 89.32%, respectively. Models D^p, D^m and D^c further improve the performance to 89.50%, 89.53% and 89.62% by enlarging the semantic loss of boundary pixels. For both the training of model C and model D, the concatenation strategy achieves the best performance compared to the element-plus and element-multiplication strategies. Fig. 6 provides some visualization results for comparison of the three branches.

Evaluation with different parameter α We inspect the performance *w.r.t.* parameter α for each category on the validation set. As Fig. 5 shows, the performance with $\alpha = 2$ is significantly better than that with $\alpha = 0$ for almost all the facial parts except the nose. We speculate this is because the nose has no clear boundary. For the categories with clear boundaries such as eyebrows, eyes and mouth, the boundary-attention semantic loss (with $\alpha > 0$) brings notable improvement. For the hair, skin and background categories, there is no obvious change with different values of α because the boundary pixels cover a very small proportion for these categories. In our experiments, the mean F1-score with $\alpha = 10$ achieves 90.07%, higher than $\alpha = 0$ by 0.75%.

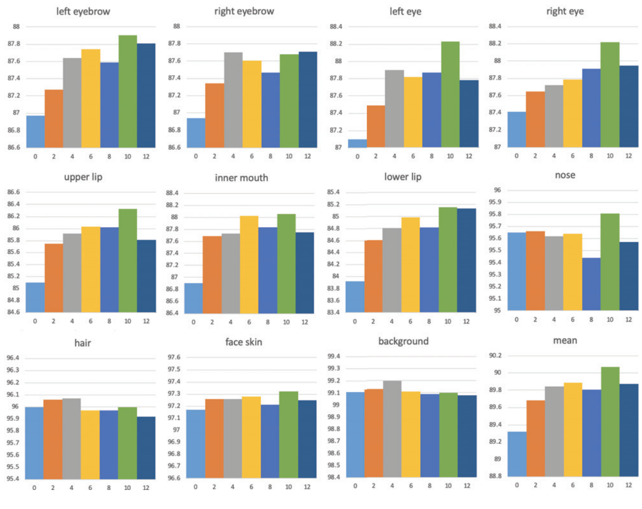
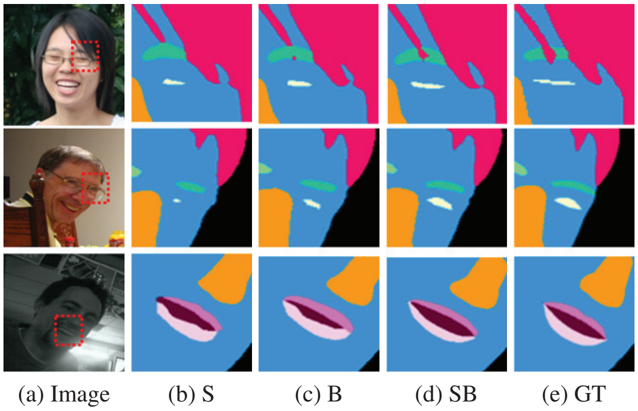


Figure 5: F1-score *w.r.t.* the parameter α. The horizontal axis refers to the value of α. The vertical axis refers to the F1-score on the validation set.

Baseline for the LaPa dataset Because the official codes of the related methods on face parsing are not available, we just report the results of our BASS method on our LaPa dataset, which could be taken as the baseline for further comparison on this dataset. The last two rows in Tab. 1 gives the results. It achieves 90.07% F1-score on the validation set and 89.79% F1-score on the test set. The parameter α is set to 10 which is determined on the validation set.

Figure 6: Visualizing the results on the LaPa dataset. (a) is the original color image, in which the red dashed rectangles indicate the challenging parts. (b)-(d) are the zoom-in results of models trained with different network structure. S, B and SB denote the semantic branch, boundary-aware boundary and boundary-attention semantic branch, respectively. (e) is the ground truth of the corresponding image.



(a) Image (b) S (c) B (d) SB (e) GT

Comparison with State-of-the-art Methods

Dataset and ground truth Since the original Helen dataset (Le et al. 2012) is made for facial landmark localization instead of parsing, Smith *et al.* (Smith et al. 2013) take several steps to convert densely labeled landmarks into segmentation maps. Specifically, ground truth segments of facial parts are automatically generated from manually-annotated contours. For facial skin, the jawline contour is used as the lower boundary, while for the upper boundary, an automatic matting algorithm (Levin, Rav-Acha, and Lischinski 2008) is used to separate the forehead from hair. The same matting strategy is adopted to recover the hair region. To make the dataset adaptive to semantic segmentation tasks, we first transform the confidence maps ranging from 0 to 255 to label maps by selecting the category with the maximum confidence value. However, this will cause incorrect ground truth for certain cases, especially for the hair category. As the number of training images are limited, many methods adopt exemplar based approach, which needs to split some images as exemplar during the training and test phase. As in (Liu et al. 2015a; Guo et al. 2018), 230 images are split as exemplars. In contrast, our method could directly output the prediction result for each pixel by the network, thus we use all the training images and exemplar images for training model, and test images are the same as (Liu et al. 2015a; Guo et al. 2018).

Experimental results As the previous works (Smith et al. 2013; Liu et al. 2015a; 2017; Guo et al. 2018; Wei et al.

Table 1: Ablation study and baselines on the LaPa dataset. **Model A** is trained only by the semantic branch. **Model B** is trained by Model A plus the boundary-aware branch. **Model C**^{x} is trained by Model B plus the boundary-attention semantic branch with softmax loss($\alpha = 0$ in Eq. (1)). **Model D**^{x} is trained by Model B plus the boundary-attention semantic branch with boundary-attention semantic loss ($\alpha = 1$). The superscript x denotes different combination strategies for the two kinds of features, where p means element plus, m means element multiplication, and c means concatenation on the channels. The performance of each category, together with the mean F1-score over the 10 foreground categories on the validation set are listed. The last two rows are the results of our BASS method on the validation set and test set.

	hair	skin	left	right	left	right	nose	upper	inner	lower	background	mean
			eyebrow	eyebrow	eye	eye		lip	mouth	lip		
A	95.53	96.88	86.09	85.59	86.26	86.25	95.33	83.35	86.57	84.37	99.01	88.62
B	95.92	97.08	86.31	86.20	86.70	86.88	95.56	83.99	87.12	85.04	99.10	89.08
C^p	95.98	97.16	86.78	86.56	87.05	87.08	95.64	84.20	86.84	85.16	99.11	89.25
C^m	95.95	97.15	86.82	86.75	86.94	87.24	95.48	83.97	87.18	85.16	99.11	89.26
C^c	96.00	97.17	86.97	86.94	87.10	87.41	95.65	83.92	86.90	85.10	99.11	89.32
D^p	96.04	97.17	87.02	87.01	87.24	**87.62**	95.62	84.49	87.29	85.52	99.12	89.50
D^m	96.03	97.21	87.18	86.99	87.39	87.58	95.61	84.34	**87.43**	85.56	99.12	89.53
D^c	**96.09**	**97.24**	**87.31**	**87.16**	**87.58**	87.56	**95.66**	**84.57**	87.39	**85.63**	**99.13**	**89.62**
Val.	96.00	97.32	87.90	87.68	88.23	88.22	95.81	85.15	88.06	86.33	99.10	90.07
Test	96.31	97.24	87.67	87.55	88.06	87.91	95.47	84.36	87.64	85.70	99.17	89.79

Table 2: Comparison with state-of-the-art methods on the Helen dataset. To keep consistent with other methods, the performance of the hair category and other fine-grained categories (*e.g.* left/right eyes) is omitted. The overall scores are computed by combining the merged brows/eyes/mouth and nose categories.

	skin	nose	upper-lip	inner-mouth	lower-lip	brows	eyes	mouth	overall
Smith *et al*. (Smith et al. 2013)	88.2	92.2	65.1	71.3	70.0	72.2	78.5	85.7	80.4
Liu *et al*. (Liu et al. 2015a)	91.2	91.2	60.1	82.4	68.4	73.4	76.8	84.9	85.9
Liu *et al*. (Liu et al. 2017)	92.1	93.0	74.3	79.2	81.7	77.0	86.8	89.1	88.6
Guo *et al*. (Guo et al. 2018)	93.8	94.1	75.8	83.7	83.1	80.4	87.1	92.4	90.5
Wei *et al*. (Wei et al. 2019)	**95.6**	95.2	80.0	86.7	86.4	82.6	89.0	93.6	91.6
Lin *et al*. (Lin et al. 2019)	94.5	95.6	79.6	86.7	89.8	83.1	89.6	95.0	92.4
BASS (ours)	94.9	**95.8**	**83.7**	**89.1**	**91.4**	**83.5**	**89.8**	**96.1**	**93.1**

2019) do not report the performance of hair and fine-grained categories (*i.e.* left/right eyebrow and left/right eye) on the Helen dataset, the mean score of foreground categories cannot be computed as Tab. 1 shows. To keep consistent with the previous methods, we report our results on the skin, nose, upper-lip, inner-mouth, lower-lip, merged brows, merged eyes and merged mouth categories. The overall scores are computed by combining the merged brows, merged eyes, merged mouth and nose categories without considering fine-grained categories. As Tab. 2 shows, our model achieves the overall score of 93.1%, surpassing state-of-the-art methods.

Conclusion

In this paper, we develop a high-efficiency framework for face parsing annotation, which significantly simplifies the pixel-level semantic annotation for face parsing with high accuracy. Benefiting from this novel framework, we construct a new dataset for face parsing. It consists of more than 22,000 facial images and each image is provided with an 11-category semantic label map along with coordinates of 106-point landmarks. Furthermore, we propose an effective boundary-attention semantic segmentation method for face parsing, which boosts the performance by fully utilizing the boundary information in both network improvement and loss development. Experiments on Helen and the proposed LaPa datasets demonstrate the effectiveness of our method.

References

Baltrusaitis, T.; Robinson, P.; and Morency, L.-P. 2013. Constrained local neural fields for robust facial landmark detection in the wild. In *ICCV Workshops*, 354–361.

Bulat, A., and Tzimiropoulos, G. 2017. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 1021–1030.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(4):834–848.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 4690–4699.

Fu, C.; Wu, X.; Hu, Y.; Huang, H.; and He, R. 2019. Dual variational generation for low-shot heterogeneous face recognition. In *NeurIPS*.

Guo, T.; Kim, Y.; Zhang, H.; Qian, D.; Yoo, B.; Xu, J.; Zou, D.; Han, J.-J.; and Choi, C. 2018. Residual encoder decoder network and adaptive prior for face parsing. In *AAAI*, 6861–6869.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Jackson, A. S.; Valstar, M.; and Tzimiropoulos, G. 2016. A cnn cascade for landmark guided semantic part segmentation. In *ECCV*, 143–155.

Kae, A.; Sohn, K.; Lee, H.; and Learned-Miller, E. 2013. Augmenting crfs with boltzmann machine shape priors for image labeling. In *CVPR*, 2019–2026.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*.

Kemelmacher-Shlizerman, I.; Seitz, S. M.; Miller, D.; and Brossard, E. 2016. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, 4873–4882.

Koestinger, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2011. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV workshops*, 2144–2151.

Le, V.; Brandt, J.; Lin, Z.; Bourdev, L.; and Huang, T. S. 2012. Interactive facial feature localization. In *ECCV*, 679–692.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2019. Maskgan: Towards diverse and interactive facial image manipulation. *arXiv preprint arXiv:1907.11922*.

Levin, A.; Rav-Acha, A.; and Lischinski, D. 2008. Spectral matting. *TPAMI* 30(10):1699–1712.

Lin, J.; Yang, H.; Chen, D.; Zeng, M.; Wen, F.; and Yuan, L. 2019. Face parsing with roi tanh-warping. In *CVPR*, 5654–5663.

Liu, S.; Yang, J.; Huang, C.; and Yang, M.-H. 2015a. Multi-objective convolutional learning for face labeling. In *CVPR*, 3451–3459.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015b. Deep learning face attributes in the wild. In *ICCV*, 3730–3738.

Liu, S.; Shi, J.; Liang, J.; and Yang, M.-H. 2017. Face parsing via recurrent propagation. In *BMVC*, 1–10.

Luo, P.; Wang, X.; and Tang, X. 2012. Hierarchical face parsing via deep learning. In *CVPR*, 2480–2487.

Merget, D.; Rock, M.; and Rigoll, G. 2018. Robust facial landmark detection via a fully-convolutional local-global context network. In *CVPR*, 781–790.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *ECCV*, 483–499.

Ou, X.; Liu, S.; Cao, X.; and Ling, H. 2016. Beauty emakeup: A deep makeup transfer system. In *ACM Multimedia*, 701–702.

Ruan, T.; Liu, T.; Huang, Z.; Wei, Y.; Wei, S.; and Zhao, Y. 2019. Devil in the details: Towards accurate single and multiple human parsing. In *AAAI*, volume 33, 4814–4821.

Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; and Pantic, M. 2013. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV Workshops*, 397–403.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823.

Smith, B. M.; Zhang, L.; Brandt, J.; Lin, Z.; and Yang, J. 2013. Exemplar-based face parsing. In *CVPR*, 3484–3491.

Sun, Y.; Chen, Y.; Wang, X.; and Tang, X. 2014. Deep learning face representation by joint identification-verification. In *NeurIPS*, 1988–1996.

Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 1701–1708.

Warrell, J., and Prince, S. J. 2009. Labelfaces: Parsing facial features by multiclass labeling with an epitome prior. In *ICIP*, 2481–2484.

Wei, Z.; Sun, Y.; Wang, J.; Lai, H.; and Liu, S. 2017. Learning adaptive receptive fields for deep image parsing network. In *CVPR*, 2434–2442.

Wei, Z.; Liu, S.; Sun, Y.; and Ling, H. 2019. Accurate facial image parsing at real-time speed. *TIP* 28:4659–4670.

Wu, Y.; Hassner, T.; Kim, K.; Medioni, G.; and Natarajan, P. 2017. Facial landmark detection with tweaked convolutional neural networks. *TPAMI* 40(12):3067–3074.

Wu, W.; Qian, C.; Yang, S.; Wang, Q.; Cai, Y.; and Zhou, Q. 2018. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2129–2138.

Yamashita, T.; Nakamura, T.; Fukui, H.; Yamauchi, Y.; and Fujiyoshi, H. 2015. Cost-alleviative learning for deep convolutional neural network-based facial part labeling. *IPS TCVA* 7:99–103.

Zafeiriou, S.; Zhang, C.; and Zhang, Z. 2015. A survey on face detection in the wild: past, present and future. *CVIU* 138:1–24.

Zhang, S.; Zhu, X.; Lei, Z.; Shi, H.; Wang, X.; and Li, S. Z. 2017. S3fd: Single shot scale-invariant face detector. In *ICCV*, 192–201.

Zhang, H.; Riggan, B. S.; Hu, S.; Short, N. J.; and Patel, V. M. 2018. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *IJCV* 1–18.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *CVPR*, 2881–2890.

Zhou, E.; Fan, H.; Cao, Z.; Jiang, Y.; and Yin, Q. 2013. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *ICCV Workshops*, 386–391.

Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; and Li, S. Z. 2016. Face alignment across large poses: A 3d solution. In *CVPR*, 146–155.