The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Residual Encoder Decoder Network and Adaptive Prior for Face Parsing

Tianchu Guo,[1] **Youngsung Kim,**[2] **Hui Zhang,**[1] **Deheng Qian,**[1] **ByungIn Yoo**[2]
Jingtao Xu,[1] **Dongqing Zou,**[1] **Jae-Joon Han,**[2] **Changkyu Choi**[2]
[1]Beijing Samsung Telecommunication, SAIT China Lab
[2]Samsung Advanced Institute of Technology, S/W Solution Lab

Abstract

Face parsing assigns every pixel in a facial image with a semantic label, which could be applied in various applications including face expression recognition, facial beautification, affective computing and animation. While lots of progress have been made in this field, current state-of-the-art methods still fail to extract real effective feature and restore accurate score map, especially for those facial parts which have *large variations of deformation* and *fairly similar appearance*, e.g. mouth, eyes and thin eyebrows. In this paper, we propose a novel pixel-wise face parsing method called Residual Encoder Decoder Network (RED-Net), which combines a feature-rich encoder-decoder framework with adaptive prior mechanism. Our encoder-decoder framework extracts feature with ResNet and decodes the feature by elaborately fusing the residual architectures into deconvolution. This framework learns more effective feature comparing to that learnt by decoding with interpolation or classic deconvolution operations. To overcome the appearance ambiguity between facial parts, an adaptive prior mechanism is proposed in term of the decoder prediction confidence, allowing refining the final result. The experimental results on two public databases demonstrate that our method outperforms the state-of-the-arts significantly, achieving improvements of F-measure from 0.854 to 0.905 on the Helen dataset, and pixel accuracy from 95.12% to 97.59% on the LFW dataset. In particular, convincing qualitative examples show that our method parses eye, eyebrow and lip regions more accurately.

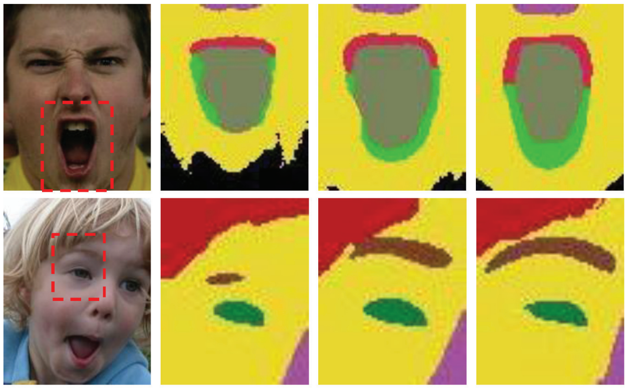

Figure 1: Challenging cases for face parsing. Existing methods fail to parse the facial parts with large variations of deformation and fairly similar appearance, e.g. mouth and light eyebrows. The first column is the original image, the red dashed frame indicates the challenging parts. The next three columns are zoom-in versions of parsing results of the red dashed frame. The second column shows the results of existing method, i.e. VGG-Deconv Net(Noh, Hong, and Han 2015) reproduced by us. The third column shows the results of our RED-Net. The last column shows the ground truth.

Introduction

Face parsing is both fundamental and important to a variety of computer vision and animation areas ranging from classic tasks such as facial beautification, affective computing and face animation, to modern applications like face Augmented Reality (AR) and expression transfer (Ou et al. 2016; Zhang et al. 2017). The problem of the face parsing is to assign every pixel with a semantic label, thus the face parsing is much more valuable and challenging comparing with face landmark detection.

Though tremendous strides have been made in face parsing, current state-of-the-art algorithms of face parsing (Kae et al. 2013; Khan, Mauro, and Leonardi 2015; Liu et al. 2015; Luo, Wang, and Tang 2012; Smith et al. 2013; Tsogkas et al. 2015; Warrell and Prince 2009; Yamashita et al. 2015; Zhou, Hu, and Zhang 2015)(or semantic segmentation) often have difficulty with extremely similar appearance and parts having large variation of deformation. On the one hand, the similar appearance of parts results in that the distinguish ability of features that learned or hand-crafted is limited. So the absolute boundary of facial parts with extremely similar appearance cannot be recognized clearly. For example, as shown in Fig. 1, segmenting child's eyebrows from its skin is tough due to the light color of eyebrows. On the other hand, large deformation of parts will cause that the features for different states of the same part, e.g., mouth open or close, are very different. Consequently, the features of such parts vary so large that the robustness of existing methods is low. It is very difficult to achieve accurate parsing results while maintaining robustness for various cases.

Recent face parsing methods parse faces usually through using deep convolutional neural network (CNN)(Kae et al. 2013; Liu et al. 2015; Luo, Wang, and Tang 2012; Tsogkas et al. 2015; Yamashita et al. 2015; Zhou, Hu, and Zhang 2015) rather than handcrafted feature (Smith et al. 2013; Warrell and Prince 2009). However, these methods still cannot process the problems we point out above for following reasons. First of all, their deeper plain network cannot provide more precise features for fine-grained segmentation due to the degradation problem (He et al. 2016a). Secondly, the traditional interpolation without any post-processing for score map reconstruction will result in over-smoothness (Chen et al. 2016; 2014). Last but not least, although facial prior correlation information is utilized implicitly in previous approaches (Liu et al. 2015), they do not apply the prior adaptively according to the face feature information to avoid the conflict between face feature and the prior.

In this paper, we present a novel Residual Encoder Decoder network (RED-Net) to pursuit higher accurate face parsing. First, an encoder with shortcut connections (He et al. 2016a) (i.e. residual branch) is utilized to encode face semantic information. With ensemble information from feature maps in different scales, more powerful rich feature could be generated. Second, we develop a novel decoder with combination of the shortcut branch and the deconvolution layer. This type of bottleneck structure decreases the total network parameters and increases its nonlinearity ability. Compared to previous interpolation methods (Chen et al. 2014; 2016), the deconvolution net could learn pixel-wise score map with more detailed semantic information automatically. Third, in order to further segment similar facial regions, an adaptive prior mechanism is designed to refine the decoder output automatically. The explicit prior for each facial part which can be considered as a kind of structure constraint is added to the score map and its weight is adjusted according to the prediction confidence of pixel-wise score map.

The proposed method is evaluated on two public datasets, Helen and LFW. Experimental results demonstrate that the proposed network outperforms state-of-the-art approaches significantly. The major contributions of this paper can be summarized as follows.

- A powerful network, RED-Net, is proposed by introducing shortcut branches to an encoder-decoder framework. With abundant information from different scales, efficient and discriminative features are generated for face parsing. An accurate pixel-wise score map could be learnt in an end-to-end fashion.
- An adaptive prior mechanism is designed to incorporate face structure constraint for further final score map refinement.
- Our method achieved state-of-the-art performance on two benchmark datasets. Especially, our method shows much higher accuracy for these facial parts which have large variation of deformation and fairly similar appearance, e.g. mouth, eye and thin eyebrows. These facial parts are extremely difficult to separate from others but crucial for popular applications including facial beautification and affective computing.

Related work

In the past few years, face parsing draws increasing attention for its wide range of applications. These methods can be classified into two categories. The first one is handcrafted feature based method. Warrell et al. (Warrell and Prince 2009) model multinomial priors of facial structure and label facial parts with a Conditional Random Field (CRF). Kae et al. (Kae et al. 2013) model the face shape prior with a restricted Boltzmann machine and combine it with a CRF for labeling on superpixels with handcrafted features. Smith et al. (Smith et al. 2013) use landmarks and SIFT features to transfer labeling masks from a set of aligned exemplars and achieve better results on rare facial parts such as eyes, nose, brows, and mouth. Khan et al. (Khan, Mauro, and Leonardi 2015) adopt color, shape and location features in a random forest model to label facial components. All these methods require manually designed specific features and cannot bring satisfying results.

More recently, deep learning frameworks are employed for face parsing to constitute another class of methods. Liu et al. (Liu et al. 2015) train a multi-objective CNN with face image patches and upsample CNN outputs to obtain pixel-wise segmentation map. Zhou et al. (Zhou, Hu, and Zhang 2015) present an interlinked CNN to solve the parsing problem in an end-to-end fashion. Yamashita et al. (Yamashita et al. 2015) utilize a weighted cost function to improve the performance of CNN during facial part labeling. Luo et al. (Luo, Wang, and Tang 2012) detect face parts and parse each facial component by learning a highly nonlinear mapping with deep belief network respectively. Although these deep learning based methods lead to promising results, they still suffer from the following limitations: 1) A plain CNN is adopted in previous works, however, it misses abundant information between feature maps in different scales and suffers from performance degradation problem (He et al. 2016a; 2016b; Huang et al. 2016; Ioffe and Szegedy 2015; Larsson, Maire, and Shakhnarovich 2016; Singh, Hoiem, and Forsyth 2016; Targ, Almeida, and Lyman 2016; Zhang et al. 2016). 2) Current methods require an additional decoder to restore pixel-wise score map through interpolation (Chen et al. 2014; 2016; Zheng et al. 2015). However, inspired by other computer vision tasks, a holistic end-to-end solution is prone to be trained as a high performance engine (Badrinarayanan, Handa, and Cipolla 2015; Noh, Hong, and Han 2015). 3) Face prior information is utilized implicitly, which increases model complexity. Although (Liu et al. 2015) integrates the face structure constraint as part of CNN input, it is still less effective for segmenting detailed facial components. Consequently, we propose a novel pixel-wise face parsing method to address these thorny issues and achieve state-of-the-art performance. In the rest of this paper, we will demonstrate how these issues are largely resolved in details.

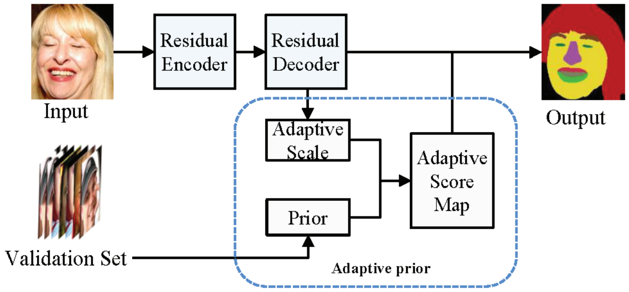

Figure 2: System overview. The proposed system consists of three parts, i.e. Residual Encoder Network, Residual Decoder Network and Adaptive prior.

Method of Face Parsing

The framework of our method is illustrated in Fig. 2. Concretely, the proposed system consists of three parts. The first part is the Residual Encoder Network (RE-Net), consisting of a series of Residual Encoder Unit (RE-Unit), which is a concise residual network to extract features of facial parts. The second part is the Residual Decoder Network (RD-Net), consisting of a series of Residual Decoder Unit (RD-Unit), which restores the feature map to original resolution to obtain a pixel-wise score map. The third part is the adaptive prior which refines the score map by adaptively adjusting the contribution of the prior.

Specifically, face parsing is the problem of assigning labels to each pixel. It can be solved by maximizing the following function,

$$F(Y|X;w) = \prod_{y_i \in Y} \mathscr{N}(P_{net}(y_i|X;w) + H(y_i|X;w)Pr(y_i|X;w)), \tag{1}$$

where X is the input image and Y is a set of random variables $y_i \in Y$ defined on every pixel i. Each y_i takes a value from a set of labels $L = \{1, 2, ..., K\}$ indicating different facial parts. w is the parameter of our parsing algorithm. The first part $P_{net}(y_i)$ is the output of RED-Net, which is a probability distribution of y_i. The second part accounts for the prior restriction which is the product of entropy $H(y_i)$ and prior distribution $Pr(y_i|X;w)$. $\mathscr{N}(\cdot)$ is a normalization function.

In the training process, the parameter w is optimized by minimizing the function of $L(X, Y, w)$,

$$L(X,Y;w) = \sum_{y_i \in Y} -log(\mathscr{N}(P_{net}(y_i = l_i|x_i;w) + H(y_i|X;w)Pr(y_i = l_i|X;w))), \tag{2}$$

where l_i is the ground truth label for pixel i.

Residual Encoder Network (RE-Net)

In the part one, we design a residual encoder network to extract informative features. The network is shown in Fig. 3. It consists of six units which are separated from each other by pooling layers. Its basic block is the residual structure (He et al. 2016a). In order to use more general facial features, we pretrain our encoder network on a face recognition task.

In the pretraining process, the parameters of the network are optimized by minimizing the following cost function,

$$Loss(Y,X) = -log(P(y = id|X, w_{pre})), \tag{3}$$

where X denotes the image. id is the identity of X. w_{pre} is the parameter of the pretrained network. $P(y = id|X; w_{pre})$ is the confidence score that y belongs to label id:

$$P(y = id|X, w_{pre}) = \frac{\exp\{y'_{id}\}}{\sum_{i=1}^{Q} \exp\{y'_i\}}, \tag{4}$$

where $y'_i = \sum_{j=1}^{d} (h_j w_{ji} + b_i)$, which is a linear combination the d-dimensional features $\{h_j\}$ as the input of neuron i. Q is the number of identities.

After the pretraining process, only the first five units are preserved in the encoder for face parsing task. The sixth unit is removed because the high level features encoded by top layers are more task-oriented and contain identity information, which are not suitable for face parsing.

Residual Decoder Network (RD-Net)

In the part two, we design the residual decoder network to translate the low-resolution feature map generated by the encoder. In order to obtain a pixel-wise labeling result, we enlarge the resolution and recover the image details gradually with a stack of RD-Units, as shown in Fig. 3.

The building blocks of RD-Net are unpooling, deconvolution and batch normalization layers. The unpooling layer can enlarge the resolution. The deconvolution layer is responsible for reconstructing details . And the batch normalization greatly accelerates the speed of convergence (Ioffe and Szegedy 2015).

The designing idea of the RD-Unit is to introduce shortcuts into a deconvolutional network. The general consideration is to replace the deconvolution layer with the bottleneck structure for reducing network parameter and increasing its nonlinearity, and then add shortcuts between them. It consists of three different modules.

Densification module. The function of this module is to obtain a denser feature map by applying a deconvolution bottleneck on the unpooling result of the previous unit's output. In more detail, the output of an unpooling layer is a sparse feature map which preserves efficient features and their positions. And the densification module tries to fill the "hole" in the neighborhood of the activated neurons. As shown in Fig. 4, the sparse ratio drops greatly after densification module in all the RD-Units. Although the result of this module is denser than its input, it is still a coarse feature map that needs further refinement.

Detail-learning module. This module is a residual architecture with two branches, i.e., the deconvolution bottleneck branch and the shortcut branch, which is an identity mapping. The outputs of these two branches are added elementwisely. Similar to the motivation of ResNet(He et al. 2016a), we regard the output H(x) of this module as a fine-grained feature map, where x is the input coarse feature map. Thus the output of the deconvolution bottleneck branch, F(x) =

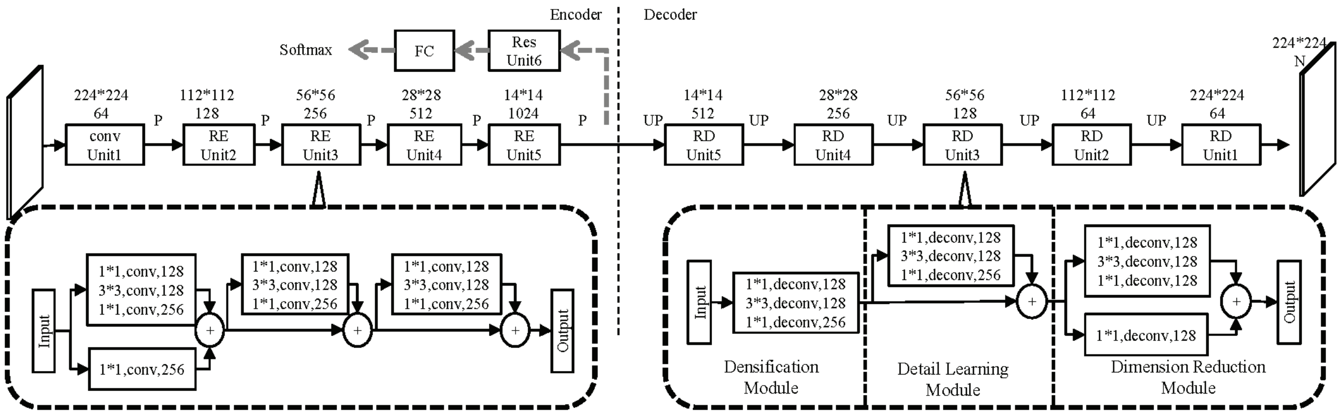

Figure 3: The configuration of the proposed RED-Net. The pretrained network is shown on the left side of the dashed vertical line. “P” in the encoder means max-pooling, and “UP” in the decoder means max-unpooling. Each conv/deconv layer is followed by a batch normalization layer and a ReLU layer. The resolutions, dimensions and kernel sizes are also described.

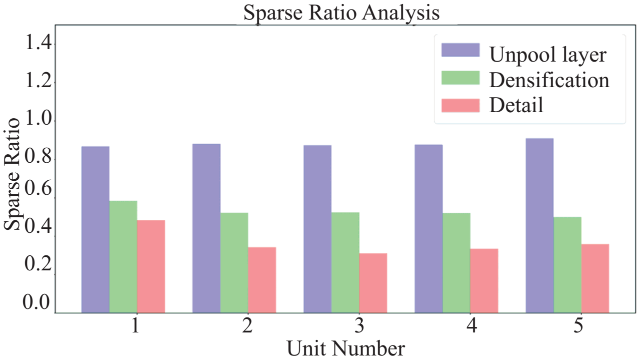

Figure 4: Sparse analysis. It computes the sparse ratio of feature map of all the RD-Units after unpooling layer, densification module and detail-learning module, respectively.

H(x)-x, represents the learned details. As shown in Fig. 4, comparing with the densification module, the sparse ratio drops slightly. It means that the feature map is dense but coarse, the detail-learning module further catch the details.

Dimension-reduction module. This module is also a residual architecture. It includes a projection branch (i.e. the 1x1 deconvolution layer) and a deconvolution bottleneck branch. Similar to the detail-learning module, their outputs are added element-wisely. The purpose of this module is to reduce the dimensions of the feature map, such that the current RD-Unit is able to be concatenated to the next one.

To design a RD-Net, we notice that there are two intrinsic requirements, i.e. upgrading sparse feature map to dense one, and achieving a clear boundaries between different facial parts. Thus, we divide the procedure of refining a coarse feature map into several steps using different modules. Three different modules could function cooperatively. As a result, more details of shapes and boundaries could be learned. Besides, benefiting from the shortcut branch and the projection branch in each RD-Unit, the RED-Net can be easily trained.

Adaptive prior

There is a strong motivation to exploit the prior of the facial structure since human faces generally share the same visual pattern. Hence, we incorporate prior information to further refine the score map explicitly. The proposed method is called adaptive prior, as shown in Fig. 5. Adaptive prior proceeds in four steps. In the first step, the prior information is obtained by following the procedure in (Liu et al. 2015). In brief, a face image is represented by five key facial points (two eye centers, nose tip, and two mouth corners) projected to a subspace, which is generated by applying principle component analysis to all images in the validation dataset. And the distance between two images is defined as Euclidean distance between their projections in this subspace. Then when a test image X is given, we can compute its distance to each validation image, select the closest N images $\{I_i\}_{i=1}^{N}$, and compute its prior information $Pr(Y|X)$ as the average of

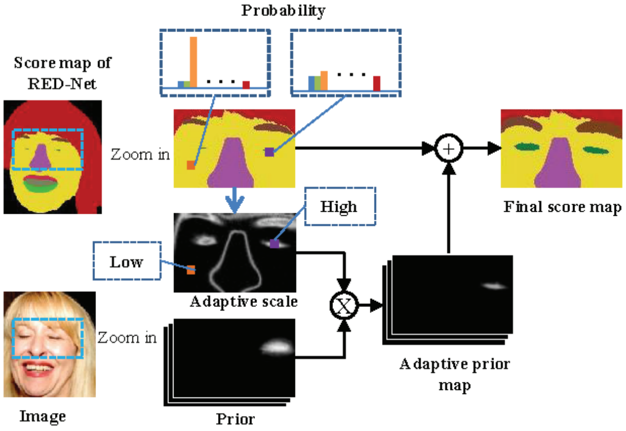

Figure 5: The adaptive prior mechanism. The exemplar prior image shown here is for the right eye region.

the ground truth labels of these closest images, i.e.,

$$Pr(Y|X) = \frac{1}{N}\sum_{i=1}^{N} gt(I_i), \tag{5}$$

where $gt(\cdot)$ is the ground truth label mask.

The second step is to adaptively compute the contribution weight of the prior (i.e. the "Adaptive scale" in Fig. 5). The weight is a matrix computed according to the entropy of the output of a decoder. Specifically,

$$As(y_i|X) = N(H(y_i|X;w)), \tag{6}$$

where $As(y_i)$ means the adaptive scale of the pixel x_i. $H(\cdot)$ means the entropy. $N(\cdot)$ scales its input to the range of $[0, 1]$. Note that w is omitted for abbreviation.

The third step is to compute the adaptive prior map. It is obtained by element-wisely multiplying the prior and the adaptive scale,

$$Apr(y_i|X) = As(y_i|X)Pr(y_i|X), \tag{7}$$

where $Apr(y_i|X)$ is the adaptive prior at pixel x_i. $Pr(y_i|X)$ means the prior information of pixel x_i.

The fourth step is to combine the adaptive prior map with the output of the decoder by the following formula,

$$P_{final}(y_i|X) = \mathscr{N}(P_{net}(y_i|X;w) + Apr(y_i|X)), \tag{8}$$

where P_{final} is the final score map. $P_{net}(\cdot)$ is the score map of the decoder.

The output of a decoder may be inaccurate, especially in regions of eyes and brows where the boundaries of facial parts are not clear. For a pixel near such boundaries, the confidence of assigning the pixel to each label is low, which causes large entropy, as illustrated in Fig. 5. For such pixels, the adaptive prior provides a compensation to refine the parsing result. For instance, the orange point is a pixel whose probability distribution has high confidence in skin category, which implies low entropy. While the purple point is an eye pixel whose probability distribution has low confidence in all categories, which implies high entropy. Such pixels with high entropy are very likely to be mis-labelled. Fortunately, after the multiplication of "Adaptive scale" and "Prior", the contribution of the prior information for them is strengthened in "Adaptive prior map", and their labels could be corrected in the "Final score map".

Experiment

In this section, we first detail the settings of our method, and then introduce the datasets on which the method is evaluated. After that, we compare our system with three algorithms. Two of them are state-of-the-arts, i.e. (Liu et al. 2015) and (Smith et al. 2013); while the third one is VGG-Deconv, an encoder-decoder network proposed in (Noh, Hong, and Han 2015). Note that we reproduce this network, and employ it to handle the face parsing problem. Some variant versions of our RED-Net are also evaluated to show the contribution from different components of the complete network, and the performance of variant decoder structures. In all experiments, the algorithms are evaluated according to the accuracy and F-measure.

Dataset and Network configuration

We evaluate our method on two public datasets, i.e. Helen(Smith et al. 2013) and LFW(Kae et al. 2013). The Helen dataset contains face labels with 11 classes annotated as hair, eyebrows, eyes, nose, lips, in mouth, skin and background. The dataset is divided into a training set with 2330 images, a testing set with 100 images and a validation set with 330 images.

In the LFW dataset there are 2927 face images collected in unconstrained environments. The resolution of each image is 250 × 250. All of them are annotated as skin, hair and background labels using superpixels. The whole dataset is divided into a training set with 1500 images, a testing set with 927 images and a validation set with 500 images. Both LFW and Helen are divided following the standard protocol (Liu et al. 2015).

For the Helen dataset, we crop the facial region and scale it to the resolution of 250×250. Such operation is done four times to obtain four images with different face sizes. Given the scaled images, we randomly crop them into patches of 224×224 and mirror the patches for data augmentation. In the LFW dataset, all the training, validation, and testing images are aligned. Thus, we use the original images for training and testing.

We first perform the pretrain, and then use the first units to initialize our encoder. After that, we add the RD-Net on the top of the encoder, as shown in Fig. 3. The training data for pretrain is CASIA-Webface (Yi et al. 2014), and the feature dimension d (see Eq.(4)) is set to 1024. The weights of the decoder are initialized randomly according to a Gaussian distribution with zero mean and standard deviation 0.01. The whole network RED-Net is fine-tuned on the face parsing task.

The training procedure is carried out using mini-batch gradient descent. The momentum, weight decay rate and the batch size are set to 0.9, 5e-4 and 64, respectively. The learning rate is initially 0.01 and decreases to its 1/10 every 5000 iterations. Besides, we set the number of closest validation images N to 5.

Quantitative analysis

We first show the results on the Helen dataset. We compare with the methods of (Smith et al. 2013) and (Liu et al. 2015) following their evaluation protocol. As shown in Table 1, our RED-Net with the adaptive prior (i.e. Ours-p) achieves the best performance for most compared items. We achieve an overall F-measure of 0.905, which is 5.1 percent improvement than Liu's work. Especially, the results of upper lip, lower lip, brows and eyes are 15.7, 14.7, 7, and 10.3 percent higher than those of Liu's work (Liu et al. 2015), respectively.

Both deconvolution based methods, i.e. VGG-Deconv and Ours in Table 1, improve the performance of the published state-of-the-arts (Liu et al. 2015; Smith et al. 2013). This is probably because the deconvolution can reconstruct the shape and boundaries more accurately than interpolation (Liu et al. 2015). Note that our RED-Net achieves higher F-measure in all the categories than VGG-Deconv.

	Skin	Nose	Upper Lip	In Mouth	Lower Lip	Brows	Eyes	Mouth all	Overall
Smith(Smith et al. 2013)	0.882	0.922	0.651	0.713	0.700	0.722	0.785	0.857	0.804
Liu's(Liu et al. 2015)	0.912	0.912	0.601	0.824	0.684	0.734	0.768	0.849	0.854
VGG-Deconv	0.931	0.925	0.673	0.809	0.800	0.761	0.796	0.89	0.874
VGG-Deconv-p	0.931	0.932	0.719	0.809	0.792	0.778	0.849	0.900	0.886
Res-Deconv	0.935	0.929	0.692	0.805	0.771	0.781	0.843	0.903	0.887
Ours	0.938	**0.941**	0.731	0.832	**0.833**	0.793	0.866	0.923	0.903
Ours-p	0.938	**0.941**	**0.758**	**0.837**	0.831	**0.804**	**0.871**	**0.924**	**0.905**
Ours-p-implicit	**0.940**	0.937	0.703	0.813	0.808	0.738	0.833	0.917	0.887

Table 1: Results on the Helen dataset. The first 2 rows record the results reported in (Smith et al. 2013) and (Liu et al. 2015), respectively. VGG-Deconv (Noh, Hong, and Han 2015) is reproduced by us for the face parsing task. Other rows show result of networks designed by us, including: VGG-Deconv-p, which is VGG-Deconv plus our adaptive prior mechanism; Res-Deconv, which has the same encoder as proposed RED-Net and the same decoder as VGG-Deconv; Ours, which is our RED-Net (Note that the structure shown in Fig. 6-(a) is adopted); Ours-p, which is Ours plus the adaptive prior mechanism; Ours-p-implicit, which is our RED-Net plus prior mechanism in (Liu et al. 2015).

Analysis of network components. First we show the benefit of the residual encoder. In Table 1 Res-Deconv has the gain of 1.3 percent vs. VGG-Deconv (88.7% vs. 87.4%), verifying the effectiveness of the residual encoder. Then we compare different decoders. The results of Res-Deconv and Ours in Table 1 show that our decoder is more powerful than the plain deconvolution network. Note that the same encoder is employed in these networks for fair comparison between decoders. Actually not all encoder-decoder combinations can work well. We experiment with the combination of plain VGG encoder and our decoder, and the performance is only 65.1%. The reasonable explanation behind this is the plain encoder cannot extract efficient representation without adequate information from different scales and our powerful decoder cannot translate it into clear score map. Thus, the design of overall network configuration is valuable.

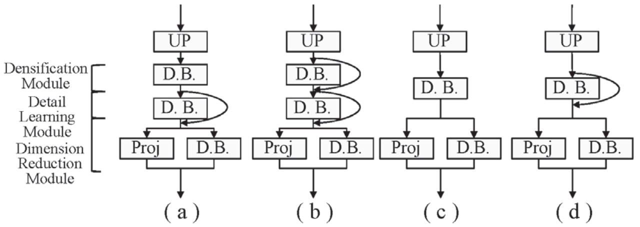

Figure 6: Schematic illustration of different structures in RD-Unit. "D.B." means the "Deconvolution Bottleneck", "UP" denotes the unpooling layer and "Proj" means the 1×1 deconvolution layers

Analysis of Adaptive Prior. The benefit of the adaptive prior can be observed from the comparison of VGG-Deconv vs. VGG-Deconv-p, and Ours vs. Ours-p, shown in Table 1. The overall F-measure of VGG-Deconv-p is 1.2 percent higher than that of VGG-Deconv. Especially, the results of upper lip, brows, and eyes are 4.6, 1.7, and 5.3 percent higher than the original results of the corresponding facial parts. The gain of Ours-p vs. Ours is also quite significant for such regions, increasing by 2.7, 1.1, and 0.5 percent, respectively. In fact, for such regions the learned feature is sometime similar to that of the skin, so the labeling confidence is low. For such cases the adaptive prior makes prior contributes more, and can correct some labeling errors. It can be observed that the gain for Ours is not so significant as that for VGG-Deconv, possibly because the feature map extracted by our network is much more discriminative. However, even for a much better network, our prior mechanism still improves the performance, and the gain for important organs is quite obvious.

Except for our explicit adaptive prior mechanism, (Liu et al. 2015) also utilizes prior as an additional network input via an implicit way. Here we use the same prior information to compare two different prior mechanisms and the results are given in Table 1. With the same network structure, our approach exceeds implicit method by a large margin. Especially for eyebrows and eyes region, implicit method's performance decreases from 80.4% to 73.8%, and 87.1% to 83.3%, respectively. And even compared to RED-Net without the prior (Ours in Table 1), it is much worse. This can be attributed to that the prior information is not accurate enough which could increase the noise of the whole network training and it may have conflict with learnt feature. However, our explicit prior mechanism largely solves this issue by incorporating prior information adaptively.

Analysis of RED-Unit. Different choices of shortcut employment formed various decoding structures are shown in Fig. 6. Among these variants, the structure of Fig. 6-(a) performs slightly better than others. RD-Nets consist of variant RD-Units are compared in Table 2. We remove all the detail-learning modules from var_a to get var_c. It is interesting that the training strategy affects the performance of var_c greatly. If the weights in decoder part are randomly initialized, the performance of the trained model is only 83.3%. In contrast, if the weights of the model are initialized from var_a, the performance is 90.2%. Thus we achieve a smaller model whose performance is comparable to that of var_a. We also remove all the densification modules from var_a to get var_d, and observe similar effect of different training strategies (var_d vs. var_d-ft). Note that var_c-ft has a gain of 1.1 percent vs. var_d-ft. So the densification module is preferred than the detail-learning module in a simplified configuration. In fact, the input here is a very sparse feature map, and the goal of the densification module is to activate the unpooled neurons properly. However, if a shortcut branch is added to convert this module to a detail-learning module, it must learn to depress the activation at unpooled neurons to make the result feature map smooth, which is hard because the activation and the depression must be fulfilled simultaneously. Var_a, which contains all three types of modules, performs slightly better than var_b, possibly benefiting from the diverse module functionalities.

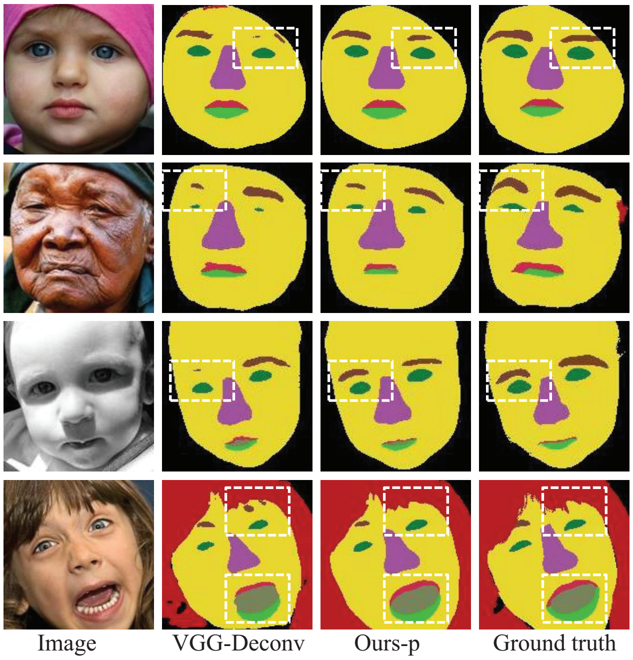

Figure 7: Comparison between VGG-Deconv (Noh, Hong, and Han 2015) and RED-Net.

	Overall	Model Size (MB)
Res-var_b	0.901	107
Res-var_c	0.833	90
Res-var_c-ft	0.902	90
Res-var_d	0.819	90
Res-var_d-ft	0.891	90
Res-var_a	**0.903**	107

Table 2: The comparison of variant decoder structures. Var_a, var_b, var_c, var_d are the network with the same encoder as shown in Fig. 3, and different RD-Unit as shown in Fig. 6-(a), (b), (c) and (d), respectively. The suffix "-ft" means that in the training process the model is initialized with the var_a result. The models are trained with Caffe platform (Jia et al. 2014).

We show results on the LFW dataset, which has three categories, i.e., skin, hair and background. In this task, we evaluate the performance according to the pixel accuracy and the F-measure of each category. Results are shown in Table 3. Comparing with Liu's work, significant improvements of the F-measure in all three categories and the pixel accuracy can be observed. Especially, our work improves by roughly 6 percent in the hair category, and the error rate is reduced by 50.6% in total pixel accuracy.

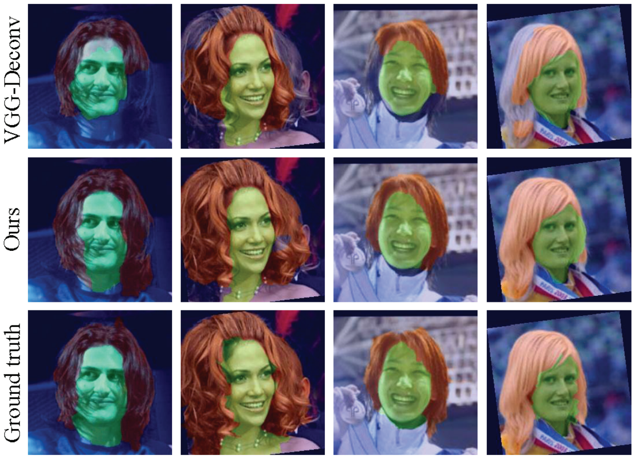

Figure 8: Comparison between VGG-Deconv (Noh, Hong, and Han 2015) and our RED-Net on LFW.

Qualitative analysis

To qualitatively show the advantages of our proposed method, we specifically compare the results of RED-Net with prior and VGG-Deconv in Fig. 7. It can be observed that our RED-Net with adaptive prior outperforms in eyebrow regions when the boundaries of eyebrows are not clear, as shown in Fig. 7 Row 1 and Row 3. Our RED-Net is more robust when there is occlusion between hair and eyebrows, and there is deformable variation in mouth region, as shown in Row 4. We get results in eye regions when the skin has complex textures and boundaries, as shown in Row 2. It can be demonstrated that facial parts with similar appearance such as light eyebrow and eyes, facial parts with large deformable variation can be segmented clearly by our method.

We also show the parsing results on LFW dataset in Fig. 8. We find that our result for hair regions is obviously better, which is consistent with the quantitative results of F-hair. Note that our results are visually better than the ground truth. The ground truth is labeled based on superpixel (Kae et al. 2013), and may be inaccurate on the boundaries.

To explain the functions of modules and advantages of our

	Acc	F-bg	F-hair	F-skin
Liu's	95.12	97.10	80.70	93.93
VGG-Deconv	97.06	97.43	78.05	94.45
Ours	**97.59**	**98.08**	**86.93**	**95.20**

Table 3: Results on the LFW dataset. The first row reports the results of (Liu et al. 2015). The second row shows the results of VGG-Deconv (Noh, Hong, and Han 2015). The third row shows the results of our RED-Net.

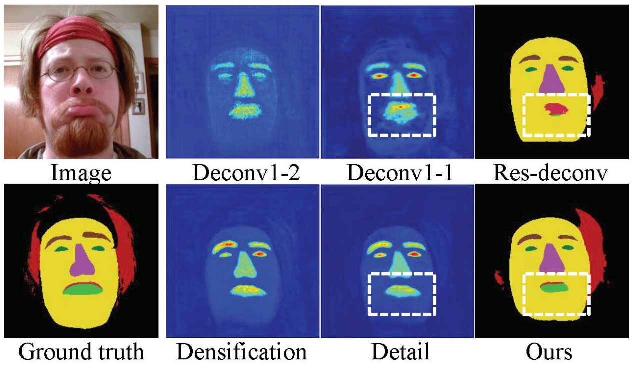

Figure 9: Outputs of last several modules (or layers) after last unpooling layer.

RD-Units, Fig. 9 shows the decoding results from last several modules (or layers) obtained from the Res-Deconv and our RED-Net,respectively. For fair comparison and clear illustration, we choose the feature maps after the last unpooling layer. It can be observed that the feature map of deconvolution layers in Res-Deconv shown in Fig. 9-(Deconv1-2) is still sparse, while our result of the densification module shown in Fig. 9-(densification) is much denser. Moreover, our result of detail-learning module shown in Fig. 9-(detail) has clearer boundaries than the result of the last deconvolution layer in Res-Deconv in Fig. 9-(Deconv1-1). And our prediction result in Fig. 9-(Ours) is more accurate than the result in Fig. 9-(Res-deconv).

To explain the contribution of adaptive prior, Fig. 10 shows the result of with and without adaptive prior. It can be observed that issue of facial parts with similar appearance can be resolved in both VGG-Deconv network or our RED-Net. It can be demonstrated that our adaptive prior mechanism has a strong generalization ability.

Analysis of convergence

Here we show that our network can converge faster than VGG-Deconv-Net(Noh, Hong, and Han 2015). All the experiments are conducted on Helen with 8000 iterations. Fig. 11 shows the pixel accuracy against the number of iterations on validation dataset for the two networks. It can be seen that RED-Net reaches 89.93% accuracy (1.29 percent less than final) with only 800 iterations, while VGG-Deconv-Net(Noh, Hong, and Han 2015) reaches 89.10% accuracy (0.95 percent less than final) with around 7400 iterations. So our network converges much faster. We also show the comparison of the training loss between RED-Net and VGG-Deconv-Net(Noh, Hong, and Han 2015) in Fig. 12. It can be seen that the training loss of our RED-Net decreases faster.

Conclusion

In this paper, we propose a deep residual encoder-decoder framework, i.e. RED-Net, and adaptive prior mechanism to tackle the face parsing problem, especially for those facial parts which have large variations of deformation and fairly similar appearance, e.g. eyes, eyebrows and mouth. With the powerful RED-Net, efficient and discriminative features are extracted. Adaptive prior mechanism incorporates face structure constraint and further refines the parsing results.

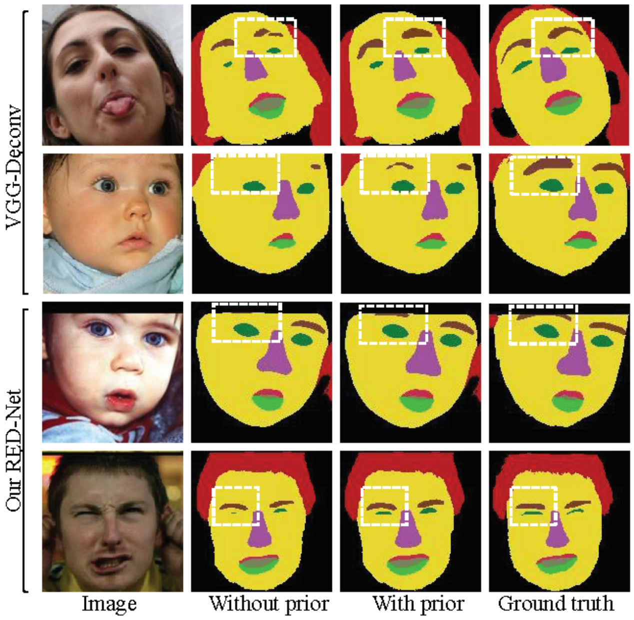

Figure 10: Prior analysis. Visualize the results of networks without and with adaptive prior mechanism.

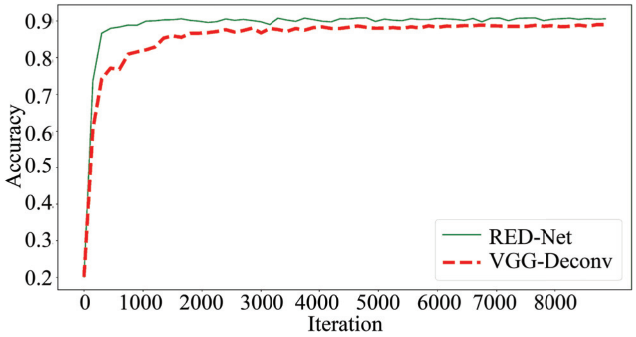

Figure 11: Validation accuracy curve.

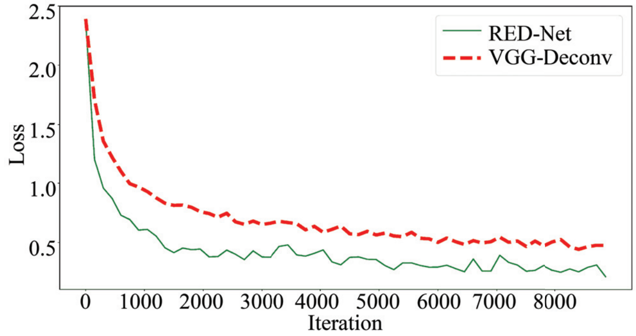

Figure 12: Training loss curve.

Experimental results demonstrate that our method outperforms the state-of-the-arts by a large margin, improving the F-measure by 5.1% on Helen dataset, and reducing the error rate by 50.6% on LFW dataset. Specifically, the results of our method in upper lip, lower lip, brows and eyes regions are 15.7, 14.7, 7 and 10.3 percent higher than existing method. Qualitative analysis also show that these regions can be segmented more clearly.

Acknowledgement

We would like to thank the anonymous reviewers for their help in improving the paper. We also want to thank Qiang Wang and He Zheng for helpful suggestions, and Weiming Li for advices on English.

References

Badrinarayanan, V.; Handa, A.; and Cipolla, R. 2015. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, 630–645. Springer.

Huang, G.; Sun, Y.; Liu, Z.; Sedra, D.; and Weinberger, K. Q. 2016. Deep networks with stochastic depth. In *European Conference on Computer Vision*, 646–661. Springer.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, 675–678. ACM.

Kae, A.; Sohn, K.; Lee, H.; and Learned-Miller, E. 2013. Augmenting crfs with boltzmann machine shape priors for image labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019–2026.

Khan, K.; Mauro, M.; and Leonardi, R. 2015. Multi-class semantic segmentation of faces. In *Image Processing (ICIP), 2015 IEEE International Conference on*, 827–831. IEEE.

Larsson, G.; Maire, M.; and Shakhnarovich, G. 2016. Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*.

Liu, S.; Yang, J.; Huang, C.; and Yang, M.-H. 2015. Multi-objective convolutional learning for face labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3451–3459.

Luo, P.; Wang, X.; and Tang, X. 2012. Hierarchical face parsing via deep learning. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2480–2487. IEEE.

Noh, H.; Hong, S.; and Han, B. 2015. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 1520–1528.

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

Singh, S.; Hoiem, D.; and Forsyth, D. 2016. Swapout: Learning an ensemble of deep architectures. In *Advances in Neural Information Processing Systems*, 28–36.

Smith, B. M.; Zhang, L.; Brandt, J.; Lin, Z.; and Yang, J. 2013. Exemplar-based face parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3484–3491.

Targ, S.; Almeida, D.; and Lyman, K. 2016. Resnet in resnet: generalizing residual architectures. *arXiv preprint arXiv:1603.08029*.

Tsogkas, S.; Kokkinos, I.; Papandreou, G.; and Vedaldi, A. 2015. Deep learning for semantic part segmentation with high-level guidance. *arXiv preprint arXiv:1505.02438*.

Warrell, J., and Prince, S. J. 2009. Labelfaces: Parsing facial features by multiclass labeling with an epitome prior. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, 2481–2484. IEEE.

Yamashita, T.; Nakamura, T.; Fukui, H.; Yamauchi, Y.; and Fujiyoshi, H. 2015. Cost-alleviative learning for deep convolutional neural network-based facial part labeling. *IPSJ Transactions on Computer Vision and Applications* 7(0):99–103.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*.

Zhang, K.; Sun, M.; Han, T. X.; Yuan, X.; Guo, L.; and Liu, T. 2016. Residual networks of residual networks: Multilevel residual networks. *arXiv preprint arXiv:1608.02908*.

Zhang, D.; Lin, L.; Chen, T.; Wu, X.; Tan, W.; and Izquierdo, E. 2017. Content-adaptive sketch portrait generation by decompositional representation learning. *IEEE Transactions on Image Processing* 26(1):328–339.

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

Zhou, Y.; Hu, X.; and Zhang, B. 2015. Interlinked convolutional neural networks for face parsing. In *International Symposium on Neural Networks*, 222–231. Springer.