Decoupled Multi-task Learning with Cyclical Self-Regulation for Face Parsing

Qingping Zheng[1], Jiankang Deng[2], Zheng Zhu[3], Ying Li[1], Stefanos Zafeiriou[2,4]
[1]Northwestern Polytechnical University, [2] Huawei, [3]Tsinghua University, [4]Imperial College London
zhengqingping2018@mail.nwpu.edu.cn, jiankangdeng@gmail.com, zhengzhu@ieee.org
lybyp@nwpu.edu.cn, s.zafeiriou@imperial.ac.uk

Abstract

This paper probes intrinsic factors behind typical failure cases (e.g. *spatial inconsistency and boundary confusion) produced by the existing state-of-the-art method in face parsing. To tackle these problems, we propose a novel Decoupled Multi-task Learning with Cyclical Self-Regulation (DML-CSR) for face parsing. Specifically, DML-CSR designs a multi-task model which comprises face parsing, binary edge, and category edge detection. These tasks only share low-level encoder weights without high-level interactions between each other, enabling to decouple auxiliary modules from the whole network at the inference stage. To address spatial inconsistency, we develop a dynamic dual graph convolutional network to capture global contextual information without using any extra pooling operation. To handle boundary confusion in both single and multiple face scenarios, we exploit binary and category edge detection to jointly obtain generic geometric structure and fine-grained semantic clues of human faces. Besides, to prevent noisy labels from degrading model generalization during training, cyclical self-regulation is proposed to self-ensemble several model instances to get a new model and the resulting model then is used to self-distill subsequent models, through alternating iterations. Experiments show that our method achieves the new state-of-the-art performance on the Helen, CelebAMask-HQ, and Lapa datasets. The source code is available at https://github.com/deepinsight/insightface/tree/master/parsing/dml_csr.*

1. Introduction

Face parsing, as a fine-grained semantic segmentation task, intends to assign a pixel-wise label for each facial component, *e.g.*, eyes, nose, and mouth. The detailed analysis of semantic facial parts is essential in many high-level applications, such as face swapping [28], face editing [15], and facial makeup [29]. Benefit from the learning capacity of deep Convolutional Neural Networks (CNNs) and the labor effort put in pixel-level annotations [15, 21, 35], methods based on Fully Convolutional Networks (FCNs) [7, 10, 18–20, 23, 36, 47, 48] have achieved a promising performance on the fully supervised face parsing. Nevertheless, the local characteristic of the convolutional kernel prevents FCNs from capturing global contextual information [25], which is crucial for semantically parsing facial components in an image.

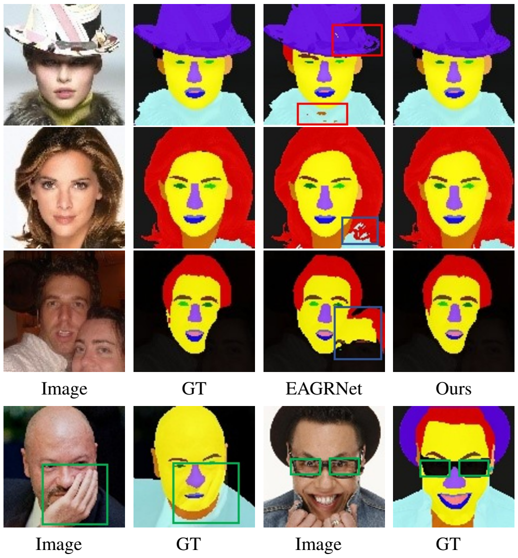


Figure 1. The first three rows show typical failure cases of spatial inconsistency and boundary confusion when applying EARGNet [36] to face parsing. The last row displays noisy labels on the training datasets.

To address this issue, most of the region-based face parsing methods [10, 20, 47] integrate CNN features into variant CRFs to learn global information. However, these methods do not consider the correlation among various objects. To this end, Te *et al*. [36] proposes the EAGRNet method to model a region-level graph representation over a face image by propagating information across all vertices on the graph. Even though EAGRNet enables reasoning over non-local regions to get global dependencies between distant facial

This work is done when Qingping Zheng is an intern at Huawei.

components and achieves state-of-the-art performance, it still faces the problems of spatial inconsistency and boundary confusion. In EAGRNet, PSP module [45] adopts an average pooling layer [22] to capture the global context prior, leading to an inconsistent spatial topology. Moreover, EAGRNet integrates additional clues of binary edges into context embedding to improve the parsing results. However, it is hard for EAGRNet to handle boundaries between highly irregular facial parts (*e.g.* hair and cloth in Figure 1) and distinguish clear boundaries between different face instances in the crowded scenarios (multi-faces in Figure 1).

Besides, learning a reliable model for face parsing requires accurate pixel-level annotations. Nonetheless, there inevitably exist careless manual labeling errors on the training dataset as shown in the last row of Figure 1. Te *et al.* [36] employ the traditional fully supervised learning scheme to train EAGRNet, failing to locate label noise because all pixels in the ground truth are processed equally. Notably, overlooking such incomplete annotations restricts the model generalization and prevents the performance from increasing to a higher level.

In this paper, we propose an end-to-end face parsing method, which is based on Decoupled Multi-task Learning with Cyclical Self-Regulation (DML-CSR). Specifically, given an input of facial image, the ResNet-101 [8] pre-trained on ImageNet is taken as the backbone to extract features from different levels. Afterwards, our multi-task model consists of three tasks, namely face parsing, binary edge detection, and category edge detection. These tasks share low-level weights from the backbone but do not have high-level interactions. Therefore, our multi-task learning approach can detach additional edge detection tasks from face parsing at the inference stage. To tackle spatial inconsistency raised by the pooling operation, we develop a Dynamic Dual Graph Convolutional Network (DDGCN) in the face parsing branch to capture long-range contextual information. The proposed DDGCN contains no extra pooling operation and it can dynamically fuse the global context extracted from GCNs in both spatial and feature spaces. To solve the boundary confusion in both single-face and multi-face scenarios, the proposed category-aware edge detection module exploits more semantic information than the binary edge detection module used in EARGNet [36].

To address the problem caused by noisy labels in training datasets, we introduce a cyclically learning scheduler inspired by self-training [3, 16, 34, 41, 42, 42, 49] to achieve advanced cyclical self-regulation. The proposed CSR contains a self-ensemble strategy that can aggregate a set of historical models to obtain a new reliable model and another self-distillation method that exploits the soft labels generated by the aggregated model to guide the successive model learning. Finally, the proposed CSR iteration alternates between these two procedures, correcting the noisy labels during training and promoting the model generalization. The proposed CSR can significantly promote the reliability of the model and labels in a cyclical training scheduler without introducing extra computation costs.

To summarize, our main contributions are as follows:

- We propose a decoupled multi-task network including face parsing, binary edge detection, and category edge detection. The face parsing branch introduces a DDGCN without any extra pooling operation to solve the problem of spatial inconsistency, and an additional category edge detection branch is designed to handle the boundary confusion.
- We introduce a cyclical self-regulation mechanism during training. The iteration alternates between one self-ensemble procedure, boosting model generalization progressively, and another self-distillation processing, regulating noisy labels.
- Our method establishes new state-of-the-art performance on the Helen [35] (93.8% overall F1 score), LaPa [21] (92.4% mean F1) and CelebAMask-HQ [15] (86.1% mean F1) datasets. Compared to EARGNet [36], our method utilizes fewer computation resources as the edge prediction modules can be decoupled from the whole network, decreasing the inference time from 89ms to 31ms but achieving much better performance.

2. Related Work

Face parsing. Most existing face parsing methods can be classified into global-based and local-based methods. Global-based methods aim to predict a pixel-wise label directly from the whole RGB face image. Early works learn spatial correlation between facial parts using various handcrafted models, such as epitome model [11] and exemplar-based method [35]. Later, many works [10, 20, 40, 47] implant the CNN-based features into the Conditional Random Field (CRF) framework, and adopt a multi-objective learning method to model pixel-wise labels and neighborhood dependencies simultaneously. Lin *et al.* [17] design a CNN-based framework with a RoI Tanh-Warping operator to use both central and peripheral information. Te *et al.* [36] introduce an edge-aware graph module to effectively reason relationship between facial regions. These global-based approaches inherently integrate the prior into the face layout, but limit accuracy due to overlook on each individual part.

Local-based methods aim to predict each facial part individually by training separated models for different facial regions. Luo *et al.* [24] exploit a hierarchical approach to segment each detected facial component separately. Zhou *et al.* [48] propose an interlinked CNN-based model to forecast pixel categories after face detection, taking a large expense of memory and computation consumption. Later, Liu *et al.* [19] combines a shallow CNN and a spatially variant RNN in two successive stages to parse a face image at

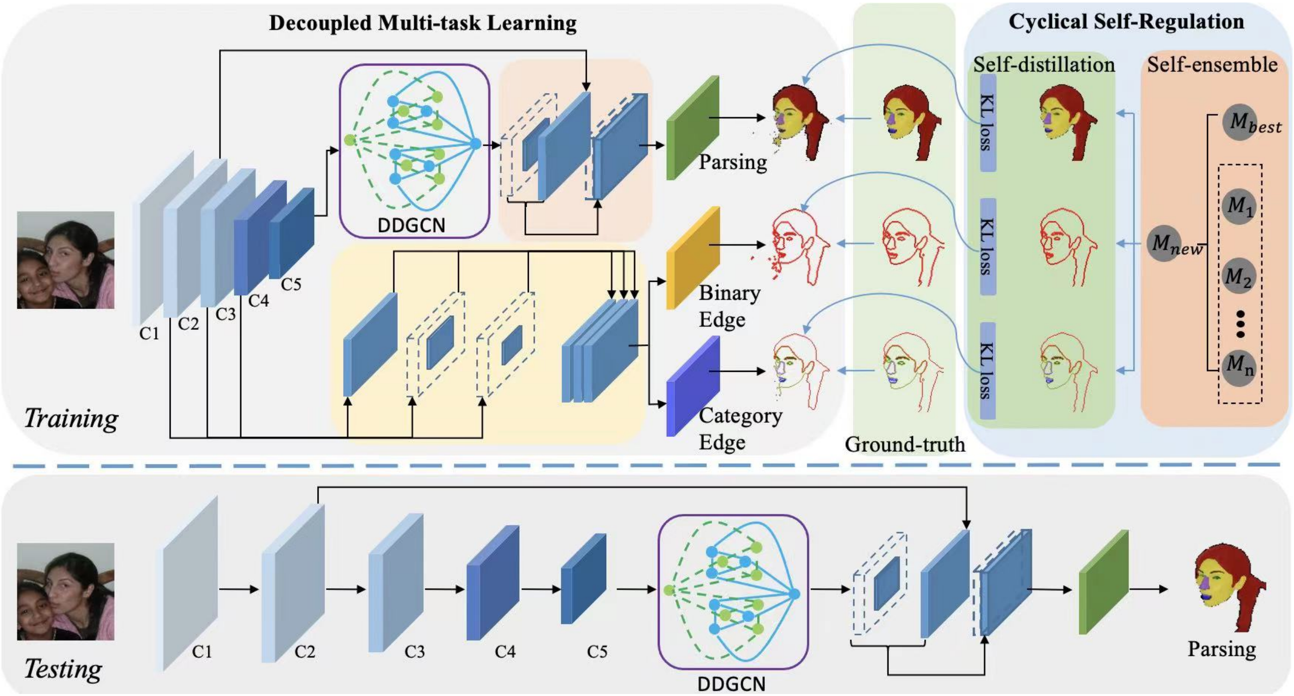


Figure 2. Overview of our proposed DML-CSR method for face parsing. At the training stage, it includes three parallel sub-models of face paring, binary edge detection and category edge detection, jointly trained by a proposed cyclical self-regulation mechanism. At the testing stage, all edge models are decoupled from the whole model.

a very fast inference speed. These local-based approaches almost take the coarse-to-fine policy with consideration of both global consistency and local precision. However, it ignores the improvement of accuracy and efficiency from backbone sharing and joint optimization.

Multi-task learning is a common strategy which jointly trains various tasks through the shared feature mechanism or hidden layers of a "backbone" model [2]. It has been widely applied for solving multiple pixel-level tasks. In the context of deep learning, multi-task learning can be categorized into hard or soft parameter sharing schemes. In hard parameter sharing based multi-task learning for image segmentation, the parameter set consists of shared and task-specific parameters. UberNet [14] is the first hard parameter sharing model for image segmentation, where a large number of low-, mid-, and high-level image vision tasks are tackled concurrently. Later, most multi-task learning models [12, 27, 37] follow the hard parameter sharing schemes and simply share the same encoder layers. In these works, each task-specific decoding head tails at the end of the shared encoder, leading to sub-optimal task groupings.

In soft parameter sharing based multi-task learning for image segmentation, each task has its own group of parameters, and a feature sharing mechanism is used to handle the cross-task communication. Cross-stitch network [26] is a typical multi-task architecture adopting the soft-parameter sharing schemes. It linearly combines the activations from every task-specific layer, regarding as soft feature fusion strategy. Afterwards, Ruder *et al*. [33] extends this method to learn the selective sharing layers. Compared to the hard parameter sharing approaches, the problem of multi-task learning based on soft parameter sharing approaches is a lack of scalability, as the growth of tasks make the size of the multi-task network increase linearly [38].

3. Methodology

This section starts with the analysis of representative failure cases when applying EARGNet [36] to face parsing. These issues motivate the proposal of a more accurate and robust training method, called Decoupled Multi-task Learning with Cyclical Self-Regulation (DML-CSR). The overall pipeline is illustrated in Figure 2.

3.1. Limitations of EAGRNet

Even though EAGRNet [36] achieves notable performance on face parsing, it has the following issues during training on public benchmark datasets (*e.g*. Helen [35], CelebAMask-HQ [15] and LaPa [21]).

Spatial Inconsistency. As shown in the first-row of Figure 1, EAGRNet improperly predicts "neck" pixels within the cloth area, resulting in spatial inconsistency of cloth. As EAGRNet employs an adaptive average pooling within PSP module [45] to capture global contextual information, the detailed spatial relationship and constraint between original pixels may be neglected. Therefore, a small part of area within a large region can be predicted as wrong classes.

Since directly adopting the general object segmentation method to face parsing is sub-optimal, we explore to avoid the unnecessary pooling operation in our model design.

Boundary Confusion. As intuitively illustrated in the second-row of Figure 1, EARGNet fails to distinguish boundaries between (1) "cloth" and "hair", and (2) the target face and the surrounding face under crowded scenario. Generally, component boundaries between different facial organs and instance boundaries between close faces can be confusing for face parsing models. As the edge network built in EARGNet simply integrates the binary edge prior into contextual features by the dot product and the pooling operation, it only recovers partial boundaries of regions.

Impact from Label Noise. As the pixel-level annotation is difficult and expensive, most of the face parsing benchmarks (*e.g*. Helen [35] and LaPa [21]) are annotated in a semi-automatic approach. Therefore, label noises inevitably exist in these datasets. As given in the last row of Figure 1, annotators mark the "eyes" as "glasses". Such annotation errors can limit the model performance, especially for tail classes (*e.g*. "necklace"). Nevertheless, the EARGNet method is a fully supervised method and lacks a regulation mechanism to tackle label noise.

3.2. Decoupled Multi-task Learning

Based on above analysis, we propose an end-to-end decoupled multi-task network to solve problems of spatial inconsistency and boundary confusion. Herein, we define three parallel tasks of face parsing, binary edge detection and category edge detection. To prevent using any pooling operation in context embedding, a customized GCN [13] module is designed to gain global contextual relationships for the parsing branch. To alleviate the boundary confusion, a binary edge detection branch as well as a category-aware semantic edge detection branch are jointly trained to gain rich edge information. During training, feature representations are simultaneously optimized for these three tasks, but the auxiliary edge prediction branches are removed during testing, without introducing any extra computation cost.

An overview of our model architecture is depicted in Figure 2. Given an input facial image, the ResNet-101 [8] pretrained on ImageNet is taken as backbone to extract features from different levels, marked as $\{C_1, C_2, C_3, C_4, C_5\}$. Afterwards, remaining parts involve: (1) a face parsing branch, which consists of a context embedding and a parsing head [36], (2) a binary edge detection branch utilizing the same edge decoder as [32], and (3) a category edge detection branch, which features abundant information of component edges. Each task shares same feature representations of first four layers in the backbone model. For the edge detection branches, feature maps from C_2, C_3 and C_4 are concatenated as input. For the parsing branch, context embedding features from C_5 are concatenated with the feature maps from C_2 as the input. Since edge branches preserve boundary information in low-level feature maps, joint edge prediction can assist high-level semantic predictions. At the testing phase, these two edge branches are decoupled from the whole model, avoiding extra computation overhead.

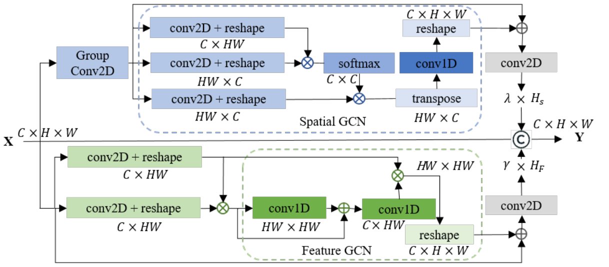


Figure 3. Illustration of the proposed DDGCN for context embedding. DDGCN is composed of two branches, and each consists of a Graph Convolutional Network (GCN) to model contextual information in the spatial-dimensions and feature-dimensions for a convolutional feature map X. No pooling step is involved in DDGCN to avoid spatial inconstancy.

Context Embedding without Pooling. Context embedding is crucial for face parsing [4, 39, 43, 45], but the pooling operation results in the problem of spatial inconsistency. To this end, we design a Dynamic Dual Graph Convolution Network (DDGCN), which exploits 1D convolution to build adjacent matrix of GCN over different 2D dimensions. As shown in Figure 3, the proposed DDGCN comprises one weighted GCN (labeled as H_S) with parameter λ in the spatial space and another weighted GCN (labeled as H_F) with parameter γ in the feature space.

$$Y = X \copyright (\lambda \times H_S) \copyright (\gamma \times H_F), \tag{1}$$

where © denotes the operation of concatenation. The parameters λ and γ are learnable weights for both H_S and H_F, respectively. Different from DGCN [44], we remove the pooling operation during coordinate space projection and, we merge spatial and channel features into the input X via a dynamic concatenation instead of the addition operation. To avoid buffer storage for gradient computation, all BN layers are replaced by Inplace-ABN [31]. As the proposed DDGCN is only applied to the C_5 feature map, our context embedding is more efficient than EAGRNet, which employs low-level features for graph representation learning.

Binary and Category Edge Assisted Face Parsing. As current training datasets for face parsing do not provide labels for the boundary detection, we first generate pseudo labels of binary and category-aware edges as illustrated in Figure 4. More specifically, binary edge pixels are identified from the pixel-wise label map by referring the neighboring four pixels. If there exists one neighboring pixel of zero value, the current pixel is regarded as an edge pixel. By employing the same criteria, the category-aware edges are generated independently for each facial component.

To learn shared features for the layers $\{C_1, C_2, C_3, C_4\}$ by simultaneously training the parsing and edge detection tasks, we design a loss function for each task and then sum them together with different weights. Different from the general semantic segmentation, face parsing features on tiny components. To retain the structure of small components, we also employ the Lovász-softmax [1] loss, which utilizes the mean intersection-over-union score to measure difference between ground truth and predicted mask. Hence, the cross-entropy [6] and Lovász-softmax [1] losses are combined together to optimize the parsing module. Additionally, the weighted cross-entropy [6] loss is employed to optimize both binary and category-aware edge detection. Consequently, the total multi-task loss is defined as

$$\mathcal{L}_{MT} = \underbrace{\lambda_0 \cdot (\mathcal{L}^p_{ce} + \mathcal{L}^p_{lovász})}_{parse} + \underbrace{\lambda_1 \cdot \mathcal{L}^b_{ce} + \lambda_2 \cdot \mathcal{L}^c_{ce}}_{edges}, \quad (2)$$

where $\mathcal{L}^b_{ce}$ and $\mathcal{L}^c_{ce}$ represent the weighted cross-entropy losses [6] corresponding to binary and category-aware semantic edges, respectively. The hyper-parameters λ_0, λ_1, and λ_2 denote the different weights for each task.

Besides the above parallel optimization, we also develop a boundary assisted semantic loss which enlarges the parsing loss of boundary pixels according to the binary and category-aware boundary maps. As edge maps are highly related to segmentation maps, it is beneficial to inject two types of edge cues into the parsing module to improve the segmentation accuracy for the components with clear contours. To this end, we define a dual edge attention loss

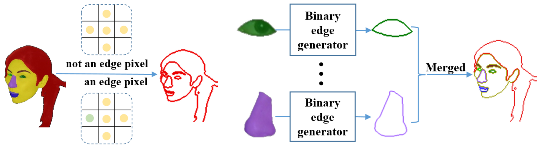
$$\mathcal{L}^b_{attn} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{b_i} * \mathcal{L}^p_i \odot B_i, \quad (3)$$

$$\mathcal{L}^c_{attn} = \frac{1}{NC}\sum_{i=1}^{N}\sum_{j=1}^{C} w_j * \frac{1}{c_{ij}} * \mathcal{L}^p_i \odot C_{ij}, \quad (4)$$

where N is the total number of images in a batch, b_i is the number of boundary pixels in a binary edge label map $B_i \in \mathbb{R}^{H\times W}$, c_{ij} is the number of boundary pixels of a specific category j in a category-aware edge label map $C_{ij} \in \mathbb{R}^{H\times W}$, w_j is a category-aware weight to emphasize a specific class j (*e.g.* the tail class of "necklace') which can increase the weights of tail classes, and $\mathcal{L}^p_i \in \mathbb{R}^{H\times W}$ is the cross-entropy between a predicted parsing result and the ground-truth. Different from the binary boundary attention loss proposed in [21], we further introduce category-aware boundary-attention semantic loss, significantly improving segmentation results of underrepresented classes.

The overall loss of our decoupled multi-task learning can be summarized as

$$\begin{aligned}\mathcal{L}_{DML} = &\underbrace{\lambda_0 \cdot (\mathcal{L}^p_{ce} + \mathcal{L}^p_{lovász})}_{parse} \\ &+ \underbrace{\lambda_1 \cdot \mathcal{L}^b_{ce} + \lambda_3 \cdot \mathcal{L}^b_{attn}}_{binary-edge} + \underbrace{\lambda_2 \cdot \mathcal{L}^c_{ce} + \lambda_4 \cdot \mathcal{L}^c_{attn}}_{category-edge},\end{aligned} \quad (5)$$



(a) Binary edge generation (b) Category-aware edge generation

Figure 4. Binary edge label generation and category-aware edge label generation from the pixel-wise label map.

where λ_3 and λ_4 correspond to weights of attention losses for binary and category-aware edges, respectively.

3.3. Cyclical Self-Regulation

To alleviate label noise, we introduce a Cyclical Self-regulation (CSR) training strategy to achieve online refinement labels. The proposed CSR depicted in Figure 2 includes two parts, self-ensemble and self-distillation.

Model Generalization via Self-Ensemble. As illustrated in the self-ensemble process of Figure 2, given a best model M_{best} from previous epochs and a set of next successive models $\{M_1, M_2, \dots, M_n\}$, a new model is obtained by aggregating weights of these models

$$M = \frac{k}{k+1}M_{best} + \frac{1}{(k+1)N}\sum_{n=1}^{N} M_n, \quad (6)$$

where k is the current cycle number and $1 \leq k \leq K$, and n is the number of models used in a cycle and $1 \leq n \leq N$. Moreover, symbols M, M_{best} and M_n represent the weights of aggregated, best and current models, respectively. In addition, all training data is forwarded into new aggregated model to re-estimate the statistical parameters in all Inplace-ABN [31] layers.

Label Refinement via Self-Distillation. As the soft labels contain dark knowledge [9] and less label noise, we explore self-distillation to improve the parsing performance. More specifically, as shown in the self-distillation process of Figure 2, the parsing results generated from the above aggregated model are exploited to supervise the multi-task learning. The total weighted loss is defined as

$$\mathcal{L}_{CSR} = \underbrace{\alpha_0 \cdot (\mathcal{L}^p_{kl} + \mathcal{L}^p_{lovász})}_{parse} + \underbrace{\alpha_1 \cdot \mathcal{L}^b_{kl} + \alpha_2 \cdot \mathcal{L}^c_{kl}}_{edges}, \quad (7)$$

where $\mathcal{L}^p_{kl}$, $\mathcal{L}^b_{kl}$, $\mathcal{L}^c_{kl}$ represent the Kullback-Leibler divergence losses [6] for face parsing, binary edge and category-aware edge tasks, respectively. They compute the difference between soft labels of the aggregated model and prediction results of the current model. Hyper-parameters α_0, α_1, α_2 are weights assigned to each task.

Finally, both self-ensemble and self-distillation processes mutually iterates in a cycle manner, promoting model generalization and correcting noisy labels progressively.

Method	Skin	Nose	U-lip	I-mouth	L-lip	Eyes	Brows	Mouth	Overall F1
Liu *et al.* [19]	92.1	93.0	74.3	79.2	81.7	86.8	77.0	89.1	88.6
Guo *et al.* [7]	93.8	94.1	75.8	83.7	83.1	80.4	87.1	92.4	90.5
Lin *et al.* [17]	94.5	95.6	79.6	86.7	89.8	89.6	83.1	95.0	92.4
Wei *et al.* [46]	95.6	95.2	80.0	86.7	86.4	89.0	82.6	93.6	91.6
Liu *et al.* [21]	94.9	95.8	83.7	89.1	**91.4**	89.8	83.5	**96.1**	93.1
Te *et al.* [36]	94.6	**96.1**	83.6	89.8	91.0	90.2	84.9	95.5	93.2
DML-CSR (Ours)	**96.6**	95.5	**87.6**	**91.2**	91.2	**90.9**	**88.5**	95.9	**93.8**

Table 1. Comparison with state-of-the-art methods on the Helen dataset in overall F1 score.

Method	Skin	Hair	L-Eye	R-Eye	U-lip	I-mouth	L-lip	Nose	L-Brow	R-Brow	Mean F1
Zhao *et al.* [45]	93.5	94.1	86.3	86.0	83.6	86.9	84.7	94.8	86.8	86.9	88.4
Liu *et al.* [21]	97.2	96.3	88.1	88.0	84.4	87.6	85.7	95.5	87.7	87.6	89.8
Te *et al.* [36]	97.3	96.2	89.5	90.0	**88.1**	90.0	89.0	97.1	86.5	87.0	91.1
DML-CSR (Ours)	**97.6**	**96.4**	**91.8**	**91.5**	88.0	**90.5**	**89.9**	**97.3**	**90.4**	**90.4**	**92.4**

Table 2. Comparison with state-of-the-art methods on the LaPa dataset in mean F1.

4. Experiments

Datasets. We use Helen [35], CelebAMask-HQ [15], and LaPa [21] for experiments. The Helen dataset contains 2,330 images with 11 labels: "background", "facial skin", "left/right brow", "left/right eye", "nose", "upper/lower lip", "inner mouth" and "hair". It is split into 2,000, 230 and 100 images for training, validation and testing. The CelebAMask-HQ dataset includes 24,183, 2,993, and 2,824 images for training, validation and testing. Apart from the 11 categories of the Helen dataset, the CelebAMask-HQ dataset adds extra 8 classes, including "left/right ear', "eyeglass", "earing", "necklace", "neck" and "cloth". The LaPa dataset features rich variations in expression, pose and occlusion, consisting of 11 categories as the Helen dataset. It is partitioned into 18,176 samples for training, 2,000 samples for validation, and 2,000 samples for testing.

Implementation Details. The proposed method is implemented by Pytorch [30], adopting the ResNet101 [8] as a backbone. The weights of the backbone are initialized with the pre-trained model on ImageNet [5]. Batch normalizations in our network are all replaced by In-Place Activated Batch Norm [31]. The input image size is 473×473 at both training and testing stages. During training, the data is augmented using: random rotation selecting an angle within (-30°, 30°) and random scaling with a factor from 0.75 to 1.25. We set the batch size as 28 and the network is trained for 200 epochs in total. The first 150 epochs are trained as initialization, following $K = 5$ cycles and each containing $N = 10$ epochs of the self-training process.

During the decoupled multi-task learning, we follow the similar training strategies as EAGRNet [36], *i.e.* Stochastic Gradient Descent (SGD) optimizer with the base learning rate 0.001 and the weight of decay of 0.0005. For the total loss function, weights of parsing, binary edge and category edge losses are set as $\lambda_0 = 1$, $\lambda_1 = 1$, and $\lambda_2 = 1$. respectively. To recover boundaries of tail classes (*e.g.* necklace and earring), weights $\lambda_3 = 4$ and $\lambda_4 = 1$ are assigned to both binary and category edge attention losses, respectively. For the cyclical self-regulation, the cosine annealing learning rate scheduler [16] with a learning rate of 10^{-5} is employed to optimize the model generalization. The weights of self-distillation losses for parsing, binary and category-aware edges are set to $\alpha_0 = 1$, $\alpha_1 = 1$ and $\alpha_2 = 0.1$.

Evaluation Metrics. To measure the performance of a face parsing model, two universally accepted evaluation metrics are employed, namely mean Intersection over Union (mIoU) and F1 score. To keep consistent comparison with the previous methods, the overall F1-score on the Helen dataset is calculated over the merged facial components: brows (left and right), eyes (left and right), nose, and mouth (upper lip, lower lip, and inner mouth). For the CelebAMask-HQ and LaPa datasets, the mean F1-score is computed over all categories excluding the background.

4.1. Comparison with State-of-the-art

In this paper, we thoroughly compare the performance of our proposed model with existing state-of-the-art methods (*i.e.* Zhao *et al.* [45], Liu *et al.* [21], Lee *et al.* [15], Luo *et al.* [23], Liu *et al.* [19], Guo *et al.* [7], Lin *et al.* [17], Wei *et al.* [46], and Te *et al.* [36]) on the Helen, LaPa and CelebAMask-HQ datasets. Statistical results in Table 1, Table 2, and Table 3 demonstrate that the proposed DML-CSR significantly outperforms other methods, achieving 93.8%, 92.4%, and 86.1% F1 scores on Helen, LaPa and CelebAMask-HQ, respectively. On the Lapa dataset, DML-CSR exhibits obvious advantages on eyebrow parsing. On the CelebAMask-HQ dataset, DML-CSR achieves much better performance on tail classes, such as "earring" and "necklace". Compared to EAGRNet [36], DML-CSR re-

Method	Face	Nose	Glasses	L-Eye	R-Eye	L-Brow	R-Brow	L-Ear	R-Ear	Mean F1
	I-Mouth	U-Lip	L-Lip	Hair	Hat	Earring	Necklace	Neck	Cloth	
Zhao *et al.* [45]	94.8	90.3	75.8	79.9	80.1	77.3	78.0	75.6	73.1	76.2
	89.8	87.1	88.8	90.4	58.2	65.7	19.4	82.7	64.2	
Lee *et al.* [15]	95.5	85.6	**92.9**	84.3	85.2	81.4	81.2	84.9	83.1	80.3
	63.4	88.9	90.1	86.6	**91.3**	63.2	26.1	**92.8**	68.3	
Luo *et al.* [23]	96.0	93.7	90.6	86.2	86.5	83.2	83.1	86.5	84.1	84.0
	93.8	88.6	90.3	93.9	85.9	67.8	30.1	88.8	83.5	
Te *et al.* [36]	**96.2**	**94.0**	92.3	88.6	88.7	**85.7**	85.2	88.0	85.7	85.1
	95.0	**88.9**	**91.2**	**94.9**	87.6	68.3	27.6	89.4	85.3	
DML-CSR (Ours)	95.7	93.9	92.6	**89.4**	**89.6**	85.5	**85.7**	**88.3**	**88.2**	**86.1**
	91.8	87.4	91.0	94.5	88.5	**71.4**	**40.6**	89.6	**85.7**	

Table 3. Comparison with state-of-the-art methods on the CelebAMask-HQ dataset in mean F1.

Baseline	DDGCN	DML	CSR	Helen		CelebAMask-HQ		LaPa	
				Mean IoU	Overall F1	Mean IoU	Mean F1	Mean IoU	Mean F1
✓				82.36	92.11	76.14	84.34	83.16	89.84
✓	✓			83.42 (+ 1.06)	92.56 (+ 0.45)	77.41 (+ 1.27)	85.33 (+ 0.99)	86.65 (+ 3.49)	92.10 (+ 2.26)
✓	✓	✓		85.48 (+ 3.12)	93.75 (+ 1.64)	77.69 (+ 1.55)	85.98 (+ 1.64)	87.00 (+ 3.84)	92.32 (+ 2.48)
✓	✓	✓	✓	**85.58** (+ 3.22)	**93.78** (+ 1.67)	**77.81** (+ 1.67)	**86.07** (+ 1.73)	**87.13** (+ 3.97)	**92.38** (+ 2.54)

Table 4. Ablation study of DML-CSR on the Helen, CelebAMask-HQ and LaPa datasets. Here, DDGCN is used for context embedding. DML denotes the multi-task learning for our decoupled model including face parsing, binary and category edge detection. CSR represents the cyclical self-regulation.

Method	Helen Overall F1	CelebAMask-HQ Mean F1	LaPa Mean F1
Baseline	92.11	84.34	89.84
+PSP [45]	92.20	84.76	90.80
+PSP-pooling	92.37	84.83	91.35
+DGCNet [44]	92.41	85.17	91.72
+DGCNet-pooling	92.45	85.20	91.99
+DDGCN	**92.56**	**85.33**	**92.10**

Table 5. Comparisons of different contextual modules on the parsing branch. Here, "+" means that the context embedding is added into the baseline, and "-pooling" denotes that the pooling operation is removed from the context embedding.

Method	Helen Overall F1	CelebAMask-HQ Mean F1	LaPa Mean F1
Baseline	92.11	84.34	89.84
+DML$_{p+b}$	93.35	85.58	92.16
+DML$_{p+b+ba}$	93.52	85.69	92.24
+DML$_{p+c}$	93.61	85.73	92.21
+DML$_{p+c+ca}$	93.71	85.87	92.28
+DML$_{p+b+c}$	93.65	85.80	92.26
+DML$_{all}$	**93.75**	**85.98**	**92.32**

Table 6. Results of our proposed multi-task learning on the Helen, CelebAMask-HQ and LaPa datasets. Here, "+" denotes adding multi-task branches into the baseline where DDGCN is used as context embedding. Losses of face parsing, binary edge detection, and category edge detection are denoted as $*_p$, $*_b$ and $*_c$ in the subscript. Binary edge attention and category edge attention losses are denoted as $*_{ba}$ and $*_{ca}$ in the subscript.

duces the parameters from 66.72M to 59.67M, and decreases the FLOP count from 51.63G to 48.54G. Given an image of the same input size as EAGRNet [36], DML-CSR dramatically shortens the inference time from 89ms to 31ms per image. In a word, DML-CSR utilizes fewer computation resources to outperform the state-of-the-art method.

4.2. Ablation Study

Analysis of Improvement. To illustrate the effect of individual modules and training strategy, the model after removing some components is trained from scratch under the same setting. The baseline method adopts the parsing module with a simple convolution unit, which includes a 3 × 3 convolution and an Inplace-ABN [31] to map features from the last layer of the backbone into new features of 256 dimensions. As shown in Table 4, our proposed DML-CSR substantially improves performance on face parsing. Compared to our baseline, adopting the DDGCN without any pooling operation as context embedding achieves a significant performance improvement. Then, appending semantic edge modules to enhance shared features has a further advance on parsing performance. The best results are obtained by training the decoupled multi-task network in a self-regulation mechanism, resulting in around 3.2% and 4.0% improvements of mean IoU on the Helen and LaPa datasets, respectively. Besides, it outperforms the baseline by around 1.7% overall F1 score improvement on the Helen dataset, and by over 2.5% mean F1 improvement on the LaPa dataset. On the CelebAMask-HQ dataset, DML-CSR also outperforms the baseline by around 1.7% improvement in both mean IoU and mean F1.

Comparison of Various Contextual Modules. To prove

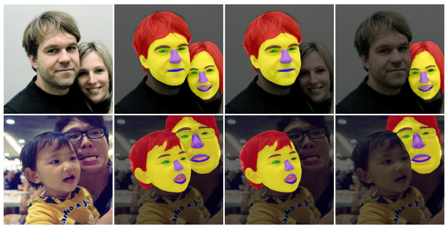


Figure 5. DML-CSR can easily distinguish different face instances under crowded scenarios due to the auxiliary category edge prediction. LaPa model is used here for visualization.

the effectiveness of our proposed DDGCN for learning contextual representation, the above-mentioned simple convolution unit in the baseline is substituted by various context embedding modules. Ablation experiments in Table 5 show that the pooling operation in PSP [45] and DGCNet [44] is harmful for the performance and the proposed DDGCN surpasses other contextual modules by dropping the pooling step and adopting dynamic feature fusion strategies.

Comparison of Different Auxiliary Tasks. Visual examples of Figure 5 show that auxiliary category edge modules can distinguish boundaries between facial components and different faces. To further explore the effect of category edge detection, several related experiments are executed. As we can see from the results listed in Table 6, both the binary edge detection branch and the category-aware edge detection branch can obviously improve the performance of face parsing. However, the category-aware edge is more informative than binary edge, thus it is more beneficial for face parsing. Besides, our proposed dual edge attention loss on the equation (4) further improves overall performance of face parsing on three benchmark testing datasets.

Analysis of Visual Results. To better understand the effect of the proposed methods step-by-step, we present visual examples in Figure 6. The second-column visual examples show that our baseline obviously address the issue of spatial inconsistency. However, examples in column (b) appear severe unclear boundaries between different facial components in the first three green boxes, and confusing contours of multi-faces in the last three green boxes. This is due to the fact that the baseline lacks a reasoning ability on global dependencies. The first three-row examples in column (c) show complete structure of individual component and almost clear boundaries between facial parts, illustrating the long-range inference ability of our proposed DDGCN. Nonetheless, the last three-row examples in the column (c) still exist different face instances, as the proposed DDGCN has a limited capability of localizing objects of similar contours. Compared to examples in columns (b)-(c), columns (d)-(e) present clear boundaries between different facial components in both single-face and multi-face scenes, due to the feature enhancement by semantic edges. Looking at the areas within green rectangles in columns (d)-(e), CSR can recover error pixels, preventing noisy labels in training datasets from degrading model generalization.

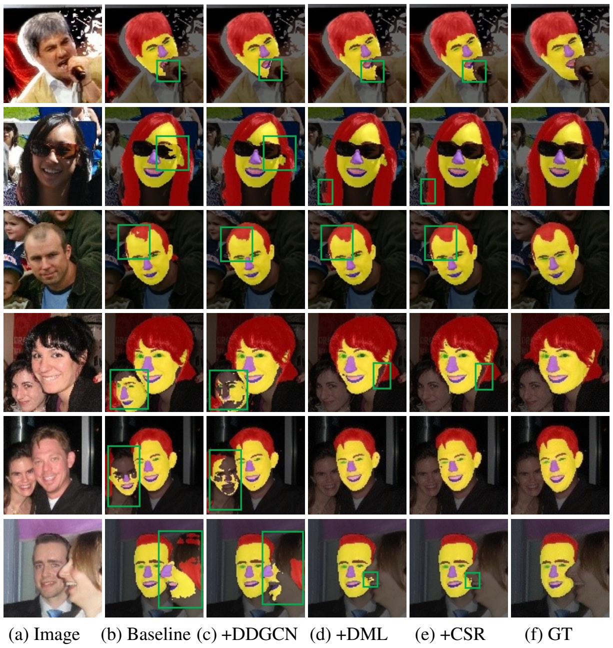


Figure 6. DML-CSR can obtain complete facial components with clear boundaries in both single-face and multi-face scenarios. Visual examples in different columns are generated by the corresponding LaPa models. Here, "+" denotes that the current component is added into the model in the previous column.

5. Conclusion

In this paper, we present DML-CSR, a decoupled multi-task learning method with cyclical self-regulation for face parsing. Comprehensive experiments on Helen, CelebAMask-HQ, and LaPa verify the effectiveness of the proposed method. The results show that DML-CSR significantly outperforms other methods on all datasets. Training details will be released to encourage further research towards face parsing.

Limitations. Our method achieves impressive results in face parsing. However, there is a slight performance degradation in low-resolution faces. This is because that we train our model on the high-resolution face dataset. Even so, we believe DML-CSR is a valuable method for training a reliable face parsing model on a large-scale dataset.

Societal Impact. We develop a general model for face parsing in this paper, and the proposed model is not used for a specific application. Therefore, this work does not directly involve societal issues.

Acknowledgements. Stefanos Zafeiriou acknowledges support from the EPSRC Fellowship DEFORM (EP/S010203/1), FACER2VM (EP/N007743/1) and a Google Faculty Fellowship. Ying Li acknowledges support from the National Natural Science Foundation of China (61871460).

References

[1] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, 2018. 5

[2] Rich Caruana. Multitask learning. *Machine Learning*, 1997. 3

[3] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin D. Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *ECCV*, 2020. 2

[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. *TPAMI*, 2018. 4

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. 5

[7] Tianchu Guo, Youngsung Kim, Hui Zhang, Deheng Qian, ByungIn Yoo, Jingtao Xu, Dongqing Zou, Jae-Joon Han, and Changkyu Choi. Residual encoder decoder network and adaptive prior for face parsing. In *AAAI*, 2018. 1, 6

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 4, 6

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015. 5

[10] Aaron S. Jackson, Michel Valstar, and Georgios Tzimiropoulos. A cnn cascade for landmark guided semantic part segmentation. In *ECCV*, 2016. 1, 2

[11] Warrell Jonathan and Simon J.D. Prince. Labelfaces: Parsing facial features by multiclass labeling with an epitome prior. In *ICIP*, 2009. 2

[12] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 3

[13] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 4

[14] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, 2017. 3

[15] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 1, 2, 3, 6, 7

[16] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *TPAMI*, 2020. 2, 6

[17] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. Face parsing with roi tanh-warping. In *CVPR*, 2019. 2, 6

[18] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Roi tanh-polar transformer network for face parsing in the wild. *IVC*, 2021. 1

[19] Sifei Liu, Jianping Shi, Liang Ji, and Ming-Hsuan Yang. Face parsing via recurrent propagation. In *BMVC*, 2017. 1, 2, 6

[20] Sifei Liu, Jimei Yang, Chang Huang, and Ming-Hsuan Yang. Multi-objective convolutional learning for face labeling. In *CVPR*, 2015. 1, 2

[21] Yinglu Liu, Hailin Shi, Hao Shen, Yue Si, Xiaobo Wang, and Tao Mei. A new dataset and boundary-attention semantic segmentation for face parsing. In *AAAI*, 2020. 1, 2, 3, 4, 5, 6

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[23] Ling Luo, Dingyu Xue, and Xinglong Feng. Ehanet: An effective hierarchical aggregation network for face parsing. *Applied Sciences*, 2020. 1, 6, 7

[24] Ping Luo, Xiaogang Wang, and Xiaoou Tang. Hierarchical face parsing via deep learning. In *CVPR*, 2012. 2

[25] Daniel Merget, Matthias Rock, and Gerhard Rigoll. Robust facial landmark detection via a fully-convolutional local-global context network. In *CVPR*, 2018. 1

[26] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, 2016. 3

[27] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Fast scene understanding for autonomous driving. *arXiv:1708.02550*, 2017. 3

[28] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *ICCV*, 2019. 1

[29] Xinyu Ou, Si Liu, Xiaochun Cao, and Hefei Ling. Beauty emakeup: A deep makeup transfer system. In *ACM MM*, 2016. 1

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zach DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NeurIPS Workshop*, 2017. 6

[31] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. In *CVPR*, 2018. 4, 5, 6, 7

[32] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *AAAI*, 2019. 4

[33] Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. Latent multi-task architecture learning. In *AAAI*, 2019. 3

[34] H.J. Scudder. Probability of error of some adaptive pattern-recognition machines. *Information Theory*, 1965. 2

[35] Brandon Smith, li Zhang, Jonathan Brandt, Zhe Lin, and Jianchao Yang. Exemplar-based face parsing. In *CVPR*, 2013. 1, 2, 3, 4, 6

[36] Gusi Te, Yinglu Liu, Wei Hu, Hailin Shi, and Tao Mei. Edge-aware graph representation learning and reasoning for face parsing. In *ECCV*, 2020. 1, 2, 3, 4, 6, 7

[37] Marvin Teichmann, Michael Weber, Marius Zöllner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *IV*, 2018. 3

[38] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van

Gool. Multi-task learning for dense prediction tasks: A survey. *TPAMI*, 2021. 3

[39] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 4

[40] Zhen Wei, Yao Sun, Jinqiao Wang, Hanjiang Lai, and Si Liu. Learning adaptive receptive fields for deep image parsing network. In *CVPR*, 2017. 2

[41] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv:1905.00546*, 2019. 2

[42] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*, 1995. 2

[43] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 4

[44] Li Zhang, Xiangtai Li, Anurag Arnab, Kuiyuan Yang, Yunhai Tong, and Philip HS Torr. Dual graph convolutional network for semantic segmentation. In *BMVC*, 2019. 4, 7, 8

[45] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2, 3, 4, 6, 7, 8

[46] Wei Zhen, Si Liu, Yao Sun, and Hefei Ling. Accurate facial image parsing at real-time speed. *TIP*, 2019. 6

[47] Lei Zhou, Zhi Liu, and Xiangjian He. Face parsing via a fully-convolutional continuous crf neural network. *arXiv:1708.03736*, 2017. 1, 2

[48] Yisu Zhou, Xiaolin Hu, and Bo Zhang. Interlinked convolutional neural networks for face parsing. In *ISNN*, 2015. 1, 2

[49] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D. Cubuk, and Quoc V. Le. Rethinking pre-training and self-training. In *NeurIPS*, 2020. 2