FaceOcc: A Diverse, High-quality Face Occlusion Dataset for Human Face Extraction

Xiangnan Yin, Liming Chen

FaceOcc: A Diverse, High-quality Face Occlusion Dataset for Human Face Extraction

Xiangnan YIN and Liming CHEN

Département Mathématique-Informatique
École Centrale de Lyon,
Laboratoire LIRIS,
36 Av. Guy de Collongue, 69134 Écully, France.
Tél: int + 33 4 72 18 60 00, Fax: int + 33 4 78 43 39 62
{yin.xiangnan, liming.chen}@ec-lyon.fr

Abstract Occlusions often occur in face images in the wild, troubling face-related tasks such as landmark detection, 3D reconstruction, and face recognition. It is beneficial to extract face regions from unconstrained face images accurately. However, current face segmentation datasets suffer from small data volumes, few occlusion types, low resolution, and imprecise annotation, limiting the performance of data-driven-based algorithms. This paper proposes a novel face occlusion dataset with manually labeled face occlusions from the CelebA-HQ and the internet. The occlusion types cover sunglasses, spectacles, hands, masks, scarfs, microphones, etc. To the best of our knowledge, it is by far the largest and most comprehensive face occlusion dataset. Combining it with the attribute mask in CelebAMask-HQ, we trained a straightforward face segmentation model but obtained SOTA performance, convincingly demonstrating the effectiveness of the proposed dataset.

Key words face segmentation, face occlusion dataset, face extraction

1 Introduction

Human face extraction is a sub-domain of image segmentation, which aims to locate pure facial regions in face images, excluding backgrounds and occlusions (e.g., hair, hands, glasses, masks, and other facial accessories). It has a wide range of applications in face-related tasks such as face alignment [11,34,32], face image deocclusion [2,18], 3D face reconstruction [4,20,31], and face recognition [35,23,5]. With the widespread use of deep learning, significant progress has been made in image segmentation. Nevertheless, existing face extraction algorithms are still not sufficiently accurate, especially on face images in the wild, where occlusions of arbitrary shapes and textures can present. We argue that the main reason for this problem is the lack of a high-quality occlusion-aware face segmentation dataset. The dataset should satisfy the following characteristics: 1) large volume of data, 2) various occlusion types, 3) accurate annotation, 4) high resolution. Unfortunately, no current dataset fulfills all four of these criteria.

(a) CelebAMask-HQ (b) Part Label (c) HELEN

Figure 1. Problems of several existing datasets (cropped for better visualization).

CelebAMask-HQ [17] comes closest to meeting these requirements among the many face segmentation datasets. It provides manually annotated masks over 30K high-resolution face images in CelebA-HQ [13], covering 18 facial attributes. Nevertheless, this dataset was designed for a GAN-based face image editing task, thus ignoring the various occlusions on the face image (except for hair and glasses), e.g., it wrongly annotates occlusions such as hands as attributes that they obscure. Furthermore, it classifies spectacles and sunglasses in the same category, which is unreasonable as the transparent lenses should not be considered an occlusion. Figure 1a illustrates the above problems. Despite the issues mentioned above, CelebAMask-HQ can be a good starting point for our work, significantly reducing the difficulty and effort of manual annotation. We examined 30K images in CelebA-HQ and annotated all occlusions ignored by CelebAMask-HQ, including hands, spectacles, microphones, scarfs, etc. As CelebA-HQ does not contain face images occluded by masks, we additionally collected such images using Google. Further, we also leverage numerous texture patches to replace the original textures of the occlusion, bringing a greater texture variety.

This paper describes how we construct the occlusion dataset and apply it to the face extraction task. And, we experimentally demonstrate the validity of the proposed method. Our contributions are as follows:

- We propose a novel face occlusion dataset entitled FaceOcc, which is the largest and most comprehensive in this domain to the best of our knowledge.
- The proposed dataset, combined with the annotation of CelebAMask-HQ, allows generating a diverse range of augmented face masking data.
- A straightforward, lightweight face extraction model was trained on the proposed database, achieving SOTA performance without fancy metrics or model structures.

Name	Volume	Quality	Occ Aware	Occ Diversity	Data Aug
COFW	1007	–	–	–	No
Part Labels	2927	Low	Partially	Low	No
HELEN	2330	Low	Partially	Low	No
ELFW	3754	Low	Partially	Low	Yes
CelebAMask-HQ	30 000	High	Partially	High	No
Ours	**30 000+**	**High**	**Yes**	**High**	**Yes**

Table 1. Comparison of publicly available relevant datasets.

2 Related Work

Caltech Occluded Faces in the wild (COFW) [1] consists of 1007 heavily occluded face images with facial landmarks. Although the dataset is designed for facial landmark detection under extreme conditions, given the large number of complex occlusions it contains, it has also been used in much of the literature [11,6,22,23] for training and testing face extraction tasks. The biggest problem with this dataset is that its data volume is too small, with a training set of only 500 images, far from enough to train a deep neural network. Therefore many methods rely on other training data as a supplement, the most commonly used are Part Labels [12] and HELEN [16,30].

Part Labels extends a subset of the Labeled Faces in the Wild (LFW) [10] dataset by manually labeling each superpixel of the face images as hair, face, or background. Its annotation quality is poor since the superpixels provided by LFW are too coarse to capture the fine structure of the face and hair edges, as evidenced by Figure 1b. Based on manually labeled facial landmarks and a hair matting algorithm [19], [30] annotated segmentation masks for 2 330 face images in the HELEN dataset. However, as shown in Figure 1c, the negligence of face occlusions and the unreliable hair matting algorithm resulted in coarse annotations similar to the Part Label. Recently, an updated version of the Part Labels, called Extended Labeled face in the wild (ELFW) [25], has been proposed. It added new data, refined the existing segmentation maps, and synthesized large amounts of occluded face images. Despite its efforts in data volume, annotation quality, and diversity of occlusions, the improvements in all aspects are marginal. All of the above datasets are only in the order of thousands, which is still insufficient for data-driven approaches, and in addition, they are not aware of certain occlusions such as glasses. Table 1 compares different aspects of the publicly available datasets of interest, including data volume, annotation quality, occlusion awareness, occlusion diversity, and whether data augmentation is applied.

Some methods also annotated their private face segmentation datasets (of unknown quality) for training. [28] segmented 5094 images from the FaceWarehouse [3] dataset using GrabCut [27] and manual inspection/correction. [23] proposed a

video-based semi-supervised face segmentation tool and generated 9818 segmented faces from 1275 videos in the IARPA Janus CS2 dataset [15]. Although they collected more training data, the labeling relies on scarce face videos, resulting in their dataset covering only 309 identities. We speculate that the dataset is of low diversity due to the videos' limited subjects and monotonous background region. [22] constructed a large face mask dataset with 598 266 images from CASIA-WebFaces [33], MS-Celeb-1M [8], and VGG Faces [24]. Despite its large volume, the images are of low resolution, and the face masks derived from the pre-trained model of [23] and the fitted 3D face silhouette are problematic.

3 Proposed Method

As mentioned above, CelebAMask-HQ is a large, high-quality face segmentation dataset with manually labeled facial attributes. Ideally, by integrating the attribute masks corresponding to the skin, eyes, nose, lips, and mouth, we obtain the mask of the whole face. However, the dataset ignores various face occlusions (except for hair and glasses) and the difference between spectacles and sunglasses, making this straightforward approach inaccurate. Consequently, we must label all of the occlusions in the dataset by hand to get an accurate face mask. This section describes how we construct the occlusion dataset and apply it to train our baseline model.

3.1 Dataset Construction

First, we aligned all face images in CelebA-HQ and resized them to a resolution of 256×256 via facial landmarks detected by [7]. This process ensures that the labeled face occlusions are of appropriate size and position for subsequent data augmentation. We then inspected the whole dataset and selected images with unlabeled or mislabeled occlusions. Next, we relabel those occlusions.

In many cases, the occlusion is not as evident and definitive as the hand or sunglasses, and different annotators may have different standards of judgment. Here list our annotation rules regarding several special occlusions:

- For spectacles with colorless lenses, only the reflections on the lenses (if present) and the frames are considered occlusions.
- For tinted but transparent lenses, we determine them as skin if their color does not differ much from the skin, and vice versa for occlusion, but the labeling will bypass the eyes.
- The tiny, heavy shadows cast by the eyeglass frames on the face are also labeled.
- Makeups with a clear border with the skin are another source of occlusion.
- The tongue out of the mouth should be treated as occlusion.

Figure 2. Results of annotation and data enhancement. The 1st row is original images; The 2nd row is augmented images with faces masked in blue.

– All occlusions should be marked as a whole, not just the area above the face.

With the occlusion mask and the originally annotated attribute masks, we get the face mask by simple truncated subtraction:

$$M_{face} = \max(\hat{M}_{face} - M_{occ}, 0), \tag{1}$$

where $\hat{M}_{face}$ denotes the integration of initially labeled attribute masks (including the face skin, eyes, nose, lips, and mouth), and M_{occ} denotes the occlusion mask.

Then we trained our first face extraction model on using M_{face} and the images of CelebA-HQ (data augmentation method and model structure in the later sections). We selected hard samples from the FFHQ [14] with this model and added the labeled occlusions to our dataset. As neither CelebA-HQ nor FFHQ contains scarf and mask type occlusions, we collected those occlusions by Google to supplement our dataset.

All annotations were done with an Apple Pencil in the Magic Eraser [1], and we ended up with a collection of over 3000 occlusions.

3.2 Data Augmentation

We apply three data augmentation techniques: 1) spatial and color transformation of the face images, 2) randomly superimposing occlusions onto the face images, 3) randomly changing the texture of the occlusions (for masks and scarfs). Although some of the synthesized images are not realistic enough, experiments demonstrate that the improvement brought by our method is significant. Moreover, instead of synthesizing fixed face images, as many existing methods do [28,23,25], we perform data augmentation on the fly as training goes, bringing more data diversity.

3.3 Baseline Model

Our baseline model uses ResNet-18 [9] as the backbone and follows the U-Net [26] structure. With the augmented data as input, the model predicts a single channel

[1] https://apps.apple.com/us/app/magic-eraser-background-editor/id989920057

probability map representing the face mask. We employ the binary cross-entropy loss combined with "online hard example mining" (OHEM) [29] to guide the training. All the above settings are the most commonly used, without bells and whistles.

4 Experiments

Method	IOU	Global	recall	FPS
Struct. Forest [11]	–	83.9	–	88.6
RPP [32]	72.4	–	–	0.03
SAPM [5]	83.5	88.6	87.1	–
Liu *et al.* [21]	72.9	79.8	89.9	0.29
Saito *et al.* [28]	83.9	88.7	92.7	43.2
Nirkin *et al.* [23]	83.7	88.8	94.1	48.6
Masi *et al.* [22]	87.0	91.3	92.4	**300**
Ours	**93.7**	**98.0**	**98.3**	257

Table 2. COFW segmentation results.

We train our model for 30 epochs with a batch size of 16, which takes about 2 hours on two Nvidia GTX-1080 GPUs. The learning rate is initialized to $1e^{-4}$ and dropped to $1e^{-5}$ after the 20th epoch. The model is optimized using Adam optimizer with a weight decay of 0 and betas of (0.9, 0.999). We evaluate the model on 507 images of the COFW test set following the practice of previous works. Since none of the previous works published their annotations for this dataset, we manually re-labeled its test set.

Table 2 reports our quantitative results compared to others (data taken from the papers of Nirkin *et al.* and Masi *et al.*), where IOU refers to the intersection over union of the predicted mask and the ground truth, Global for the prediction accuracy of all pixels, recall for the percentage of correctly predicted face pixels to ground truth face pixels, and FPS for the number of images processed per second. Our method remarkably outperforms the others in the first three metrics and is slightly slower than Masi *et al.*'s method. However, the input resolution of Masi *et al.* is 128×128, while ours is 256×256.

Experiment results show that the main factors limiting the performance of current face extraction models are the quality and quantity of training data. While using the proposed face occlusion dataset for training, even the most straightforward model can significantly outperform the state-of-the-art, strongly demonstrating the superiority of the our dataset.

5 Conclusion

This paper proposes a novel diverse, high-quality face occlusion dataset entitled FaceOcc, which contains all mislabeled occlusions in CelebAMask-HQ and complements some occlusions and textures from the internet. Together with the facial attribute masks in CelebAMask-HQ, the proposed dataset yields face masks and augmented data for training face extraction models. We validate the superiority of

the proposed dataset by training a straightforward face extraction model far exceeding the state-of-the-art.

References

1. Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE international conference on computer vision*, pages 1513–1520, 2013.
2. Jiancheng Cai, Hu Han, Jiyun Cui, Jie Chen, Li Liu, and S Kevin Zhou. Semi-supervised natural face de-occlusion. *IEEE Transactions on Information Forensics and Security*, 16:1044–1057, 2020.
3. Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013.
4. Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
5. Fei Gao and Jiangjiang Liu. Face recognition using segmentation technology. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 545–548. IEEE, 2019.
6. Golnaz Ghiasi, Charless C Fowlkes, and C Irvine. Using segmentation to predict the absence of occluded parts. In *BMVC*, pages 22–1. Citeseer, 2015.
7. Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
8. Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.
9. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
10. Gary B. Huang, Vidit Jain, and Erik Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007.
11. Xuhui Jia, Heng Yang, Kwok-Ping Chan, and Ioannis Patras. Structured semi-supervised forest for facial landmarks localization with face mask reasoning. In *BMVC*. Citeseer, 2014.
12. Andrew Kae, Kihyuk Sohn, Honglak Lee, and Erik Learned-Miller. Augmenting CRFs with Boltzmann machine shape priors for image labeling. In *CVPR*, 2013.
13. Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
14. Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.
15. Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1931–1939, 2015.
16. Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.
17. Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
18. Yu-Hui Lee and Shang-Hong Lai. Byeglassesgan: Identity preserving eyeglasses removal for face images. In *European Conference on Computer Vision*, pages 243–258. Springer, 2020.

19. Anat Levin, Alex Rav-Acha, and Dani Lischinski. Spectral matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(10):1699–1712, 2008.
20. Chunlu Li, Andreas Morel-Forster, Thomas Vetter, Bernhard Egger, and Adam Kortylewski. To fit or not to fit: Model-based face reconstruction and occlusion segmentation from weak supervision. *arXiv preprint arXiv:2106.09614*, 2021.
21. Sifei Liu, Jimei Yang, Chang Huang, and Ming-Hsuan Yang. Multi-objective convolutional learning for face labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3451–3459, 2015.
22. Iacopo Masi, Joe Mathai, and Wael AbdAlmageed. Towards learning structure via consensus for face segmentation and parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5508–5518, 2020.
23. Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 98–105. IEEE, 2018.
24. Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. *BMVC*, 2015.
25. Rafael Redondo and Jaume Gibert. Extended labeled faces in-the-wild (elfw): Augmenting classes for face segmentation. *arXiv preprint arXiv:2006.13980*, 2020.
26. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
27. Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.
28. Shunsuke Saito, Tianye Li, and Hao Li. Real-time facial segmentation and performance capture from rgb input. In *European conference on computer vision*, pages 244–261. Springer, 2016.
29. Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.
30. Brandon M Smith, Li Zhang, Jonathan Brandt, Zhe Lin, and Jianchao Yang. Exemplar-based face parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3484–3491, 2013.
31. Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard G Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *CVPR*, pages 3935–3944, 2018.
32. Heng Yang, Xuming He, Xuhui Jia, and Ioannis Patras. Robust face alignment under occlusion via regional predictive power estimation. *IEEE Transactions on Image Processing*, 24(8):2393–2403, 2015.
33. Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.
34. Yucheng Zhao, Fan Tang, Weiming Dong, Feiyue Huang, and Xiaopeng Zhang. Joint face alignment and segmentation via deep multi-task learning. *Multimedia Tools and Applications*, 78(10):13131–13148, 2019.
35. Hui Zhi and Sanyang Liu. Face recognition based on genetic algorithm. *Journal of Visual Communication and Image Representation*, 58:495–502, 2019.