Towards Learning Structure via Consensus for Face Segmentation and Parsing — Supplementary material —

Iacopo Masi Joe Mathai Wael AbdAlmageed
USC Information Sciences Institute, Marina del Rey, CA, USA
{masi,jmathai,wamageed}@isi.edu

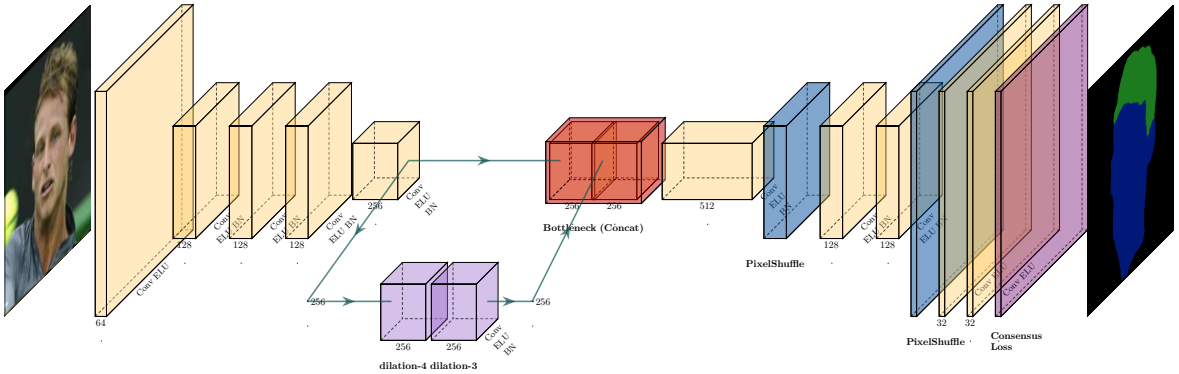

Figure 1: **Network structure at a glance.** Encoder-decoder used for face segmentation and supervised through structure via consensus.

1. Network Architecture

Architecture Details. Further details of our network architecture are provided in Table 1. Our network is similar to the encoder-decoder framework U-Net [8], but it has some modifications explained below. The input resolution is 128×128, though our model can work also with other resolutions since it is fully convolutional. The resolution is decreased only with striding without any pooling layer. The basic block of the network consists of `ReflectionPad2`, `Convolution`, `ELU` [2] and `Batch Normalization`. The first encoder downsamples the input up to 32×32 to preserve some spatial information, with a depth of 256 (layer id 19 referring to Table 1); then other convolutional layers with dilation set to 4 and 3 in a sub-encoder refine the feature maps to capture a more global scale (layer id 27 referring to Table 1). In Table 1, if not specified, convolution is computed with dilation equal to 1. The final feature maps are concatenated together for a final bottleneck feature map with dimensionality 512×32×32. The convolutional filters are initialized with the method described in [3]. The decoder part takes the bottleneck feature maps as input and upscales it back to the input dimension. We used efficient sub-pixel convolution with ratio equal to 2 applied two times in the decoder to do this upscaling, since sub-pixel convolution has been shown to work well in super-resolution applications. We used Pytorch [7] to develop the network and sub-pixel convolution has been implemented via `PixelShuffling`.[1] The entire encoder-decoder has 4,524,323 parameters. The network is supervised either with 2D softmax normalization and cross-entropy or by using our novel "structure via consensus" method. The final network structure is displayed at a glance in Fig. 1 using [4].

[1] pytorch.org/docs/stable/nn.html#torch.nn.PixelShuffle

ID	Layer (type)	Output Shape ($B\times C\times H\times W$)	Param. Size
		Encoder $\downarrow$	
1	ReflectionPad2d	[64, 3, 130, 130]	—
2	Conv2d	[64, 64, 128, 128]	1,792
3	ELU	[64, 64, 128, 128]	—
4	ReflectionPad2d	[64, 64, 130, 130]	—
5	Conv2d	[64, 128, 64, 64]	73,856
6	ELU	[64, 128, 64, 64]	—
7	BatchNorm2d	[64, 128, 64, 64]	256
8	ReflectionPad2d	[64, 128, 66, 66]	—
9	Conv2d	[64, 128, 64, 64]	147,584
10	ELU	[64, 128, 64, 64]	—
11	BatchNorm2d-	[64, 128, 64, 64]	256
12	ReflectionPad2d-	[64, 128, 66, 66]	—
13	Conv2d	[64, 128, 64, 64]	147,584
14	ELU	[64, 128, 64, 64]	—
15	BatchNorm2d-	[64, 128, 64, 64]	256
16	ReflectionPad2d	[64, 128, 66, 66]	—
17	Conv2d	[64, 256, 32, 32]	295,168
18	ELU	[64, 256, 32, 32]	—
19	BatchNorm2d	[64, 256, 32, 32]	512
		Sub-encoder $\downarrow$	
20	ReflectionPad2d	[64, 256, 40, 40]	—
21	Conv2d (dilation=4)	[64, 256, 32, 32]	590,080
22	ELU	[64, 256, 32, 32]	—
23	BatchNorm2d	[64, 256, 32, 32]	512
24	ReflectionPad2d	[64, 256, 38, 38]	—
25	Conv2d (dilation=3)	[64, 256, 32, 32]	590,080
26	ELU	[64, 256, 32, 32]	—
27	BatchNorm2d	[64, 256, 32, 32]	512
		Concat feature maps 19 and 27	
		Decoder $\uparrow$	
28	ReflectionPad2d	[64, 512, 34, 34]	—
29	Conv2d	[64, 512, 32, 32]	2,359,808
30	ELU	[64, 512, 32, 32]	—
31	BatchNorm2d	[64, 512, 32, 32]	1,024
32	*PixelShuffle* ($\times$2)	[64, 128, 64, 64]	—
33	ReflectionPad2d	[64, 128, 66, 66]	—
34	Conv2d	[64, 128, 64, 64]	147,584
35	ELU	[64, 128, 64, 64]	—
36	BatchNorm2d	[64, 128, 64, 64]	256
37	ReflectionPad2d	[64, 128, 66, 66]	—
38	Conv2d	[64, 128, 64, 64]	147,584
39	ELU	[64, 128, 64, 64]	—
40	BatchNorm2d	[64, 128, 64, 64]	256
41	*PixelShuffle* ($\times$2)	[64, 32, 128, 128]	—
42	ReflectionPad2d	[64, 32, 130, 130]	—
43	Conv2d	[64, 32, 128, 128]	9,248
44	ELU	[64, 32, 128, 128]	—
45	ReflectionPad2d	[64, 32, 130, 130]	—
46	Conv2d	[64, 32, 128, 128]	9,248
47	ELU	[64, 32, 128, 128]	—
48	Conv2d	[64, 3, 128, 128]	867
		Total # params.	4,524,323

Table 1: **Network details.** Network layers, output shapes and learnable parameters.

2. Additional Qualitative Results on COFW

We show supplementary results on the Caltech Occluded Face in the Wild data (COFW) [1] in Fig. 2. The figures augment Fig. 5 in the paper to provide further samples. The figures display the input image and its ground-truth mask; the result obtained by Nirkin *et al*. [6], obtained by aligning the faces as the mentioned in their publicly available code; our baseline with pixel-wise softmax and cross-entropy; our final approach trained with structure via consensus. Fig. 2 show again that even on a larger pool of samples, our method returns less sparse, more continuous occlusion masks for better face segmentation and parsing. As a remark, we get such clean masks, much closer visually to the ground-truth compared to other approaches, yet we do so by *still* performing pixel-wise inference: we do not use any super-pixel approach at test time nor employ any post-processing step such as CRF, morphological operations etc.

Figure 2: **Additional qualitative samples from the COFW.** Input image and its ground-truth mask; results by Nirkin *et al*. [6]; baseline with pixel-wise loss; our result. The faces are masked to remove occlusions according to each method.

3. Additional Qualitative Results on Part Labels

We show some supplementary qualitative results on the Part Labels database [5] in Fig. 3. On average our masks look more continuous and greatly improve the IoU of the hair class. Fig. 3 reports the input image, the ground-truth annotated mask, the baseline model trained with pixel-wise loss and regularization and our method with regularization. The result of each prediction for each class is used for segmenting part of the face showing the segmentation separately for face and hair. In some cases, the predictions of our model are better than the super-pixel labels (e.g. tenth row).

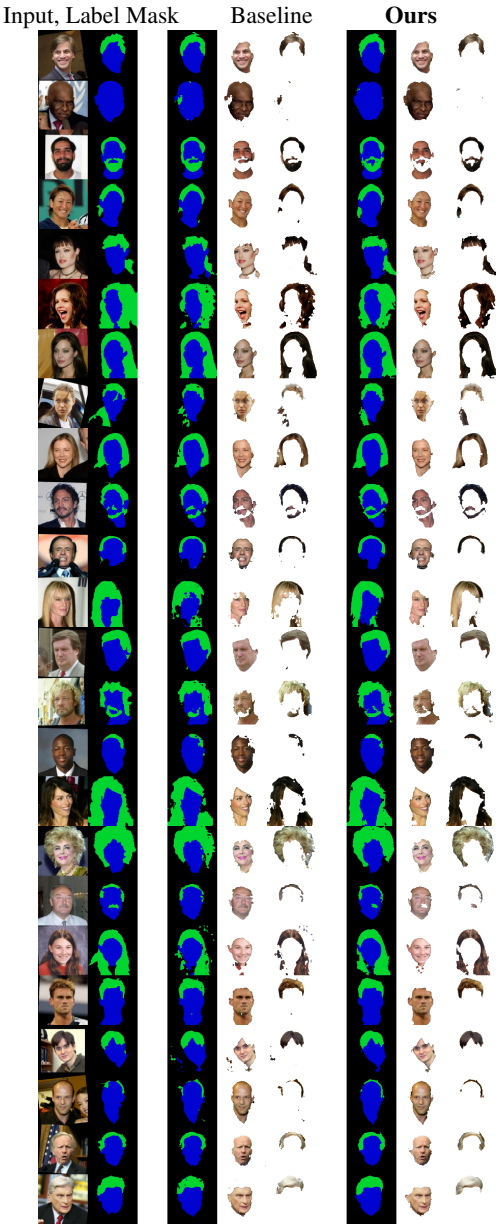

Figure 3: **Additional qualitative samples from PartLabel.** Input image and its ground-truth mask; results by the baseline with pixel-wise loss; our result. The faces are masked to decouple the face from the hair.

References

[1] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *ICCV*, pages 1513–1520. IEEE, 2013.

[2] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In *ICLR*, 2016.

[3] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256, 2010.

[4] Haris Iqbal. Harisiqbal88/plotneuralnet v1.0.0. Dec 2018.

[5] Andrew Kae, Kihyuk Sohn, Honglak Lee, and Erik Learned-Miller. Augmenting CRFs with Boltzmann machine shape priors for image labeling. In *CVPR*, 2013.

[6] Yuval Nirkin, Iacopo Masi, Anh Tran, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *AFGR*, 2018.

[7] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS Workshops*, 2017.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.