Delving into High-Quality Synthetic Face Occlusion Segmentation Datasets

Kenny T. R. Voo Liming Jiang Chen Change Loy
S-Lab, Nanyang Technological University
{kvoo001,liming002,ccloy}@ntu.edu.sg

Abstract

This paper performs comprehensive analysis on datasets for occlusion-aware face segmentation, a task that is crucial for many downstream applications. The collection and annotation of such datasets are time-consuming and labor-intensive. Although some efforts have been made in synthetic data generation, the naturalistic aspect of data remains less explored. In our study, we propose two occlusion generation techniques, Naturalistic Occlusion Generation (NatOcc), for producing high-quality naturalistic synthetic occluded faces; and Random Occlusion Generation (RandOcc), a more general synthetic occluded data generation method (Figure 1). We empirically show the effectiveness and robustness of both methods, even for unseen occlusions. To facilitate model evaluation, we present two high-resolution real-world occluded face datasets with fine-grained annotations, RealOcc and RealOcc-Wild, featuring both careful alignment preprocessing and an in-the-wild setting for robustness test. We further conduct a comprehensive analysis on a newly introduced segmentation benchmark, offering insights for future exploration. Our code and dataset are available at https://github.com/kennyvoo/face-occlusion-generation.

Figure 1. Examples of data generated by our NatOcc and RandOcc methods from CelebAMask-HQ [22]. The first two rows are naturalistic occluded faces generated by NatOcc using color transfer, image harmonization, and super-resolution techniques. The last two rows show occluded faces generated using RandOcc by overlaying random shapes with random textures and transparency.

1. Introduction

This paper explores the problem of occlusion-aware face segmentation, which involves extracting face pixels from an occluded face where occlusions are treated as the background class. Face or occlusion semantic segmentation is extensively used in face-related tasks, such as face recognition [31], face-swapping [25, 27], and facial reconstruction [32, 38]. Occlusions on faces remained less explored in the past. In reality, human faces are likely to be partially occluded by hands or wearable objects such as sunglasses or face masks. These occlusions would often degrade the performance of the face-related tasks. Therefore, occlusion-aware face segmentation is worth studying for many practical applications [18, 25, 38].

Supervised training of occlusion-aware face segmentation requires a large amount of data. However, existing real-world occluded face datasets such as [6, 19, 22] are not suitable for this purpose. They are either low in quantity, low resolution or inaccurately labeled. As data collection and annotation is very time-consuming and labor-intensive, previous studies [25, 28, 29, 38] created their synthetic datasets for training by overlaying commonly wearable objects or hand on existing face datasets with basic augmentation techniques. They often replaced the occluders' texture and color to enrich the occlusions' diversity.

Nevertheless, synthetic data generation often neglects the naturalistic aspects of the data. The color, texture, and edges of the occluders are visually unnatural, as shown in Figure 2. In addition, these synthetic data generation efforts [28, 38] usually only work for specific use cases and cannot scale to other applications. Specific types of oc-

Figure 2. Examples of synthetic occluded face images taken from previous studies. The red box shows images from [25], whereas the blue box shows images from [38], and the green box shows images from [28]. Specific types of occluders (*e.g.*, sunglasses, face masks, and hands) were overlaid at a specific location, and the color and texture of occluders often looked unnatural.

cluders (*e.g.*, sunglasses, and mask) were overlaid at a fixed location, as shown in Figure 2. The question then arises whether more general synthetic data generation could perform as well, if not better than, or at least comparable with these specific methods. Moreover, the commonly used validation set, Caltech Occluded Faces in the Wild (COFW) dataset [6] might not be a suitable benchmark since not all the faces are occluded due to different occlusion definitions. Thus, a new standard benchmark is required.

In this study, we proposed two data generation techniques: Naturalistic Occlusion Generation (NatOcc), an effective method to produce naturalistic synthetic occluded face from CelebAMask-HQ [22] using various techniques such as color transfer, image harmonization, and super-resolution; and Random Occlusion Generation (RandOcc), a more general method by overlaying random shapes with random transparency and textures from the Describable Textures Dataset (DTD) [10]. To evaluate the effectiveness and robustness of our synthetic datasets, they were compared with a real-world occluded face dataset. The real-world occluded face dataset was prepared by manually correcting approximately 4,000 incorrectly labels from CelebAMask-HQ [22]. The dataset is further split into occluded and non-occluded categories. Our synthetic datasets were generated from the non-occluded images. Some examples of the synthetic data generated by NatOcc and RandOcc are shown in Figure 1. Both data generation methods performed at a level or even better than real-world occluded face dataset. Besides, our methods were proved to be effective and robust to handle unseen occlusions and faces in the wild (see Section 4.1). Furthermore, to facilitate model evaluation, we present two high-resolution real-world occluded face datasets collected from Pexels and Unsplash with fine-grained manually annotated masks: RealOcc, which consists of 550 aligned and cropped face images; and RealOcc-Wild, which comprises 270 occluded face images in the wild for robustness test.

Our contributions are summarized as follows. **1)** We propose two synthetic occluded data generation techniques, NatOcc to produce naturalistic synthetic occluded faces, and RandOcc, a general synthetic occluded data generation method. **2)** We provide manually corrected annotation masks and new categories (occluded and non-occluded) for widely used CelebAMask-HQ [22]. **3)** To facilitate model evaluation, we contribute two real-world occluded face datasets with manually annotated masks, RealOcc (aligned and cropped) and RealOcc-Wild (in the wild). **4)** We further benchmark several representative baselines [9, 35, 39] and present a comprehensive analysis, verifying the effectiveness of our method (even for unseen occlusions) and providing insights for future exploration.

2. Related Work

Real-World Occluded Face Datasets. Existing real-world occluded face datasets [6, 17, 22, 36] have the following shortcomings: they are either low in quantity, low resolution, incorrectly labeled, or lack of annotated masks. To our knowledge, there are only a limited number of real-world occluded face datasets [6, 22] that cover various occlusions (*e.g.*, sunglasses, hats, foods, and hands). For instance, Real World Occluded Faces (ROF) [13] consists of 1,686 sunglasses-occluded faces and 678 mask-occluded faces. The datasets that are specifically targeted for face mask occlusion are Interactive Systems Labs (ISL) Unconstrained Face Mask Dataset (ISL-UFMD) [14] with 10,618 face images, and the Face Mask Label Dataset (FMLD) [4] with 63,072 face images. Additionally, Specs on Faces (SoF) [2] consists of 42,592 glasses-occluded face images and Interactive Systems Labs (ISL) Unconstrained Face Hand Interaction Dataset (ISL-UFHD) [14] consists of 10,004 hand-occluded images. Many of these datasets do not have annotated masks.

The only large-scale, high-quality dataset with masks is CelebAMask-HQ [22]. This dataset contains 30,000 face images (1,024×1,024), each containing masks of different parts of the face and accessories such as sunglasses, hats, and earrings. Out of these 30,000 images, approximately 4,000 images are occluded faces. However, the occluded faces suffer incorrect annotations as some objects (*e.g.*, microphones, food, and hands) that overlap the face area are annotated as part of the face rather than the background. The COFW dataset [6] used to have 1,007 annotated masks, but the full data is no longer publicly available. A previous study [15] has provided the 500 annotated masks for the training set of the COFW [6] dataset. There are a few shortcomings with this dataset. First, the quantity and resolution of images in the training set are too low, and the masks of the test partition [6], which were used as a benchmark in different studies, are not available. Lastly, Part Labels [19], which is a subset from LFW [17], contains only 2,927 images with annotated masks. However, these masks

are coarse, not accurately annotated and lacked variety in occlusion. Hence, it is not suitable to be both training and test set.

Synthetic Occluded Face Datasets. Numerous approaches have been proposed to generate synthetic occluded faces. Each method has its own considerations, leading to varying numbers of classes and definitions of occlusions. Many studies such as [25, 28, 29, 38] generated their synthetic datasets by overlaying common occlusions such as sunglasses, masks, hands onto faces. However, most of them are built on a low-resolution dataset, and the synthetic occluded faces do not look natural due to the occluders' color, texture, and edges. Besides, these data generation methods usually overlaid specific types of occluders at specific locations and orientations. Since most of their code or datasets are not made public, it is challenging to reproduce, cross-check, and improve on previous data generation methods. The existing occlusion augmentation techniques such as [12, 40] are not used in previous studies. Instead, they produce unnatural occlusions by covering a region of the image with rectangles.

Extended Labeled Faces in-the-Wild (ELFW) [28] has 3,754 labeled faces (250×250), which is an extension from Part Labels [19]. The original Part Labels [19] has many incorrect annotations and lacks occluded faces. Thus, the contribution of [28] went beyond simply improving the masks of the original dataset by adding new categories (*e.g.*, sunglasses, head-wearables and face masks), but also increased the occluded face images by overlaying objects and hands over the original images. They followed the approach described in [26] to perform occlusion augmentation of hands from both Hand2Face [26] and HandOverFace [33]. Although the synthetic dataset is not provided, their code to perform occlusion augmentation is released. The main issue with this dataset is that the resolution and quantity are low.

The more recent work [38] built their synthetically occluded training dataset based on CelebAMask-HQ [22] where they occluded the dataset with 300 different occluders (*e.g.*, masks, scarfs) and replaced the texture on the original occluders from 800 textures to create more variation. A similar concurrent work to our study is FaceOcc [37], where they fixed the masks of CelebAMask-HQ [22] and produced a synthetic occluded face dataset with it. However, they have different definitions of occlusions. At the time of this study, they had yet to release their code or dataset. Therefore, further comparison was not included.

3. Methodology

3.1. Dataset Preparation

This section shows our rationale in dataset selection and methods to prepare different datasets. First of all, it is essential to set some crucial definitions as each previous study had a different interpretation. Our study has two classes, background and face, the grey or ambiguous area of which is treated as that category (see Table 1) at the time of this study.

Table 1. Definition of classes in our dataset.

Class	Definition
Face	The skin of the head includes eyes, nose, mouth but excluding ears.
Face (Grey Area)	Tattoo, shadow, moustache, and beard overlapped with face, as well as skin of the bald person are considered as part of the face.
Background (Occlusions)	Non-face area and any objects such as sunglasses, shirt, hair, microphone, hands that are physically covering (overlap) the face. Words from magazines or copyright labels fall into this category as well.
Background (Grey Area)	Transparent/Translucent glasses

Figure 3. Examples of the wrongly annotated masks in CelebAMask-HQ [22] as occlusions are treated as part of the face.

Face Dataset. CelebAMask-HQ [22] is a suitable and relevant face dataset since it contains 30,000 high-quality images with refined masks. The initial face masks were obtained by subtracting the skin masks with the masks of sunglasses, neck, and head in the dataset. After that, we manually corrected the wrong annotated masks (see Figure 3) such as hands and microphone using an online annotation platform, Segments.ai[1]. Next, the dataset was split into occluded and non-occluded categories. The occluded category refers to faces that are overlapped or intersected with any objects. Most of the corrupted and cartoon images were excluded from the dataset. A subset of images (716 images) made of 80 identities was extracted from the non-occluded category to act as the test set. The partition is summarized in Table 2.

Table 2. Summary of our partition of CelebAMask-HQ [22].

Category	Quantity
CelebAMask-HQ-WO (Train)	24603
CelebAMask-HQ-WO (Test)	716
CelebAMask-HQ-O	4597
Excluded Images	86

WO - Without occlusion O - Occluded

[1] https://segments.ai/

Figure 4. Comparison of the images from CelebAMask-HQ [22] overlaid with different hands datasets. The hands from 11k Hands [1] have finer details and higher resolution than EgoHands [3].

Occluders Dataset. Unlike other studies [25, 28, 29, 37, 38] that use very specific occluders such as sunglasses and face mask, we extracted 128 common objects across 20 categories (*e.g.*, food, bottles, cellphones, and cups) from the Microsoft Common Objects in Context (COCO) dataset [23] with COCOAPI [23]. The original resolutions of the COCO objects were low, so we performed super-resolution ($\times$4) of the original images with GLEAN [8]. In contrast to other occlusions, hands have similar color to faces, making them harder to detect as occlusion. There are several existing hand datasets [3, 33]. Initially, we used the 200 hands of EgoHands [3] that were sampled by [25]. However, the image resolution of the hands is low and blurry. The edges and details of the hands are not preserved after resizing and overlaid onto CelebAMask-HQ [22] images, as shown in Figure 4. Therefore, we sampled 200 images from another hand dataset, 11k Hands [1], which has a higher resolution of 1,600$\times$1,200. Due to the lack of fine masks, we manually annotated the 200 hands images. The only drawback of this hands dataset is the lack of different postures, but it is good enough for our study. The comparison of the hands overlaid by both datasets is shown in Figure 4, the hands from 11k hands [1] have finer details and higher resolution than EgoHands [3].

Validation Set. To facilitate model evaluation, we introduce a new validation set, RealOcc consisting of 550 high resolution (1,024$\times$1,024) occluded face images from websites such as Pexels and Unsplash, covering different occlusions (*e.g.*, hands, masks, food and sunglasses). The face was aligned and cropped as shown in Figure 5 using the same method used in FFHQ [20]. However, due to occlusions, approximately 265 occluded face images failed to be detected by dlib [21]. With reference to the aligned and cropped images, we manually aligned and cropped the 265 face images. To speed up the labeling process, coarse masks were produced with the models that are trained with our synthetic dataset, followed by correcting the mask using LabelMe [34] and Segments.ai. We introduced another validation dataset for model robustness test, RealOcc-Wild, consisting of 270 occluded faces in the wild (without cropping and alignment) that are mainly made of the images failed to be detected by dlib [21]. Due to our definition of occlusion (*e.g.*, transparent glasses are occlusion, and mustache overlapping faces is not occlusion), the masks in the COFW [6] dataset were manually modified.

Figure 5. Examples of aligned and cropped occluded face images in RealOcc with different methods.

3.2. Naturalistic Occlusion Generation (NatOcc)

Due to the lack of real-world, large-scale occluded face datasets, previous studies [25, 28, 29, 37, 38] have proposed different synthetic occluded face generation techniques. In this work, we propose a NatOcc method to produce high-quality naturalistic occluded face images from CelebAMask-HQ-WO (Train) (see Table 2).

Algorithm 1: Color transfer via Sliced Optimal Transport (SOT) [5] with custom preprocess

$sQuantity \leftarrow source < lowerThresh$;
$tQuantity \leftarrow target == 0$;
$blackRatio \leftarrow sQuantity/tQuantity$;
if $blackRatio > 1$ **then**
  $mean(rgb) \leftarrow$ Mean of non-black $source$ pixels in each channel ;
  Replace $blackRatio - 1$ ratio of black pixels in $source$ with $mean(rgb)$;
end
Clipped pixels value of $source$ to $upperThresh$;
ColorTransferViaSOT($source$,$target$);

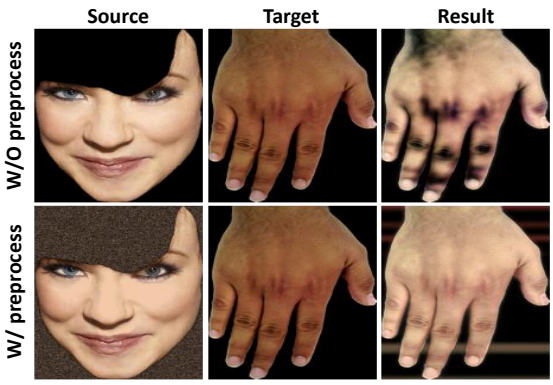

Figure 6. Comparison of the color transfer via Sliced Optimal Transport (SOT) [5] with or without custom preprocess. The preprocess mitigates the issue of black pixels imbalance between the source and the target images, thus improving the quality of the color transfer.

Color Transfer. In reality, most hand occlusions have a similar color to the face. To simulate this scenario, colors from the face were transferred to the hands using color transfer via Sliced Optimal Transport (SOT) [5]. The size of the source (face) and target (hand) images must be the same. If the quantity of the black pixels of the source (face) is larger than those of the target (hand), some area of the hands will appear black due to black pixels imbalance, as shown in Figure 6. On the other hand, if a part of the pixels of the source (face) are too bright, the hands will look unnatural as well. Therefore, preprocess is necessary to address this issue. To resolve this issue, a ratio of black pixels at the source (face) were replaced with the average of each RGB channel of the non-black pixels at the source (face) correspondingly, and the pixels values of the source (face) image were clipped at a certain threshold. The steps of the preprocess before the color transfer is shown in Algorithm 1. The comparison of the color transfer with and without preprocess is shown in Figure 6. This simple preprocess can solve most of the cases. However, future work can look into color transfer via Sliced Partial Optimal Transport (SPOT) [5] that might handle the above issues better. To speed up the long process of this method, we have utilized GPU and multiprocessing. Samples of the synthetic images with or without color transfer can be found in Figure 7.

Augmentation. Affine augmentation, image compression, random brightness and contrast were applied to both faces and occluders using Albumentations [7]. Besides, occluders were randomly resized to 0.5-1 of the face size. Moreover, the edges of occluders were carefully considered to produce a more naturalistic composite image. This was done by applying Gaussian blur to the mask of the occluders before alpha blending, a method of overlaying a foreground image over a background image. After alpha blending, the

Figure 7. The effects of color transfer via SOT [5] on 11k Hands [1] with CelebAMask-HQ-WO (Train) images.

Figure 8. The effects of image harmonization on the occluders using RainNet [24]. The images after image harmonization look more natural.

intersection of the occluders and the faces in the composite image was blurred again. The occluders were randomly positioned around the faces.

Additional Augmentation (Hand Only). Besides color transfer, hands were positioned so that the fingers always point to the face, followed by a small random rotation to simulate hands in real-life scenarios.

Image Harmonization. To further improve the naturalistic aspect of synthesized images, we applied RainNet [24] to harmonize the foreground (occluder) to match the background (face image). COCO [23] occluders look more natural after image harmonization (see Figure 8). However, it was not the same case for the hand occluder as the color of the hand was changed after image harmonization, defeating the purpose of color transfer. Thus, image harmonization was not applied to the hand occluder.

3.3. Random Occlusion Generation (RandOcc)

To overlay hands onto the faces, studies such as [26, 28] find the matching target face images with the same posture as the source face image, followed by overlaying the source image's hands onto the target face image. Occluders such as sunglasses and face masks were overlaid onto eyes and mouth, respectively. Such a synthetic dataset generation method cannot be applied to most applications. Thus, we propose a more general occlusion method, Random Oc-

Figure 9. Examples of CelebAMask-HQ-WO (Train) occluded with random shape and texture from DTD [10]. 30 percent of the occluders are slightly transparent.

clusion Generation (RandOcc), which creates synthetic occluded data with minimal effort. The performance and robustness of this RandOcc dataset will be compared with the real occluded dataset and our NatOcc dataset.

Occlusion Augmentation. RandOcc overlaid random shape with random transparency and texture from Describable Textures Dataset (DTD) [10], which consists of 5,640 images from 47 categories. The same augmentation process (see Section 3.2) was applied in RandOcc.

Transparent/Translucent Object Simulation. To simulate transparent or translucent objects, the alpha mask of alpha blending was randomly set between 0.5 to 0.8. This is applied to approximately 30 percent of the dataset. Examples of the RandOcc dataset can be found in Figure 9.

4. Experiments

4.1. Settings

We provide quantitative and qualitative results on different variants of our dataset. The training was carried out with two CNNs, PSPNet [39] and DeepLabv3+ [9] with pre-trained ResNet-101 backbone [16], and a Vision Transformer, SegFormer [35] with pre-trained MIT-B5 backbone. The trained models were tested on two datasets, COFW (train set) [6], and OFD-W to compare the robustness of each model. The test results of CelebAMask-HQ-WO (Test) are provided in the supplementary material.

Implementation Details. All experiments were carried out using MMSegmentation [11]. Every model was trained on 4 Tesla V100 GPUs, with an input image size of 512×512 and a batch size of 8 for 30k iterations. The optimizer for PSPNet [39] and DeepLabv3+ [9] is SGD with learning rate of 0.01, momentum of 0.9 and decay rate of 0.0005. The optimizer for SegFormer [35] is AdamW with a learning rate of 6e-05 and weight decay rate of 0.01. Evaluations take place every 400 iterations on the RealOcc dataset. In all the experiments, images were resized to 512×512. Basic data augmentations, such as random horizontal flips, 30 degrees of random rotation, and photometric distortion were applied. The loss function for all the experiments is the binary cross-entropy loss with online hard example mining (OHEM) [30].

Datasets. The definition of training datasets is shown in Table 3. Different combinations of training datasets can be found in Table 4. All the synthetic datasets were generated using C-WO, *i.e.*, CelebAMask-HQ-WO (Train) (see Table 2). The validation sets are the RealOcc, while the test sets are COFW (training set) [6] and RealOcc-Wild.

Evaluation Metrics. We evaluate the model performance by comparing mIoU, *i.e.*, the mean Intersection over Union, across all the classes on the validation dataset.

4.2. Results and Analysis

We evaluate our two data generation methods with different approaches. NatOcc focuses on producing naturalistic occluded faces, while RandOcc produces occlusions with a general approach.

NatOcc. Table 4 shows that the models trained with our NatOcc dataset (C-WO-NatOcc-SOT, C-WO-NatOcc) performed at a level or better than the real-world occluded face dataset (C-CM), demonstrating the effectiveness of our method, especially with the CNN models (PSPNet [39] and DeepLabv3+ [9]). As shown in Figure 11, the CNN models trained with NatOcc datasets can segment the faces more accurately than the ones trained with the C-CM. Adding C-CM to the NatOcc dataset further enhanced the model's performance. In particular, SegFormer [35] trained on C-WO-NatOcc-SOT and C-CM achieved the highest score of 98.02. Although our synthetic datasets do not cover most of the occlusions (*e.g.*, glasses, face masks, camera, and food), the models trained with our NatOcc dataset can generalize well to these unseen occluded faces. We include more results in our supplementary material.

RandOcc. The RandOcc dataset (C-WO-RandOcc) brought huge improvement over the one without occlusions (C-WO), and it was not far behind the real-world occluded dataset (C-CM). Figure 11 shows that SegFormer [35] trained with RandOcc datasets was able to generalize to unseen occlusions despite the unnatural occluders of RandOcc datasets (see Figure 9). The advancement in deep learning models has made this type of general data generation practical and possible. C-WO-Mix, the mixture of RandOcc and NatOcc further improved the performance closer to C-CM. Further exploration can be done in future work to create a more robust and effective general synthetic occluded data generation method that can save human labor and time preparing synthetic data for every application.

Hard Occluders (Transparent/Translucent). SegFormer [35] trained on the dataset without any real transparent/translucent occluders (C-WO-NatOcc) failed to accurately detect transparent glasses, as shown in Figure 10. Although C-WO-RandOcc contains synthetic transparent/translucent occluders generated by RandOcc in Sec-

Table 3. Definition of datasets in our experiments.

Class	Definition
C-Original	CelebAMask-HQ-WO (Train) and CelebAMask-HQ-O with original masks.
C-CM	CelebAMask-HQ-WO (Train) and CelebAMask-HQ-O with corrected masks.
C-WO	CelebAMask-HQ-WO (Train).
C-WO-NatOcc	One set of hand-occluded (without color transfer) face dataset and one set of COCO-occluded face dataset generated by NatOcc with C-WO.
C-WO-NatOcc-SOT	One set of hand-occluded (with color transfer) face dataset and one set of COCO-occluded face dataset generated by NatOcc with C-WO.
C-WO-RandOcc	Two sets of occluded face dataset generated by RandOcc with C-WO.
C-WO-Mix	Half set of C-WO-RandOcc and one set of C-WO-NatOcc.

Table 4. **Overall Performance:** Results of PSPNet [39], DeepLabv3+ [9] and SegFormer [35] with different combination of datasets. The best results for each validation set are marked in bold. The metrics are mIoU (higher is better).

	Quantity	**RealOcc** (mIoU)			**COFW (Train)** (mIoU)			**RealOcc-Wild** (mIoU)		
		PSPNet	DeepLabv3+	SegFormer	PSPNet	DeepLabv3+	SegFormer	PSPNet	DeepLabv3+	SegFormer
C-Original	29,200	89.52	88.13	88.33	89.64	88.62	91.36	85.21	82.05	85.24
C-CM	29,200	96.15	96.13	97.42	91.82	92.77	**94.87**	91.33	91.01	95.16
C-WO	24,602	89.38	89.01	91.36	89.53	88.97	92.24	83.86	84.14	86.72
C-WO + C-WO-NatOcc	24,602 + 49,204	96.65	96.51	97.30	90.71	91.21	94.30	91.34	91.70	94.17
C-WO + C-WO-NatOcc-SOT	24,602 + 49,204	96.35	96.59	97.18	**92.32**	91.74	93.55	**93.26**	92.69	94.27
C-WO + C-WO-RandOcc	24,602 + 49,204	95.09	95.21	96.53	90.82	91.35	93.14	89.54	89.68	92.84
C-WO + C-WO-Mix	24,602 + 73,806	96.55	96.66	97.37	90.99	91.20	93.74	92.14	91.84	94.40
C-CM + C-WO-NatOcc	29,200 + 49,204	**97.28**	**97.33**	97.95	91.61	92.66	94.86	92.13	**93.81**	**95.43**
C-CM + C-WO-NatOcc-SOT	29,200 + 49,204	97.17	97.29	**98.02**	92.07	**92.91**	94.60	92.84	93.73	94.53

tion 3.3, the model trained with C-WO-RandOcc alone was not able to detect transparent glasses as occlusion. However, the model trained on C-WO-Mix, *i.e.*, a mixture of the datasets generated by NatOcc and RandOcc, was able to detect transparent glasses as occlusion, as shown in Figure 10, demonstrating the possibility to simulate transparent/translucent objects. The examples in Figure 10 show that RandOcc is complementary to NatOcc to potentially detect transparent/translucent objects.

Impact of an Unclean Dataset. Table 4 and Figure 11 show that the C-Original that has incorrect masks (ignored some occlusions) performed poorly and surprisingly worse than the C-WO, which does not have any occluded faces. This indicates that unclean datasets have a significant negative impact on the model's performance even with SegFormer [35], the better deep learning model. Thus, it shows the importance of clean data.

Class Imbalance. Occlusions would increase the already large amount of background pixels in an image. Therefore, the IoU of background significantly affects the score of the mIoU, resulting in the high mIoU of the dataset with incorrect annotations (C-Original) and the dataset without occlusions (C-WO). For instance, the background and face IoU of PSPNet [39] trained on C-original are 95.04 and 83.99, respectively, and this result in a mIoU of 89.52. Despite having high mIOU, PSPNet trained on C-original perform badly on occlusion-aware segmentation as shown in Figure 11. Therefore, better metrics such as frequency weighted IoU can be explored in future work to better evaluate the model performance.

Figure 10. The visual results on transparent/translucent occluders of SegFormer [35] trained on C-WO-NatOcc, C-WO-RandOcc and C-WO-Mix. Models trained on C-WO-NatOcc and C-WO-RandOcc failed to detect transparent glasses as occlusion accurately. In contrast, the model on C-WO-Mix can detect transparent glasses as occlusions, proving that RandOcc is complementary to NatOcc to detect transparent/translucent objects.

Impact of Color Transfer. Despite the minor differences between the model performance trained with C-WO-NatOcc and C-WO-NatOcc-SOT (*i.e.*, with or without color

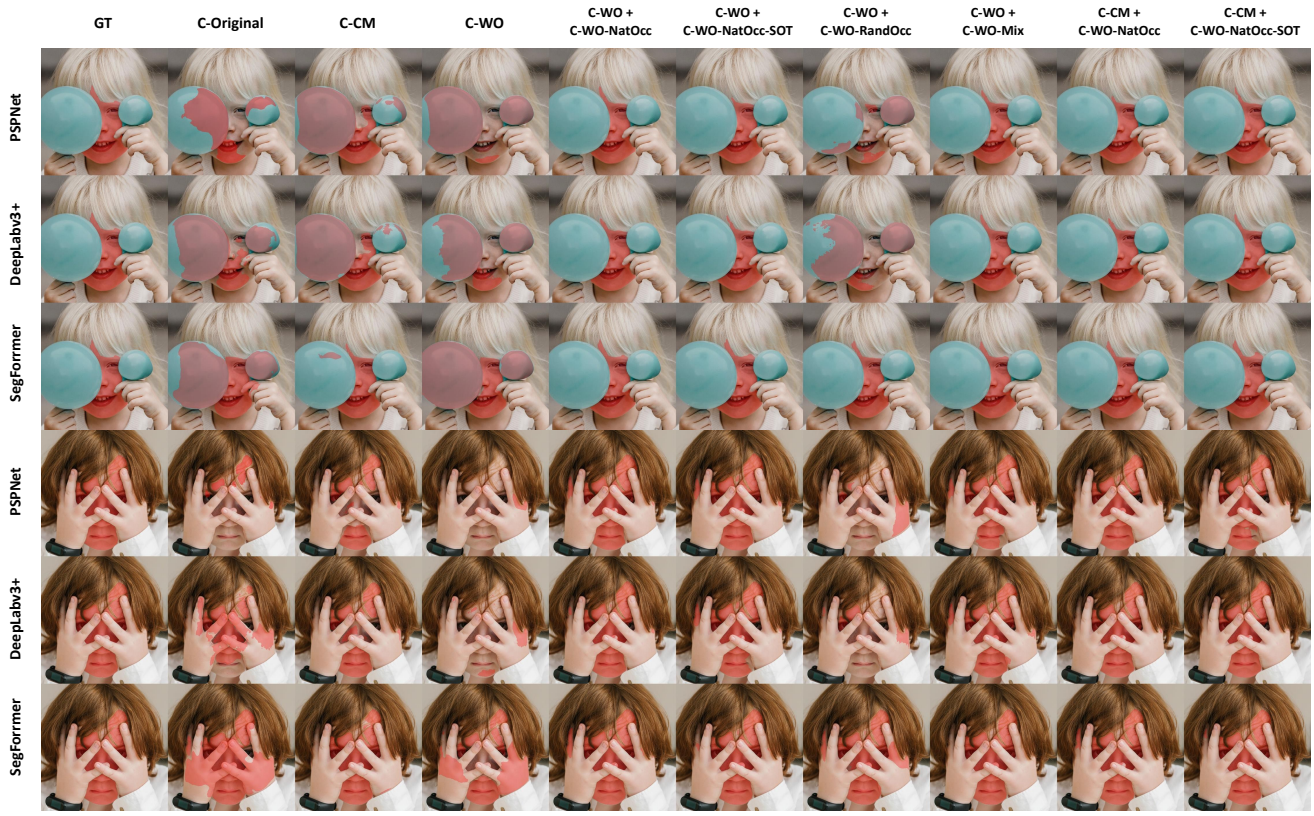

Figure 11. Occlusion-aware face segmentation results of PSPNet [39], DeepLabv3+ [9] and SegFormer [35] trained on different variations of datasets. Overall, the models trained with our NatOcc datasets obtain comparable or even better results than the models trained with real-occluded dataset (C-CM). SegFormer [35] trained with our RandOcc shows a huge improvement over C-Original and C-WO.

transfer), no conclusion can be made on which one is better. Without a hand-occluded face dataset with annotation masks, the effectiveness of color transfer cannot be evaluated fairly. In reality, faces might be occluded by the hands of another person, which are very different in color from that of the faces. A mixture of hands with and without color transfer might be able to complement each other and achieve higher performance.

Robustness Analysis. The robustness of the trained models was evaluated on real-world occluded faces in the wild, specifically COFW (Train) [6] and RealOcc-Wild. Overall, the models trained with our synthetic datasets performed better and can boost the performance of the real-world occluded face dataset (C-CM). In addition, the results on CelebAMask-HQ-WO (Test) shows that despite the improvement on segmenting occluded faces, NatOcc and RandOcc datasets did not bring any negative side effect on segmenting non-occluded faces. Please refer to our supplementary material for additional test results.

5. Conclusion and Future Work

In this paper, we proposed NatOcc and RandOcc, two occlusion generation methods that are proven to be effective for occlusion-aware segmentation, even for unseen occlusions. Besides, we contributed the corrected masks and new categories of CelebAMask-HQ [22]. To facilitate model evaluation, we introduce two high-resolution real-world occluded face datasets, RealOcc and RealOcc-Wild. Further, we benchmark several representative baselines and provide insights for future exploration. As for future work, devising more advanced techniques to produce higher-quality synthetic data to better simulate real-world data could be interesting. Moreover, improving generalization of synthetic data can be further studied. In addition, our benchmark can be enriched with higher-quality synthetic data and more baselines for a more comprehensive analysis. Last but not least, practical applications of occlusion-aware face segmentation in face-related tasks (*e.g.*, face recognition and face-swapping) could be further explored.

Acknowledgments. This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Mahmoud Afifi. 11k hands: gender recognition and biometric identification using a large dataset of hand images. *Multimedia Tools and Applications*, 2019. 4, 5

[2] Mahmoud Afifi and Abdelrahman Abdelhamed. Afif4: Deep gender classification based on adaboost-based fusion of isolated facial features and foggy faces. *Journal of Visual Communication and Image Representation*, 2019. 2

[3] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1949–1957, 2015. 4

[4] Borut Batagelj, Peter Peer, Vitomir Štruc, and Simon Dobrišek. How to correctly detect face-masks for covid-19 from visual information? *Applied Sciences*, 11(5), 2021. 2

[5] Nicolas Bonneel and David Coeurjolly. Sliced partial optimal transport. *Association of Computing Machinery Transactions on Graphics (Proceedings of Special Interest Group on Computer Graphics and Interactive Techniques)*, 38(4), jul 2019. 4, 5

[6] 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. 1, 2, 4, 6, 8

[7] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. 5

[8] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 4

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, 2018. 2, 6, 7, 8

[10] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2, 6

[11] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 6

[12] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

[13] Mustafa Ekrem Erakın, Uğur Demir, and Hazım Kemal Ekenel. On recognizing occluded faces in the wild. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2021. 2

[14] Fevziye Irem Eyiokur, Hazım Kemal Ekenel, and Alexander Waibel. A computer vision system to help prevent the transmission of covid-19. *arXiv preprint arXiv:2103.08773*, 2021. 2

[15] Golnaz Ghiasi and Charless C. Fowlkes. Using segmentation to predict the absence of occluded parts. In *British Machine Vision Conference*, 2015. 2

[16] 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. 6

[17] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 2

[18] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[19] Andrew Kae, Kihyuk Sohn, Honglak Lee, and Erik Learned-Miller. Augmenting CRFs with Boltzmann machine shape priors for image labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 1, 2, 3

[20] 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. 4

[21] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 4

[22] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 3, 4, 8

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4, 5

[24] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9361–9370, 2021. 5

[25] 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. 1, 2, 3, 4

[26] Behnaz Nojavanasghari, Charles E Hughes, Tadas Baltrušaitis, and Louis-Philippe Morency. Hand2face: Automatic synthesis and recognition of hand over face occlusions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction*, pages 209–215. IEEE, 2017. 3, 5

[27] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020. 1

[28] Rafael Redondo and Jaume Gibert. Extended labeled faces in-the-wild (elfw): Augmenting classes for face segmentation. *arXiv preprint arXiv:2006.13980*, 2020. 1, 2, 3, 4, 5

[29] 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. 1, 3, 4

[30] 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. 6

[31] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 773–782, 2019. 1

[32] Hitika Tiwari, Min-Hung Chen, Yi-Min Tsai, Hsien-Kai Kuo, Hung-Jen Chen, Kevin Jou, K. S. Venkatesh, and Yong-Sheng Chen. Self-supervised robustifying guidance for monocular 3d face reconstruction, 2021. 1

[33] Aisha Urooj and Ali Borji. Analysis of hand segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4710–4719, 2018. 3, 4

[34] Kentaro Wada. Labelme: Image Polygonal Annotation with Python. 4

[35] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 6, 7, 8

[36] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[37] Xiangnan Yin and Liming Chen. Faceocc: A diverse, high-quality face occlusion dataset for human face extraction. *arXiv preprint arXiv:2201.08425*, 2022. 3, 4

[38] Xiangnan Yin, Di Huang, Zehua Fu, Yunhong Wang, and Liming Chen. Segmentation-reconstruction-guided facial image de-occlusion. *arXiv preprint arXiv:2112.08022*, 2021. 1, 2, 3, 4

[39] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 6, 7, 8

[40] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13001–13008, Apr. 2020. 3