RoI Tanh-polar Transformer Network for Face Parsing in the Wild

Yiming Lin, Jie Shen*, Yujiang Wang, Maja Pantic

Department of Computing, Imperial College London, UK

Abstract

Face parsing aims to predict pixel-wise labels for facial components of a target face in an image. Existing approaches usually crop the target face from the input image with respect to a bounding box calculated during pre-processing, and thus can only parse inner facial Regions of Interest (RoIs). Peripheral regions like hair are ignored and nearby faces that are partially included in the bounding box can cause distractions. Moreover, these methods are only trained and evaluated on near-frontal portrait images and thus their performance for in-the-wild cases has been unexplored. To address these issues, this paper makes three contributions. First, we introduce iBugMask dataset for face parsing in the wild, which consists of 21,866 training images and 1,000 testing images. The training images are obtained by augmenting an existing dataset with large face poses. The testing images are manually annotated with 11 facial regions and there are large variations in sizes, poses, expressions and background. Second, we propose RoI Tanh-polar transform that warps the whole image to a Tanh-polar representation with a fixed ratio between the face area and the context, guided by the target bounding box. The new representation contains all information in the original image, and allows for rotation equivariance in the convolutional neural networks (CNNs). Third, we propose a hybrid residual representation learning block, coined HybridBlock, that contains convolutional layers in both the Tanh-polar space and the Tanh-Cartesian space, allowing for receptive fields of different shapes in CNNs. Through extensive experiments, we show that the proposed method improves the state-of-the-art for face parsing in the wild and does not require facial landmarks for alignment.

Keywords: Face parsing, in-the-wild dataset, head pose augmentation, Tanh-polar representation

1. Introduction

Face parsing is a fundamental facial analysis task: it predicts per-pixel semantic labels in a target face. It provides useful features for many downstream applications, such as face recognition [1, 2], face beautification [3], face swapping [4, 5], face synthesis [6, 7, 8, 9, 10], facial attribute recognition [11, 12, 13], and facial medical analysis [14]. Recently, methods based on deep Convolutional Neural Networks (CNNs), especially Fully Convolutional Networks (FCNs) [15], have achieved impressive results on this task.[16, 17, 18, 19, 20, 21, 22].

Two unique aspects distinguish face parsing from generic image parsing. The first is that face parsing does not parse the entire image but only the target face specified by a bounding box, whereas image parsing predicts a label for every pixel in the image. How to preprocess the input with bounding boxes remains an underexplored problem. Most methods [23, 24] crop the face with a fixed margin and resize the cropped patch to the same dimension, which we refer to as *crop-and-resize* the face area. Such methods ignore hair because the margin around the hair area is hard to determine. If the selected margin is too small, the hair region would be cut off. If it is too big, too many background pixels and / or nearby faces could be included in the cropped patch, causing significant distractions to the model. Another pre-processing method is to use facial landmarks for face alignment [16] such that the face is appropriately rotated. We refer to this method as *align*. The landmarks can be jointly obtained with the face bounding boxes [25]. The main problem is that a good template has to be carefully chosen for alignment.

The other challenge is that in-the-wild images are

*Corresponding author

Email addresses: yiming.lin15@imperial.ac.uk (Yiming Lin), jie.shen07@imperial.ac.uk (Jie Shen), yujiang.wang14@imperial.ac.uk (Yujiang Wang), maja.pantic@gmail.com (Maja Pantic)

underrepresented in existing benchmarks. Four most widely used face parsing datasets are Helen [26], LFW-PL [27], CelebAMask-HQ [28] and LaPa [29]. Figure 2 shows exemple images from these datasets and Section 3 compares them in detail. Most images from Helen and LaPa are portraits, which means that only one large face in frontal view is present near the centre. CelebAMask-HQ contains images synthesised from CelebA [30] using super-resolution. All faces are aliened using landmarks and resized to the same size. As such, the resulted images contain very little context information. Similarly, LFW-PL is a subset of LFW [31] and the faces are aligned using landmarks. Hence, how models perform under in-the-wild conditions remains unexplored.

To tackle the first challenge, we propose RoI Tanh-polar transform (RT-Transform), that transforms the entire image into a fixed-size representation in the Tanh-polar coordinate system based on the target bounding box. Figure 6 illustrates the transform process. As a fully invertible transform, it preserves all contextual information. Moreover, regardless of the face's actual size in the input image, the ratio between the face and the background remains fixed at $76\% : 24\%$ in the transformed representation. In the Tanh-polar coordinate system, planar convolutions correspond to group-convolutions [32] in rotation. Thus, Convolutional Neural Networks (CNNs) applied in the tanh-polar space would produce a representation that is equivariant to rotations in the original Cartesian space.

We further introduce Hybrid Residual Representation Learning Block (HybridBlock) that uses RT-Transform to create hybrid representations in the residual blocks. A HybridBlock consists of two 3×3 convolutional layers, one in the Tanh-polar coordinate system and the other in the Tanh-Cartesian coordinate system. They are operating on different-shaped receptive fields and their outputs are concatenated in the Tanh-polar system to obtain a hybrid representation. By stacking HybridBlocks, we arrive at HybridNet, a backbone network that takes as input an image transformed by RT-Transform and the target bounding box, and outputs a hybrid representation for face parsing. We then add the vanilla FCN decoder and inverse RT-Transform. The resulting framework, called RoI Tanh-polar Transformer Network (RT-Net), is shown in Figure 1.

To tackle the second challenge of lacking suitable benchmarks, we present iBugMask dataset that contains $22,866$ in-the-wild images. For the training set of $21,866$ images, we use a face profiling method [33] to rotate the faces from images in Helen dataset with respect to the yaw angle, creating many large-pose and profile faces. For the $1,000$ testing images, per-pixel manual annotations for 11 regions including hair are provided. The curated images contain large variations in pose, expression, size and background clutter (see Figure 2). Extensive experiments show that iBugMask dataset is more challenging than other benchmarks and models trained on iBugMask improves performance under both intra-dataset and cross-dataset evaluation.

In summary, we offer the following contributions:

- We propose RoI Tanh-polar Transform for face parsing in the wild that transforms the target face to the Tanh-polar coordinate system based on the bounding box, preserving the context and allowing CNNs to learn representations equivariant to rotations.
- We propose Hybrid Residual Representation Learning Blocks, that extracts a hybrid representation by applying convolutions in both Tanh-polar and Tanh-Cartesian coordinates.
- We present iBugMask dataset, a novel in-the-wild face parsing benchmark that consists of more than 22 thousand images.
- We conduct extensive experiments and show that the overall framework RTNet improves the state-of-the-art on all benchmarks.

2. Related Work

Increasing research effort has been devoted to face parsing due to its potential application in various face analysis tasks. In this section, we briefly review four groups of relevant works, *i.e.* 1) the face parsing benchmarks, 2) the face parsing methods, 3) works on scene parsing and 4) representation learning in polar space.

2.1. Face Parsing Benchmarks

Publicly available face parsing benchmarks are comparatively scarce, mainly due to the significant amount of effort required for pixel-level annotations. Currently, two most widely-used benchmarks are LFW-PL [27] and Helen [26].

The Helen dataset [34, 26] includes $2,330$ facial images with 194 landmark annotations that are obtained through Amazon Mechanical Truck. In this dataset, the inner facial components including eyes, eyebrows, nose, inner mouth and upper/lower lips are manually annotated by human, while the ground-truths for the rest facial parts, *i.e.* facial skin and hairs, are generated via

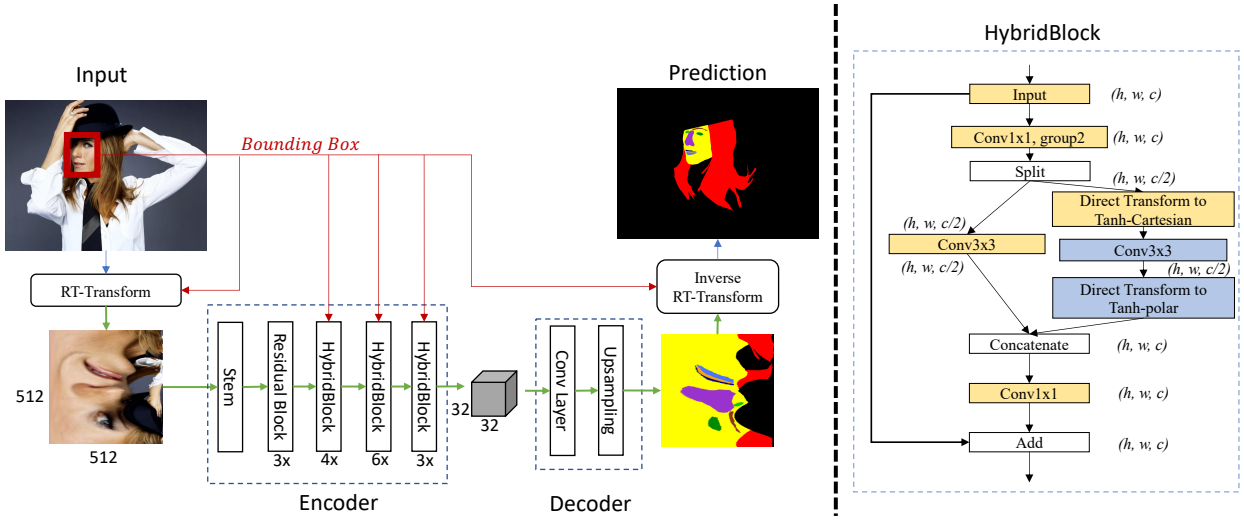


Figure 1: *Left:* RoI Tanh-polar Transformer Network (RTNet): a facial image is transformed to Tanh-polar coordinates with RT-Transform. The encoder consists of a Stem layer, one stage of Residual Blocks and three stages of HybridBlocks. Bounding box is used in RT-transform and HybridBlocks to warp tensors between Tanh-polar and Tanh-Cartesian coordinates. The decoder consists of a Conv layer and a Bilinear Upsampling layer. The output mask is transformed back to Cartesian coordinates using inverse RT-Transform. *Right:* HybridBlock. Yellow rectangles are layers in Tanh-polar space while blue ones are layers in Tanh-Cartesian space. Tuples (h, w, c) are the shape of the output tensor for each operation. "Split" and "Concatenate" operations are performed along the channel dimension (see Section 4.2).

image matting algorithms [26] and may not be fully accurate. Despite such disadvantages, Helen was still the only publicly-available face parsing dataset with an acceptable amount of training data for several years, and thus it has been a popular choice for evaluating face parsing methods [16, 24, 35].

The LFW-PL dataset [27] consists of 2,972 facial images selected from the Labeled Faces in the Wild (LFW) dataset [31]. To obtain dense annotations, each facial image is first automatically segmented into superpixels and those superpixels are subsequently labelled as one of the following categories: facial skin, hair and background.

Recently, two large-scale face parsing datasets were released, which are CelebAMask-HQ [28] and LaPa [29]. Although the number of annotated samples are greatly increased, the facial images included in those two datasets are not strictly in-the-wild, since they have already been pre-processed in an unrecoverable way. In CelebAMask-HQ, the resolutions of facial images are intentionally enlarged through the super-resolution technique [36], while most faces are aligned to be frontal and centralised. Besides, the background region usually comprises a small portion of the whole facial image, *i.e.* most environmental information has been discarded. Similar situations can be discovered in the Lapa dataset in which faces are also cropped and aligned with limited background information preserved.

Compared to those datasets, our proposed iBugMask dataset is the only face parsing benchmark consisting of fully in-the-wild images. The facial samples are neither cropped nor aligned, and we also preserve almost all the background information. It covers large variations in poses, illuminations, occlusions and scenes. A detailed comparison between the iBugMask dataset and the existing benchmarks is provided in Sec. 3.

2.2. Face Parsing Methods

Face parsing is the task of pixel-wisely labelling given facial images. Earlier works [37, 26] on face parsing usually leveraged holistic priors and hand-crafted features. Warrell *et al.* [37] modelled the spatial correlations of facial parts with Conditional Random Fields (CRFs). Smith *et al.* [26] applied SIFT features to select exemplars in facial parts and propagate the labels of these exemplars to generate complete segmentation maps. A hybrid method was proposed in [27] that combined the strength of both CRFs and Restricted Boltzmann Machine in a single framework to model global and local facial structures. The idea of utilising engineering-based features can also been seen in other works [38, 39, 40]. Those approaches are typically time-consuming and cannot generalise well to different scenarios, and thus they have been gradually replaced by deep-learning-based methods with encouraging performance.

(a) CelebAMask-HQ (b) LFW-PL (c) Helen (d) LaPa (e) iBugMask

Figure 2: Examples from benchmarks with colour-coded labels (best viewed in colour). CelebAMask-HQ [28] and LFW-PL [27] contain well-aligned faces and little context information. Helen [26] contains mostly portrait images where faces are big and near the centre. LaPa [29] contains face images with some variations in pose and occlusion but the faces are cropped and centred. By contrast, iBugMask contains large variations in expression, pose and background and all background information is preserved (see Section 3).

State-of-the-art performance on face parsing is mostly achieved by deep learning methods. Liu *et al.* [18] incorporated CNNs into CRFs and proposed a multi-objective learning method to model pixel-wise likelihoods and label dependencies jointly. An interlinked CNN was present in [41] to detect different facial parts, while this architecture cannot generate semantic labels for large-scale components like facial skin. Luo *et al.* [17] applied multiple Deep Belief Networks to detect facial parts and accordingly built a hierarchical face parsing framework. Jackson *et al.* [21] employed facial landmarks as a shape constraint to guide Fully Convolution Networks (FCNs) for face parsing. Multiple deep methods including CRFs, Recurrent Neural Networks (RNNs) and Generative Adversarial Networks (GAN) were integrated by authors of [42] to formulate an end-to-end trainable face parsing model, while the facial landmarks also served as the shape constraints for segmentation predictions. The idea of leveraging shape priors to regularise segmentation masks can also be found in the Shape Constrained Network (SCN) [43] for eye segmentation. In [24], a spatial Recurrent Neural Networks was used to model the spatial relations within face segmentation masks. A spatial consensus learning technique was explored in [19] to model the relations between output pixels, while graph models was adopted in [35] to learn the implicit relationships between facial components. To better utilise the temporal information of sequential data, authors of [44] integrated ConvLSTM [45] with the FCN model [15] to simultaneously learn the spatial-temporal information in face videos and to obtain temporally-smoothed face masks. In [46], a Reinforcement-Learning-based key scheduler was introduced to select online key frames for video face segmentation such that the overall efficiency can be globally optimised.

Most of those methods assume the target face has already been cropped out and are well aligned. Moreover, they often ignore the hair class due to the unpredictable margins for cropping the hair region. The most related work to our paper is the RoI Tanh Warping [16] which proposed to warp the entire image using the Tanh function. However, there are several limitations in this work. The warping operation requires not only the facial bounding boxes but also the facial landmarks, which can be overly redundant. Additionally, the warped image is still in Cartesian space and cannot benefit from the rotation-equivariant property in polar space. Moreover, multiple sub-networks are employed to learned the shapes of inner facial parts and hair separately, and these sub-networks need to be trained with different loss functions, making the pipeline trivial.

2.3. Scene Parsing

Scene parsing is to segment an image into different image regions associated with semantic scene labels such as roads, pedestrians, cars, etc. Fully Convolutional Networks (FCNs) [15] is the first critical milestone of applying deep learning techniques in this field. Via replacing fully connected layers with convolutional ones, FCNs successfully adapt classical CNN classification models like VGG-16 [47] or ResNet [48] to solve scene parsing tasks. Following FCNs [15], a

wide variety of scene parsing methods have been developed, including the application of dilated (*atrous*) convolutions [49, 50, 51, 52], the encoder and decoder structures [53, 54, 55], the spatial pyramid architectures [56, 54, 57] the involvement of attention mechanisms [58, 59, 60], utilising Neural Architecture Search (NAS) techniques [61, 62], etc. PSPNet [56] proposed a spatial pyramid pooling module that adopts a set of spatial pooling operations of different sizes to increase the variety of receptive fields of the network. SPNet [57] extended the pooling module by introducing the strip pooling module to capture long-range banded context. Deeplab family [51, 54] devised an Atrous Spatial Pyramid Pooling (ASPP) module that consists of three parallel dilated convolutional layers to capture multi-scale context. UNet [63] introduced skip connections between the encoder and the decoder sub-networks to preserve low-level details, while the details of high resolution features were maintained in HRNet [64] by branching the backbone network. BiSeNet [65] proposed a bilateral network consisting of a context branch and a spatial branch. Such a two-branch architecture allows BiSeNet to operate with satisfying efficiency while achieving the state-of-the-art performance. Readers are referred to [66] for a more detailed review of scene parsing techniques.

2.4. Polar Representation Learning

Compared with Cartesian coordinate system, polar or log-polar space are not sensitive to certain transformations such as rotations and scaling, and therefore polar representations have been widely studied in image processing and computer vision. Early applications of polar transformations included face detection [67], face tracking [68], face recognition [69], the aggregation of hand-crafted descriptors [70, 71], etc.

Recently, how to integrate polar representations with deep CNN models have been increasingly explored. The traditional CNN architectures can be insensitive to translations, i.e. translation equivalence, yet it is not the case for other transformations such as rotations and scaling. Representation learning in polar space, on the other hand, can effectively overcome such limitations through its equivariances to rotations and scales. Polar Transformer Networks (PTN) [72] is one of the first attempts to construct a CNN model that maps Cartesian-based images into polar coordinates for better tolerances to transformations like rotations and dilation. In PTN, a shallow network consisting of several 1×1 convolutional layers first scans the whole image to predict a polar origin. This predicted origin together with the input image are then fed into a differentiable polar transformer module to generate image representations in log-polar systems. The obtained polar representation is invariant with respect to the original object locations while rotations and dilations are now shifts, which are handled equivariantly by a conventional classifier CNN. Ebel *et al.* [73] utilises PTN to extract polar-based local descriptors for key-point matching, leading to more robust performance. Different from those works, our RoI Tanh-polar Transformer network warps the whole image into a Tanh-polar representation that can emphasise the Region of Interests (RoI) through oversampling in RoI areas and undersampling in the rest.

3. Dataset

In this section, we introduce a new in-the-wild face parsing benchmark, iBugMask, that consists of a pose-augmented training set and a manully-annotated testing set. We compare their characteristics with existing face parsing datasets. The main motivation for the new benchmark is that existing benchmarks only contain faces with limited variations in expression, pose and context information, which makes them less suitable for capturing characteristics of real-world face images. Moreover, large scale training data is key to the success of CNN-based models, but existing face parsing datasets do not provide sufficient training data for such challenging cases.

3.1. Overview of Existing Benchmarks

Figure 2 shows exemplar images from different face parsing benchmarks with their colour-coded labels overlaid.

CelebAMask-HQ [28] contains $30,000$ synthesised faces from the CelebA dataset [30]. All images are scaled to 512×512 and faces are well-aligned at the centre using facial landmarks. Background information is either removed or blurred. 19 facial classes are labelled: background, skin, left/right brow, left/right eye, upper/lower lip, left/right ear, nose, inner mouth, hair, hat, eyeglass, earring, necklace, neck, and cloth. The dataset is divided into $24,183$ images for the training, $2,993$ images for the validation, and $2,824$ images for the testing.

LFW-PL [27] contains $2,927$ images of resolution 250×250 with faces aligned in the centre. Face and hair regions are annotated using superpixel-based methods, thus resulting in inaccurate labels. The dataset is divided into $1,500$ images for the training, 500 images for the validation, and 927 images for the testing.

Helen [26] is the most popular face parsing benchmark and it contains $2,330$ real-world images with rich

context information. 11 semantic labels are annotated: background, skin, left/right brow, left/right eye, upper/lower lip, inner mouth, nose and hair. There are significant annotation errors for facial skin and hair classes in the training set, as discussed in [16], because these labels were automatically generated using image matting. The authors of Helen only cleaned the testing set to guarantee fair comparison in the test set. Helen dataset is divided into 2,000 images for the training, 230 images for the validation and 100 images for the testing.

LaPa [29] is a face parsing dataset containing more than 22,176 facial images with relatively more variations in expression, pose and occlusion. The same 11 semantic classes are annotated as in Helen and the annotation process was guided by 106-point facial landmarks. The dataset is divided into 18,176 images for the training, 2,000 images for the validation, and 2,000 images for the testing. The faces contain some variations in pose and occlusion but the background and hair region are largely removed since the faces are cropped with a hand-picked margin.

3.2. *iBugMask: An In-the-wild Face Parsing Benchmark*

The proposed iBugMask consists of two parts: a training set obtained by pose augmentation and a manually curated testing set. The training set contains 21,866 images while the testing set contains 1,000 images. We describe these two parts in detail below.

3.2.1. *A Large-Pose Augmented Training Set*

For machine learning models to learn to parse faces with large variations in head poses, the training set needs to contain a balanced distribution over poses. However, existing datasets contain faces mostly with absolute yaw angles less than 45 degrees. This means that models trained on these datasets cannot handle faces with extreme poses.

We propose to solve this problem by synthesising training faces with large poses. First, we examined the training set Helen, and manually corrected the labelling errors as in [16]. Next, we augmented the data with a face profiling method [33] that has been applied to augment face alignment datasets. One major advantage of this method is that it creates 3D meshes for both internal face and external face regions, which preserve the unpredictable hair regions as well as important context information for face parsing. Through face profiling, we augment the training set of Helen to a large scale one with many faces having large variation in head poses.

Figure 3: Examples of face data augmentation using 3DDFA [33]. The first column shows the original images and the other three columns show synthesised images with different Δyaw until $yaw = 90°$.

With the fitted 3D model, we gradually enlarge the yaw angle of image at the step of 5° until 90°. Considering that the fidelity of a synthesised face is negatively related with the Δyaw, we resample the augmented images of each face with probabilities $0.8^{\Delta yaw/5°}$. In Table 1, we compare our training set with other training sets and show that ours contains much more variations in pose, facial expression, and background. We conduct extensive experiments in Sec. 5. The results show that all models trained on our augmented training set improve over their counterparts trained on other datasets for in-the-wild face parsing.

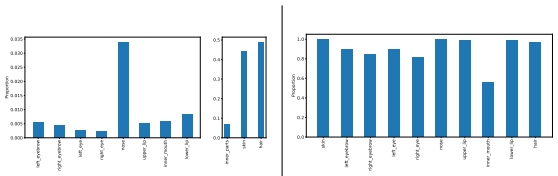
Benchmark	Number of images	In-the-wild	Non-neutral & non-smile	$\|yaw\|$ ≥ 30°	$\|yaw\|$ ≥ 60°
Helen [26]	2,000	✓	40.1%	120	4
CelebAMask-HQ [28]	27,176	✗	44.6%	1,565	60
LFW-PL [27]	2,000	✗	74.7%	125	0
LaPa [29]	18,176	✗	39.7%	3,961	194
iBugMask (ours)	21,866	✓	34.6%	14,692	6,880

Table 1: Comparison of training sets. Ours has large variations in pose, expression and background.

3.2.2. *A Manually Curated Testing Set*

In in-the-wild images, faces can appear at any location in an image, with various distracting contextual information around it. In existing benchmarks, the target faces are cropped and centred by the data providers, largely removing background, part of hair and other faces. This introduces bias and evaluating methods on pre-processed images does not honestly reveal their robustness to the distracting context noise.

To fairly evaluate face parsing models under in-the-wild conditions, we present iBugMask dataset that contains 1,000 challenging face images and manually-annotated labels for 11 semantic classes: background,



(a) Pixel distribution for facial parts

(b) Region distribution for facial parts

Figure 4: Distributions for facial parts in iBugMask. Left: pixel distribution for facial parts. We compare different facial parts in the first subplot. We merge the inner parts to compare with skin and hair in the second subplot. Right: region distribution for facial parts. Inner mouth the least seen region in the dataset.

facial skin, left/right brow, left/right eye, nose, upper/lower lip, inner mouth and hair. The images are curated from challenging in-the-wild face alignment datasets, including 300W [74] and Menpo [75]. Compared to the existing face parsing datasets, iBugMask contains in-the-wild scenarios such as "party" and "conference", which include more challenging appearance variations or multiple faces. There is a larger number of profile faces. More expressions other than "neutral" and "smile" are also included (*e.g.* "surprise" and "scream"). Examples can be found in the rightmost column of Figure 2. Table 2 compares characteristics of different benchmarks. We use 3DDFA [33] to estimate the yaw angles with facial landmarks obtained by FAN [76]. We use the facial expression classifier proposed by Want *et al.* [77] to estimate the facial expressions. Figure 5 shows the absolute yaw angle distributions of benchmarks. Finally, Figure 4 shows the pixel and region distributions in the testing set.

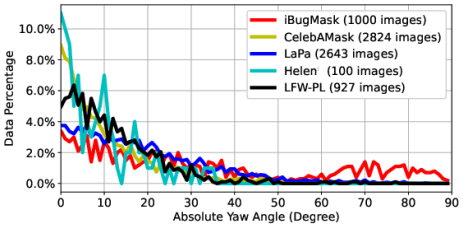


Figure 5: Absolute yaw angle distributions of different testing sets. Yaw is estimated with 3DDFA [33].

4. Methodology

We introduce the RoI Tanh-polar Transformer Network for face parsing in the wild. Figure 1 shows the overall framework: given an in-the-wild image in Cartesian coordinates and a bounding box of the target face,

Benchmark	Number of images	In-the-wild	Non-neutral & non-happy	$\|yaw\|$ ≥ 30°	$\|yaw\|$ ≥ 60°
Helen [26]	100	✓	40.0%	4	0
CelebAMask-HQ [28]	2,824	✗	42.4%	180	6
LFW-PL [27]	927	✗	73.3%	53	7
LaPa [29]	927	✗	40.8%	525	48
iBugMask (ours)	1,000	✓	56.1%	413	241

Table 2: Comparison of existing benchmark datasets. iBugMask has large variations in pose, expression and background.

the whole image is first projected into the Tanh-polar space through the proposed RoI Tanh-Polar Transform in Section 4.1. We further introduce a deep CNN encoder named HybridNet to extract semantic features of the Tanh-Polar-warped image. Consisting of several Hybrid Residual Representation Learning blocks (Section 4.2), the proposed HybridNet takes advantages of both Tanh-Cartesian and Tanh-Polar coordinate systems and thus can generate more robust spatial features. Those features are fed into a FCN decoder [15] to obtain Tanh-polar-based segmentation masks which are then mapped back into the Cartesian coordinate as the final output.

4.1. RoI Tanh-Polar Transform

4.1.1. To Crop or Not To Crop?

In in-the-wild face parsing, the target face is specified by a bounding box and is often not centralised. A common pre-processing step is to extend the facial bounding box with a certain margin and then to crop out the facial images, which are further resized into a certain resolution depending on the employed deep models. We refer to this pre-processing technique as *crop-and-resize*. In this pre-processing approach, however, the cropping margin needs to be carefully selected. An overly loose margin may introduce irrelevant and distracting information, e.g. other faces, while a margin that are too narrow can lead to the ignorance of useful image regions like hairs, both of which are undesirable in the face parsing task. Another pre-processing method is to use facial landmarks for face alignment [30, 16] such that the face is appropriately rotated. We refer to this method as *align*. The landmarks can be jointly obtained with the face bounding boxes [25].

To overcome the limitations in the *crop-and-resize* method and eliminate the need for facial landmarks, we propose the RoI Tanh-polar transform that warps the whole image to a canonical representation in the Tanh-polar space. Compared with the classical *crop-and-resize* and *align*, the only prerequisite of our method is the detected bounding box. Besides, our mapping can

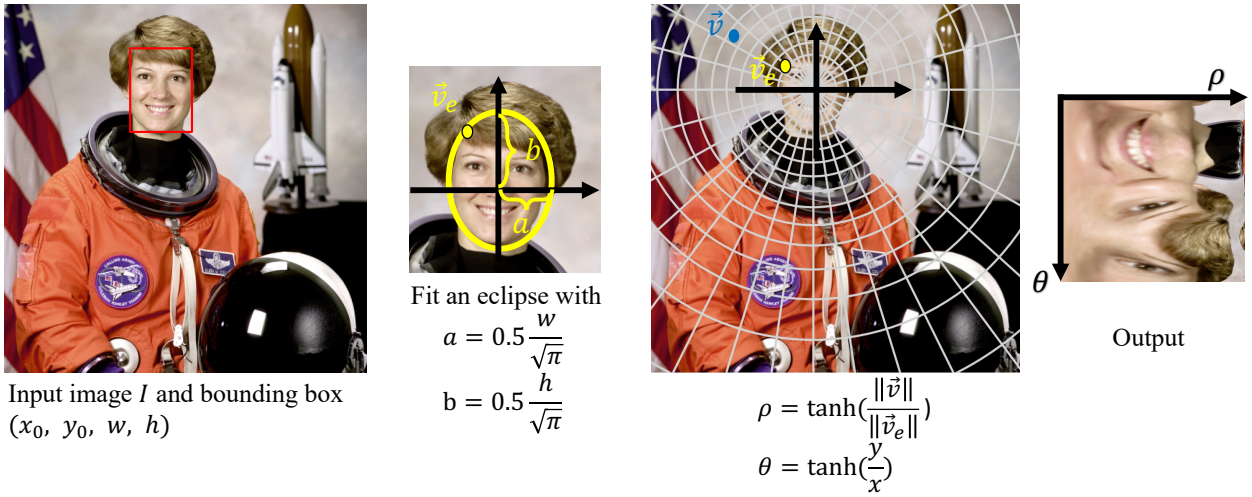


Figure 6: RoI Tanh-polar transform (RT-Transform). (1) Input is an image and the target bounding box. (2) An eclipse e is fitted to the box. (3) The grey patterns depict the Tanh-polar sampling grid. $\vec{v}$ is an arbitrary vector and $\vec{v}_e$ is on the eclipse e in the same direction as $\vec{v}$. (4) Transformed image. Due to the normalisation by $\vec{v}_e$, the boundary of the face is located at $\rho = \tanh(1) = 0.76$ regardless of the face size. All information is preserved and the proportion between face and background is fixed.

also introduce rotation equivariance to CNN models because of the polar-based representations.

The RT-Transform is illustrated in Figure 6. Let $\vec{v} = (x, y)$ represent the Cartesian coordinate of a point in the original image, and let w and h represent the width and height of the bounding box, respectively. We select the centre of the bounding box as the polar origin. We first fit an ellipse e to the target bounding box to the bounding rectangle, described by

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \tag{1}$$

where $a = 0.5\frac{w}{\sqrt{\pi}}$, $b = 0.5\frac{h}{\sqrt{\pi}}$, and w and h are the width and height of the bounding box. We then define the Tanh-polar coordinate system by an injective map f:

$$f(\vec{v}) := (\tanh(\frac{y}{x}), \tanh(\frac{\|\vec{v}\|}{\|\vec{v}_e\|})) \tag{2}$$

where $\vec{v}_e = (x_e, y_e)$ is the vector on the broader of the target face ellipse e and $\vec{v}_e$ and $\vec{v}$ are *parallel*. A new representation is constructed by resampling the input image over a rectangular grid in the Tanh-polar coordinate system. Following typical transformer networks [78, 79], we use bilinear interpolation for points that do not coincide with the pixel locations in the input image. We name this as RoI Tanh-polar transform (RT-transform). It can be observed that: 1) compared to representations obtained by crop-and-resize, all information in the input image is preserved in the new representation; 2) the normalisation with $\vec{v}_e$ ensures that the target face always occupies around 76% (since $\tanh(1) = 0.76$). It is worth noting that the proposed RT-Transform is invertible and differentiable. Therefore, not only can the input RGB images be transformed but also the intermediate feature maps in CNNs.

Rotation Equivariance. To handle rotation of the target face, previous face parsing works [35, 16] rely on transforming facial landmarks to canonical locations correct the rotation of the target face. We show that using the Tanh-polar representation can eliminate such pre-processing step.

The Tanh-polar coordinate system by definition is a canonical coordinate system [80] for the *rotation group* $\mathcal{SO}(2)$ with angle $\theta \in [-\pi, \pi]$. This is because for the rotation transformation $T_\theta\vec{v} = (x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta)$, the Tanh-polar coordinate system satisfies [79]

$$f(T_\theta\vec{v}) = f(\vec{v}) + \mathbf{e}_\theta, \tag{3}$$

where $\mathbf{e}_\theta = (\theta, 0)$. Thus, a rotation transformation T_θ appears as a translation by $(\theta, 0)$ under the Tanh-polar coordinate system f. As a result, the planar convolution that is self-consistent with respect to translation [79] in f is now equivalent to $\mathcal{SO}(2)$ group-convolution [32, 72] in the Cartesian space.

Scale Invariance. Equation 2 shows that the warped image would always have a fixed ratio between the face area and the background area regardless of the face's original size in the input images as long as the provided bounding box has the correct size. This means a model trained in the ROI Tanh-polar space would per-

form equally well on small faces as well as on large faces in the input image.

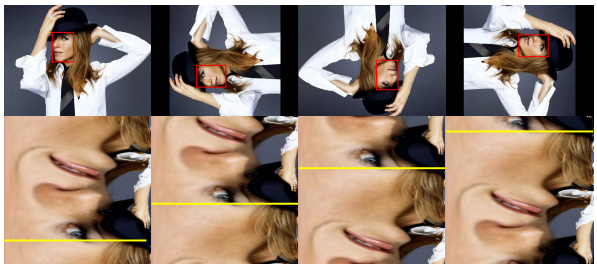


Figure 7: Rotation equivariance. Rotation is reduced to translation in the Tanh-polar coordinate system.

4.2. Hybrid Residual Representation Learning Block

Using the Tanh-polar representation as input to CNNs, rotation equivariance is achieved but translation equivariance may be lost. To overcome this, we propose Hybrid Residual Representation Learning Block, dubbed as *HybridBlock*, a CNN building block similar to the Residual Block [48].

The incentive of designing HybridBlock is to have two branches of convolutions learn representations that are complementary. One branch (Tanh-polar branch) learns the rotation equivariant representations while the other branch (Tanh-Cartesian branch) learns translation equivariant representations. The detailed components of a HybridBlock is depicted on the right in Figure 1.

We define Tanh-Cartesian coordinate system by

$$f_{TC}(\vec{v}) := (\tanh(\frac{x}{||\vec{v}_e||}), \tanh(\frac{y}{||\vec{v}_e||})). \quad (4)$$

The input to HybridBlock is a Tanh-polar representation $\mathbf{X}_{TP}$ of shape (h, w, c). The residual path uses a stack of 1×1, 3×3 and 1×1 convolutions following the *bottleneck* design [48]. The first 1×1 conv layer is used to reduce the channel dimension and its output feature maps are transformed to the Tanh-Cartesian space. In each coordinate system a 3×3 conv layer is used to compute feature maps, which are then concatenated in the Tanh-polar space. The last 1×1 conv layer restores the channel dimension so the residual representation can be added to the input $\mathbf{X}_{TP}$.

Direct Transformation from Tanh-polar to Tanh-Cartesian. To obtain Tanh-Cartesian representations, a naive approach is to inverse-transform from f to Cartesian and then resample with Equ. 4. However, iterated resampling will degrade image quality and amplify the influence of interpolation artefacts. To circumvent this issue, we find the correspondence between the sampling grids in both coordinates and directly resample the Tanh-polar representation.

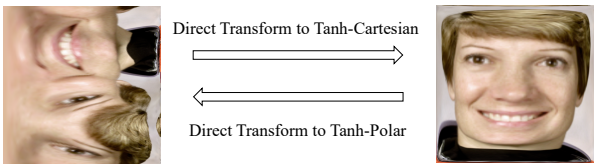


Figure 8: Direct transform between Tanh-polar and Tanh-Cartesian coordinates. We do not sample on the original image but directly transform between two coordinates. Translation equivariance is recovered in Tanh-Cartesian coordinates.

Hybrid Receptive Field. The receptive field (RF) is the region in the input space that a particular neuron is looking at. The two 3×3 convolution layers in different coordinate systems have RFs of different shapes. The Tanh-polar one has the arc-shaped RF while the other has the rectangle-shaped RF.

50-layer HybridNet. We follow the design of ResNets [48] and stack HybridBlocks to create a new backbone network HybridNet50. Thanks to the grouped conv1x1 and the conv3x3 with halved channels, the overall number of parameters are less than the ResNet50 backbone (23.5 M versus 17.8M).

4.3. RTNet: the Overall Framework

With the previously introduced components, we now describe the overall framework of RoI Tanh-polar transformer network (RTNet) for face parsing in the wild. As in Figure 1, RTNet is based on the FCN framework [15]. An input image I of arbitrary resolution with the target bounding box is transformed to I_{tp} in the Tanh-polar space. By default, the size of I_{tp} is set to be 512×512. Next, HybridNet-50 is used to extract features from I_{tp}, followed by a naive FCN decoder head to predict the segmentation mask in the Tanh-polar space. Finally, the segmentation mask is inverse-transformed to Cartesian as the final output that has the same resolution with the input image I.

FCN Decoder. We use the FCN-8s [15] decoder to predict the masks. More advanced decoders like ASPP [54] require dedicated hyperparameter-tuning that may largely affect the performance. The adopted decoder consists of two conv3x3 layers and a bilinear upsampling layer to map the feature maps to pixel-wise prediction logits.

Loss function. We use Cross-Entropy loss $loss_{CE}$ and Dice loss $loss_{dice}$ [81]. Two losses are jointly optimised with a factor of λ and the overall loss l is

$$l = \lambda loss_{CE} + (1 - \lambda) loss_{dice}. \quad (5)$$

The losses are computed on the Tanh-polar coordinates since the outputs are of the same size and the computation can be batched and accelerated.

Mixed padding. Zero-padding is used in most CNNs to keep feature map size.This is not for the Tanh-polar representation, as it is periodic about the angular axis. We use wrap-around padding on the vertical dimension and replication padding on the horizontal dimension.

Bounding box augmentation. To improve robustness of RTNet, we augment the input bounding box during training time by adding a random shift and a random scaling. The augmentation can also be conducted during test time to the input image multiple times, and the inverse-transformed prediction masks can be averaged for smoother results.

5. Experiments

5.1. Experiment Setup

5.1.1. Baseline Methods

We adopt the following criteria to select baseline methods. First, the model should be able to parse inner facial components as well as hair. Second, it is open-sourced and we are able to re-produce the reported performance by re-training the model from scratch. Third, the number of hyper-parameters has to be relatively small so that the performance does not rely on advanced training techniques. This allows that the same training setup can be applied and training can finish with reasonable computing resources and time. The selected models include the classic models like FCN [15], as well as the advanced ones, such as Deeplabv3+ [54] and SPNet [57]. We collected their open-sourced codes and built an unified benchmarking codebase such that the same training and evaluation procedures are ensured.

5.1.2. Training and Evaluation

Data. Each model is trained with four datasets, *i.e.* Helen [26], CelebAMask-HQ [28], LaPa [29] and ours, among which Helen, LaPa and ours have the same set of labelling classes so the models trained on them can be evaluated directly. CelebAMask-HQ has more labelling classes and we assigned those additional classes to the background during training and evaluation. The target bounding box in each image is generated from the groundtruth mask to eliminate the bias from face detection.

Evaluation. We adopt two popular metrics, intersection over union (IoU), and F1 score (F1) for evaluation. We report the metrics for all foreground classes and their mean. The predicted masks are evaluated on the original image scale. For methods with crop-and-resize pre-processing, we resize the predicted masks to the size of the cropped image and then zero-pad it to match the original image resolution. For our RTNet, we apply inverse RT-Transform to the predicted masks. We did not employ other common evaluation techniques such as multi-scale, flipping or multi-cropping.

5.1.3. Implementation Details

We use PyTorch [82] to implement all baselines and our methods. The backbone networks are pre-trained on ImageNet [83]. We use Stochastic Gradient Descent (SGD) to optimise the losses. The initial learning rates are set to 0.01 and the poly learning rate annealing schedule is adopted with *power* = 0.9. All methods are trained for 50 epochs and early stopping is adopted if the mean IoU on the validation set stops growing for 15 epochs. For all methods, we apply random scaling in the range of $[0.5, 2.0]$, random horizontal flip and random brightness as data augmentation methods during training. For our methods, we transform the entire image to 512×512 with our RT-Transform. Batch size is set to 4 in all experiments. All training and evaluation are conducted on two RTX 2080 Ti GPUs.

5.2. Results on iBugMask

We compare our model with different baselines that use *align* as the input pre-processing method. The alignment templates are adopted from open-sourced ArcFace [84] library[1]. These templates have been shown successful in face recognition tasks. We report results for all facial parts. Eyebrows, eyes, lips and inner mouth are merged to Inner Parts.

Table 3 shows the benchmarking results. Our first observation is that iBugMask is challenging and cannot be readily solved. Compared with existing benchmarks, the models' performance on iBugMask is not saturated. For example, the mean F1 score on Helen has reached over 90% but our best results on iBugMask are around 86%. We believe iBugMask can serve as a challenging benchmark for face parsing in the wild.

Our second observation is that when using face alignment for pre-processing, the baseline models perform comparably on inner parts. However, the performance on hair is largely degraded because the templates cannot handle different hairstyles. In contrast, RT-transform allows our model to capture complete hair and face regions without being cut out.

[1] `https://git.io/JOrvm`

Lastly, without landmarks and alignment, our RTNet perform better than other methods in eyes, eyebrows, skin and hair regions, and comparably in nose, lips and mouth. When compared to the baseline FCN, we observe a large improvement in eyebrows and eyes. This could be attributed to the fact that the hybrid representation can better capture elongated regions.

5.3. Qualitative Results

Figure 9 visualises the prediction results of different methods, and our RTNet can better capture the varying hair styles, profile poses, occlusions, *etc*., which again verifies the superior performance of our method under in-the-wild scenarios.

5.4. Ablation Study

We conduct extensive ablation studies to better understand the working mechanisms in RTNet. All variants in this section were trained on pose-augmented images and evaluated on iBugMask.

5.4.1. Effectiveness of RT-Transform

We compare the performance of 4 pre-processing techniques:
1) *Resize*: resizing all the input images to 512×512 with zero-padding to preserve the aspect ratios;
2) *Crop-and-resize*: cropping the face out with 40% margin and then resizing the cropped face to 512×512;
3) *Align*: we use 5 landmarks returned by RetinaFace [25] to align the target faces using the open-source library and warp to 512×512;
4) *RT-Transform*: warping the whole image to a representation of size 512×512 in the Tanh-polar space with the proposed RT-Transform.

Table 4 shows the F1 scores of different pre-processing methods on iBugMask. It can be seen that resizing the input images to the same size gives the lowest accuracy. This is in line with our expectation, as faces vary largely in size and uniformly resizing them will cause confusions. As for the crop-and-resize approach, only a small amount of improvement is observed, especially for the Hair class. This is potentially because the pre-defined cropping margin cannot guarantee a full coverage of the hair region, which will cause accuracy loss on such regions. The alignment method gives competitive results on inner parts and facial skin. However, the performance on hair has degraded by a large amount because the warping template cannot account for different hair styles. In contrast, our RT-Transform achieves the best performance on all three categories, and this can be attributed to the proposed Tanh-Polar transform that can emphasise the facial region while preserving all the contextual background information.

5.4.2. Design of HybridBlock

We also conduct ablation studies to verify the design of HybridBlocks. We started with the Resnet50 backbone and gradually replace the residual blocks with HybridBlocks at different places of the network. In particular, the backbone comprises a stem layer and 4 stages of residual blocks, and we followed common practice [85] to replace blocks in the last three stages with the proposed HybridBlock which has fewer parameters. Results of different replacing stages are reported in Table 5, and we can observe that HybridBlocks can always introduce performance improvement with fewer parameters. Besides, the highest accuracy is achieved when using HybridBlocks in all the three stages, which demonstrate the effectiveness and the generality of the proposed HybirdBlocks.

5.4.3. Bounding box augmentation and mix-padding

Table 6 quantifies the performance gains of the bounding box augmentation and mix-padding described in Sec. 4.3. The box augmentation can make the model more robust to the bounding box noise. And mix-padding is necessary as the Tanh-polar representation is periodic about the angular axis.

5.5. Effectiveness of Pose-augmented Training Set

To show the effectiveness of the pose augmentation, we train 6 on 4 different training sets. For simplicity and faster training, we use *crop-and-resize* with 40% margin to pre-process the input image to obtain a 512×512 facial image for the baseline models. We make the following observations:

Training on pose-augmented images improves all methods. We can also observe that training on pose-augmented images improved all methods, especially on the inner facial parts. This can be reasonably be attributed to that pose-augmented images is constructed in a way that the numbers of faces are balanced across different poses and that in-the-wild information is also preserved. In contrast, CelebAMask-HQ is a synthesised dataset with limited variations in pose and background. Although CelebAMask-HQ contains a larger number of facial images, models trained on this dataset achieve less competitive performance than trained on others.

	L-brow	R-brow	L-eye	R-eye	Nose	U-lip	I-mouth	L-lip	Inner Parts	Skin	Hair
Deeplabv3 [51]	70.6	69.5	78.6	78.4	**90.2**	**75.7**	82.2	**78.6**	**85.8**	91.0	58.1
Deeplabv3+ [54]	71.8	72.1	77.8	78.9	90.0	75.2	82.4	78.4	**85.8**	91.1	57.7
PSPNet [56]	70.2	70.0	78.6	79.1	89.5	75.3	82.2	78.1	85.3	90.7	58.2
SPNet [57]	73.2	71.9	77.9	78.0	90.0	**75.7**	81.7	78.5	85.5	90.1	57.9
FCN (baseline)	71.0	70.6	78.2	78.5	89.6	75.6	**82.7**	78.0	85.4	90.9	57.7
RTNet (ours)	**76.0**	**73.0**	**79.6**	**79.9**	89.3	75.5	82.5	77.6	**85.8**	**91.8**	**81.8**

Table 3: Results on iBugMask. F1 scores are reported in percentage. Eyebrows, eyes, lips and inner mouth are merged to Inner Parts.

Figure 9: Qualitative results on iBugMask of four methods: FCN [15], Deeplabv3+ [54], SPNet [57] and ours. Our method can handle large variations in head pose, hair styles, expressions and occlusions.

Input Pre-processing	Inner Parts	Facial Skin	Hair
Resize	63.2	73.1	63.9
Crop-and-resize	79.0	81.3	71.3
Align	85.2	90.7	56.7
RT-Transform	**85.8**	**91.8**	**81.8**

Table 4: Performance of our model with different pre-processing methods. F1 scores are reported in percentage.

BBox augment	Mix-padding	Inner Parts	Facial Skin	Hair
N	Y	83.8	91.4	80.7
Y	N	85.0	91.4	80.8
Y	Y	**85.8**	**91.8**	**81.8**

Table 6: Ablation study. Random bounding box augmentation during training time, and mix-padding all contribute to improve the F1 scores (in percent).

Hybrid Stages	Inner Parts	Facial Skin	Hair	# Params (in millions)
-	85.6	90.7	81.2	31.4
Stage_2	85.8	90.9	81.4	31.2
Stage_3	85.6	91.6	81.2	29.9
Stage_4	85.8	91.4	81.4	29.0
Stage_All	**85.8**	**91.8**	**81.8**	**27.3**

Table 5: The effectiveness of HybridBlock in different stages of the backbone. F1 scores in percentage are reported.

RTNet consistently outperforms other methods. The results on iBugMask show that our approach outperforms all other methods. Moreover, when trained with our proposed pose-augmented images, RTNet significantly outperform all baselines, especially on the hair class, which indicates that: 1) Compared with other benchmarking datasets, pose-augmented images can better benefit the in-the-wild learning of segmentation models, despite that most of its facial images were generated through the pose augmentation technique, and 2) Different from the baselines, our RTNet can learn from the in-the-wild data more effectively and thus can demonstrate more robust performance on the unconstrained iBugMask dataset.

5.6. Comparison with the State-of-the-arts

In additional to the self-collected iBugMask dataset, we also train and evaluate our method on various face parsing benchmarks.

Results on Helen. Table 10 compares our RTNet with other state-of-the-art methods on Helen [26, 16]. Our model achieves slightly better performance on Facial Skin while significantly outperforms others on the Inner Parts and Hair classes.

Results on LFW-PL. Table 8 compares our RTNet with other state-of-the-art methods on LFW-PL [27]. Our model achieves comparable results in Inner Parts, and outperforms other methods in Facial Skin and Hair.

Results on LaPa. Table 11 compares results from different methods on LaPa [29] dataset and we can easily spot that our RTNet consistently exhibits the highest performance on all three categories, which further demonstrates the generality of our method.

5.7. Model Efficiency

We also examine the running efficiency of different models by evaluating 1) the number of model parameters, 2) Floating Point Operations per Second (FLOPS) and 3) the actual inference time per sample. All models are measured on the same machine with a GTX1080Ti GPU with an $(512, 512, 3)$ input size. To ensure a fair comparison, we repeat the evaluation process 100 runs for each method and report the average. As shown in Table 9, our model has the smallest model size and also operates with the fewest FLOPS when compared with three representative face parsing approaches. Although the inference time of our models are slightly longer than that of FCN and SPNet due to the direct sampling between two coordinates, we believe the time difference (16 ms) is tolerable as our method has shown improved performance over others.

6. Conclusion

In this paper, we have approached in-the-wild face parsing from three aspects: data, representation and model. We have proposed a novel benchmark, iBugMask, for training and evaluating face parsing methods in unconstrained environment. We have created a large-scale training set using pose augmentation and shown its effectiveness. We have solved the dilemma of face cropping and eliminated the need for facial landmarks by proposing a new Tanh-polar representation obtained by the proposed RoI Tanh-polar transform. Equivareriance with respect to rotations has also been achieved with the new representation. HybridBlock is introduced to extract features in both Tanh-polar and Tanh-Cartesian coordinates. We have achieved the state-of-the-art performance on iBugMask as well as other existing face parsing benchmarks. We expect our RT-Transform to be

Training Set	Region	Deeplabv3 [51]	Deeplabv3+ [54]	FCN [15]	PSPNet [56]	SPNet [57]	RTNet (ours)
Helen [26]	Inner Parts	70.3	71.7	68.1	70.4	69.8	**74.3**
	Hair	72.8	72.8	71.3	72.1	71.5	**78.7**
	Skin	90.7	90.5	88.7	90.2	89.6	**91.9**
CelebAMask-HQ [28]	Inner Parts	73.6	73.7	73.9	74.0	74.4	**76.1**
	Hair	74.3	72.9	74.0	73.1	74.6	**77.8**
	Skin	88.8	88.6	89.1	88.6	89.7	**91.8**
LaPa [29]	Inner Parts	74.2	74.1	75.1	74.4	75.9	**77.6**
	Hair	75.8	75.4	75.6	75.9	75.8	**79.8**
	Skin	89.8	89.3	90.1	89.7	89.9	**92.2**
iBugMask-train (ours)	Inner Parts	77.9	78.7	76.8	78.3	78.9	**85.8**
	Hair	72.4	72.7	64.6	72.0	72.9	**81.8**
	Skin	91.7	91.7	91.1	91.7	91.5	**91.8**

Table 7: Effectiveness of the pose-augmented training set iBugMask-train. Baseline models use *crop-and-resize* for pre-processing. The mean F1 scores are reported (in percentage).

Methods	Skin	Hair	Background	accuracy
Liu *et al.*[18]	93.93%	80.70%	97.10%	95.12%
Long *et al.*[15]	92.91%	82.69%	96.32%	94.13%
Chen *et al.*[51]	92.54%	80.14%	95.65%	93.44%
Chen *et al.*[86]	91.17%	78.85%	94.95%	92.49%
Zhou *et al.*[87]	94.10%	85.16%	96.46%	95.28%
Liu *et al.*[24]	**97.55**%	83.43%	94.37%	95.46%
Lin *et al.*[16]	95.77%	88.31%	98.26%	96.71%
RTNet	95.85%	**90.08**%	**98.55**%	**97.11**%

Table 8: Comparison with state-of-the-art methods on LFW-PL. F1 scores for each region and the overall pixel accuracy are reported.

Measurement	FCN	Deeplabv3+	SPNet	Ours
Params (M)	32.95	39.64	39.13	**27.29**
FLOPS (GMac)	26.55	31.39	29.60	**21.99**
Inference Time (ms)	**54**	74	63	70

Table 9: Efficiency comparison between four methods: FCN, Deeplabv3+, SPNet and ours. Input images are of size $(512, 512, 3)$. Models are profiled on the same machine and values are the mean of 100 runs. Lower values indicate better efficiency. Our model is more efficient in the number of parameters and FLOPS. M stands for Million, GMac for Giga Multiply–accumulate operations, ms for milliseconds.

applicable to other face analysis tasks, where the heuristic pre-processing steps, such as cropping with bounding boxes and rotation correction with landmarks, are unavoidable.

Acknowledgements

All datasets used in the experiments were obtained by, and all training, testing, and ablation studies have been conducted at, Department of Computing, Imperial College London, UK.

References

[1] C. Chen, A. Ross, Matching thermal to visible face images using a semantic-guided generative adversarial network, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–8.

[2] X. Cheng, J. Lu, B. Yuan, J. Zhou, Face segmentor-enhanced deep feature learning for face recognition, IEEE Transactions on Biometrics, Behavior, and Identity Science 1 (4) (2019) 223–237.

[3] X. Ou, S. Liu, X. Cao, H. Ling, Beauty emakeup: A deep makeup transfer system, in: Proceedings of the ACM International Conference on Multimedia, 2016, pp. 701–702.

[4] Y. Nirkin, I. Masi, A. Tran Tuan, T. Hassner, G. Medioni, On face segmentation, face swapping, and face perception, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, pp. 98–105.

[5] Y. Nirkin, Y. Keller, T. Hassner, Fsgan: Subject agnostic face swapping and reenactment, in: 2019 IEEE International Conference on Computer Vision (ICCV), 2019, pp. 7183–7192.

[6] S. Banerjee, W. J. Scheirer, K. W. Bowyer, P. J. Flynn, On hallucinating context and background pixels from a face mask using multi-scale gans, in: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 289–298.

[7] Y. Li, S. Liu, J. Yang, M.-H. Yang, Generative face completion, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3911–3919.

[8] H. Zhang, B. S. Riggan, S. Hu, N. J. Short, V. M. Patel, Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks, International Journal of Computer Vision 127 (6-7) (2019) 845–862.

[9] Z. Chen, C. Wang, B. Yuan, D. Tao, Puppeteergan: Arbitrary portrait animation with semantic-aware appearance transformation, in: 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[10] P. Zhu, R. Abdal, Y. Qin, P. Wonka, Sean: Image synthesis with semantic region-adaptive normalization, in: 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[11] M. M. Kalayeh, B. Gong, M. Shah, Improving facial attribute prediction using semantic segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6942–6950.

[12] M. M. Kalayeh, M. Shah, On symbiosis of attribute prediction

Methods	Eyes	Brows	Nose	I-mouth	U-lip	L-lip	Mouth	Inner parts	Facial skin	Hair	Foreground mean
Smith *et al.*[26]	78.5	72.2	92.2	71.3	65.1	70.0	85.7	80.4	88.2	-	-
Zhou *et al.*[88]	87.4	81.3	95.0	83.6	75.4	80.9	92.6	87.3	-	-	-
Liu *et al.*[18]	76.8	71.3	90.9	80.8	62.3	69.4	84.1	84.7	91.0	-	-
Liu *et al.*[24]	86.8	77.0	93.0	79.2	74.3	81.7	89.1	88.6	92.1	-	-
Wei *et al.*[23]	84.7	78.6	93.7	-	-	-	91.5	90.2	91.5	-	-
Wei *et al.*[20]	89.0	82.6	95.2	86.7	80.0	86.4	93.6	91.5	91.5	-	
Lin *et al.*[16]	**89.6**	83.1	**95.6**	86.7	79.6	89.8	95.0	92.4	94.5	83.5	88.6
RTNet	89.3	**84.9**	94.9	**89.9**	**94.1**	**90.9**	**95.6**	**92.7**	**96.2**	**90.6**	**91.8**

Table 10: Comparison with state-of-the-art methods on Helen. F1 scores are reported in percentage. **Bold** values are for the best results.

Methods	L-Eye	R-Eye	U-lip	I-mouth	L-lip	Nose	L-Brow	R-Brow	Skin	Hair	Mean
Zhao *et al.* [56]	86.3	86.0	83.6	86.9	84.7	94.8	86.8	86.9	93.5	94.1	88.4
Liu *et al.* [29]	88.1	88.0	84.4	87.6	85.7	95.5	87.7	87.6	97.2	96.3	89.8
Te *et al.* [35]	89.5	90.0	88.1	90.0	89.0	97.1	86.5	87.0	97.3	96.2	91.1
Ours	**91.5**	**90.9**	**88.7**	**90.5**	**90.5**	**96.9**	**90.1**	**89.1**	**97.8**	**96.5**	**92.5**

Table 11: Comparison with state-of-the-art methods on the LaPa benchmark. F1 scores are reported in percentage.

and semantic segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[13] K. Khan, M. Attique, R. U. Khan, I. Syed, T.-S. Chung, A multi-task framework for facial attributes classification through end-to-end face parsing and deep convolutional neural networks, Sensors 20 (2) (2020) 328.

[14] X. Li, D. Yang, Y. Wang, W. Zhang, F. Li, W. Zhang, Tcminet: Face parsing for traditional chinese medicine inspection via a hybrid neural network with context aggregation, IEEE Access 8 (2020) 93069–93082.

[15] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (04) (2017) 640–651.

[16] J. Lin, H. Yang, D. Chen, M. Zeng, F. Wen, L. Yuan, Face parsing with roi tanh-warping, in: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[17] P. Luo, X. Wang, X. Tang, Hierarchical face parsing via deep learning, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2480–2487.

[18] Sifei Liu, J. Yang, Chang Huang, M. Yang, Multi-objective convolutional learning for face labeling, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3451–3459.

[19] I. Masi, J. Mathai, W. AbdAlmageed, Towards Learning Structure via Consensus for Face Segmentation and Parsing, in: 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[20] Z. Wei, S. Liu, Y. Sun, H. Ling, Accurate facial image parsing at real-time speed, IEEE Transactions on Image Processing 28 (2019) 4659–4670.

[21] A. S. Jackson, M. Valstar, G. Tzimiropoulos, A cnn cascade for landmark guided semantic part segmentation, in: Computer Vision – ECCV 2016, Springer, Springer International Publishing, Cham, 2016, pp. 143–155.

[22] T. Guo, Y. Kim, H. Zhang, D. Qian, B. Yoo, J. Xu, D. Zou, J.-J. Han, C. Choi, Residual encoder decoder network and adaptive prior for face parsing, in: AAAI Conference on Artificial Intelligence, 2018.

[23] Z. Wei, Y. Sun, J. Wang, H. Lai, S. Liu, Learning adaptive receptive fields for deep image parsing network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3947–3955.

[24] L. J. Sifei Liu, Jianping Shi, M.-H. Yang, Face parsing via recurrent propagation, in: Proceedings of the British Machine Vision Conference (BMVC), 2017.

[25] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-shot multi-level face localisation in the wild, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5202–5211. doi:10.1109/CVPR42600.2020.00525.

[26] B. M. Smith, L. Zhang, J. Brandt, Z. Lin, J. Yang, Exemplar-based face parsing, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[27] A. Kae, K. Sohn, H. Lee, E. Learned-Miller, Augmenting crfs with boltzmann machine shape priors for image labeling, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2019–2026.

[28] C.-H. Lee, Z. Liu, L. Wu, P. Luo, Maskgan: Towards diverse and interactive facial image manipulation, in: 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[29] Y. Liu, H. Shi, H. Shen, Y. Si, X. Wang, T. Mei, A new dataset and boundary-attention semantic segmentation for face parsing., in: AAAI Conference on Artificial Intelligence, 2020, pp. 11637–11644.

[30] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision (ICCV), 2015.

[31] G. B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007).

[32] T. Cohen, M. Welling, Group equivariant convolutional networks, in: International Conference on Machine Learning, 2016, pp. 2990–2999.

[33] X. Zhu, X. Liu, Z. Lei, S. Z. Li, Face alignment in full pose range: A 3d total solution, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[34] V. Le, J. Brandt, Z. Lin, L. Bourdev, T. S. Huang, Interactive facial feature localization, in: Computer Vision – ECCV 2012, Springer, Springer International Publishing, Cham, 2012, pp. 679–692.

[35] G. Te, Y. Liu, W. Hu, H. Shi, T. Mei, Edge-aware graph representation learning and reasoning for face parsing, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 258–274.

[36] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, International Conference on Learning Representations.

[37] J. Warrell, S. J. D. Prince, Labelfaces: Parsing facial features by multiclass labeling with an epitome prior, in: 2009 IEEE International Conference on Image Processing (ICIP), 2009, pp. 2481–2484.

[38] C. Scheffler, J.-M. Odobez, Joint adaptive colour modelling and skin, hair and clothing segmentation using coherent probabilistic index maps, in: Proceedings of the British Machine Vision Conference (BMVC), 2011.

[39] Y. Yacoob, L. S. Davis, Detection and analysis of hair, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (7) (2006) 1164–1169.

[40] K.-c. Lee, D. Anguelov, B. Sumengen, S. B. Gokturk, Markov random field models for hair and face segmentation, in: 2008 IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, 2008, pp. 1–6.

[41] Y. Zhou, X. Hu, B. Zhang, Interlinked convolutional neural networks for face parsing, in: International Symposium on Neural Networks, Springer, 2015, pp. 222–231.

[42] U. Güçlü, Y. Güçlütürk, M. Madadi, S. Escalera, X. Baró, J. González, R. van Lier, M. A. van Gerven, End-to-end semantic face segmentation with conditional random fields as convolutional, recurrent and adversarial networks, arXiv preprint arXiv:1703.03305.

[43] B. Luo, J. Shen, S. Cheng, Y. Wang, M. Pantic, Shape constrained network for eye segmentation in the wild, in: 2020 IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 1952–1960.

[44] Y. Wang, B. Luo, J. Shen, M. Pantic, Face mask extraction in video sequence, International Journal of Computer Vision 127 (6-7) (2019) 625–641.

[45] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, Convolutional lstm network: A machine learning approach for precipitation nowcasting, in: Advances in neural information processing systems, 2015, pp. 802–810.

[46] Y. Wang, M. Dong, J. Shen, Y. Wu, S. Cheng, M. Pantic, Dynamic face video segmentation via reinforcement learning, in: 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[47] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, International Conference on Machine Learning.

[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[49] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv preprint arXiv:1511.07122.

[50] F. Yu, V. Koltun, T. Funkhouser, Dilated residual networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 472–480.

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

[52] Y. Wang, M. Dong, J. Shen, Y. Lin, M. Pantic, Dilated convolutions with lateral inhibitions for semantic image segmentation, arXiv preprint arXiv:2006.03708.

[53] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (12) (2017) 2481–2495.

[54] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 833–851.

[55] G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1925–1934.

[56] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[57] Q. Hou, L. Zhang, M.-M. Cheng, J. Feng, Strip Pooling: Rethinking spatial pooling for scene parsing, in: 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[58] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: Criss-cross attention for semantic segmentation, in: 2019 IEEE International Conference on Computer Vision, 2019, pp. 603–612.

[59] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3141–3149.

[60] Z. Zhu, M. Xu, S. Bai, T. Huang, X. Bai, Asymmetric non-local neural networks for semantic segmentation, in: 2019 IEEE International Conference on Computer Vision, 2019, pp. 593–602.

[61] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, L. Fei-Fei, Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation, in: 2019 IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 82–92.

[62] L.-C. Chen, M. D. Collins, Y. Zhu, G. Papandreou, B. Zoph, F. Schroff, H. Adam, J. Shlens, Searching for efficient multi-scale architectures for dense image prediction, in: Neural Information Processing Systems, 2018.

[63] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.

[64] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, B. Xiao, Deep high-resolution representation learning for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[65] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Bisenet: Bilateral segmentation network for real-time semantic segmentation, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 334–349.

[66] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: A survey, arXiv preprint arXiv:2001.05566.

[67] K. Hotta, T. Kurita, T. Mishima, Scale invariant face detection method using higher-order local autocorrelation features ex-

tracted from log-polar image, in: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 70–75.

[68] F. Jurie, A new log-polar mapping for space variant imaging. Application to face detection and tracking, Pattern Recognition 32 (5) (1999) 865–875.

[69] Y. Bao, B. Qi, F. Gu, Facial recognition using partial Log-Polar transformation, in: 2011 IEEE/SICE International Symposium on System Integration (SII), 2011, pp. 74–77.

[70] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (10) (2005) 1615–1630.

[71] E. Tola, V. Lepetit, P. Fua, A fast local descriptor for dense matching, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.

[72] X. Z. Carlos Esteves, Christine Allen-Blanchette, K. Daniilidis, Polar transformer networks, International Conference on Learning Representations.

[73] P. Ebel, E. Trulls, K. M. Yi, P. Fua, A. Mishchuk, Beyond cartesian representations for local descriptors, in: 2019 IEEE International Conference on Computer Vision (ICCV), 2019, pp. 253–262.

[74] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 faces in-the-wild challenge: The first facial landmark localization challenge, in: 2013 IEEE International Conference on Computer Vision Workshops, 2013, pp. 397–403.

[75] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, S. Zafeiriou, The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking, International Journal of Computer Vision (2018) 1–26.

[76] A. Bulat, G. Tzimiropoulos, How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks), in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[77] K. Wang, X. Peng, J. Yang, S. Lu, Y. Qiao, Suppressing uncertainties for large-scale facial expression recognition, in: 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[78] M. Jaderberg, K. Simonyan, A. Zisserman, k. kavukcuoglu, Spatial transformer networks, in: Neural Information Processing Systems, 2015.

[79] K. S. Tai, P. Bailis, G. Valiant, Equivariant Transformer Networks, in: International Conference on Machine Learning, 2019.

[80] J. Segman, J. Rubinstein, Y. Y. Zeevi, The canonical coordinates method for pattern deformation: theoretical and computational considerations, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (12) (1992) 1171–1183.

[81] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 International Conference on 3D vision (3DV), IEEE, 2016, pp. 565–571.

[82] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc., 2019, pp. 8026–8037.

[83] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.

[84] J. Deng, J. Guo, X. Niannan, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: CVPR, 2019.

[85] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[86] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, A. L. Yuille, Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4545–4554.

[87] L. Zhou, Z. Liu, X. He, Face parsing via a fully-convolutional continuous crf neural network, arXiv preprint arXiv:1708.03736.

[88] H. Zhao, X. Qi, X. Shen, J. Shi, J. Jia, Icnet for real-time semantic segmentation on high-resolution images, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 418–434.