OpenForensics: Large-Scale Challenging Dataset For Multi-Face Forgery Detection And Segmentation In-The-Wild

Trung-Nghia Le[1], Huy H. Nguyen[2], Junichi Yamagishi[1,2], and Isao Echizen[1,2,3]

[1]National Institute of Informatics, [2]The Graduate University for Advanced Studies (SOKENDAI), [3]University of Tokyo
https://sites.google.com/view/ltnghia/research/openforensics/

Figure 1. Examples from our OpenForensics dataset (best viewed online in color with zoom-in). Can you spot the forged faces and identify the manipulated areas in these images? The answers are in the supplementary material.

Abstract

The proliferation of deepfake media is raising concerns among the public and relevant authorities. It has become essential to develop countermeasures against forged faces in social media. This paper presents a comprehensive study on two new countermeasure tasks: multi-face forgery detection and segmentation in-the-wild. Localizing forged faces among multiple human faces in unrestricted natural scenes is far more challenging than the traditional deepfake recognition task. To promote these new tasks, we have created the first large-scale dataset posing a high level of challenges that is designed with face-wise rich annotations explicitly for face forgery detection and segmentation, namely OpenForensics. With its rich annotations, our OpenForensics dataset has great potentials for research in both deepfake prevention and general human face detection. We have also developed a suite of benchmarks for these tasks by conducting an extensive evaluation of state-of-the-art instance detection and segmentation methods on our newly constructed dataset in various scenarios.

1. Introduction

Continuing advances in deep learning have led to impressive improvements in deepfake methods (*i.e*., deep learning-based face forgery), which can change the target person's identity [32, 1, 64, 42]. Emerging techniques such as autoencoder (AE) models and generative adversarial networks (GANs) enable transferring one person's face to another person while retaining the original facial expression and head pose [68, 67, 56, 66]. The realistic appearance synthesized with deepfake methods is drawing much attention in the fields of computer vision and graphics because of the potential application of such methods in a wide range of areas [18, 26, 30, 79, 39]. Moreover, falsified AI-synthesized images/videos have raised serious concerns about individual harassment and criminal deception [6, 62, 12]. To address threats posed by spoofing and impersonation attacks, it is essential to develop countermeasures against face forgeries in digital media.

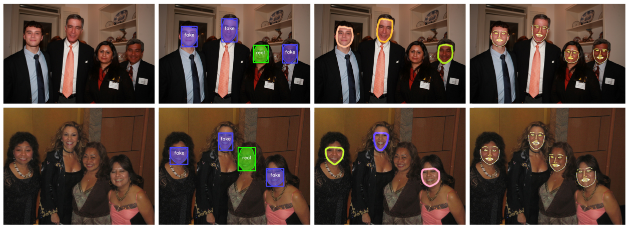


Figure 2. Face-wise multi-task ground truth in OpenForensics dataset (best viewed online in color with zoom-in). From left to right, original images followed by overlaid ground truth bounding box and segmentation mask, forgery boundary, and general facial landmarks.

Conventional face forgery recognition methods [2, 54, 53] require the input of given face regions. Therefore, they can process only one face at a time, and processing multiple faces sequentially is time-consuming. Moreover, their performance greatly depends on the accuracy of the independent face detection method used. Given that these methods have been evaluated only in laboratory environments using images with a simple background and a single clear front face [31, 78], they are not ready for deployment in the real world, where the contexts are much more diverse and challenging than simple staged scenarios.

It has thus become essential to develop methods that can

Table 1. Basic information about deepfake datasets. “Cls.”, “Det.” and “Seg.” stand for classification, detection, and segmentation, respectively. Pristine scenarios are originally collected images/videos used to generate fake data. Unique fake scenarios are fake images/videos ignoring perturbations. Released scenarios are number of real/fake (or both) images/videos publicly released by authors.

Dataset	Year	Task	GT Type	Fake Identity	#Face Per Image	Face Occlusion	#Pristine Scenario	#Unique Fake Scenario	#Released Scenario	Data Augmentation
DF-TIMIT [31]	2018	Cls.	Image label	Other videos	1	✗	320	320	640	✗
UADFV [78]	2019	Cls.	Image label	Other videos	1	✗	49	49	98	✗
FaceForensics++ [61]	2019	Cls.	Image label	Other videos	1	✗	1,000	4,000	5,000	✗
Google DFD [16]	2019	Cls.	Image label	Other videos	1	✗	363	3,068	3,431	✗
Facebook DFDC [14]	2020	Cls.	Image label	Other videos	1	✗	48,190	104,500	128,154	✓
Celeb-DF [46]	2020	Cls.	Image label	Other videos	1	✗	590	5,639	6,229	✗
DeeperForensics [27]	2020	Cls.	Image label	Hired actors	1	✗	1,000	1,000	10,000	✓
WildDeepfake [84]	2020	Cls.	Image label	N/A	1	✗	0	707	N/A	✗
OpenForensics	2021	Det. / Seg.	BBox/Mask	GAN	> 1	✓	45,473	70,325	115,325	✓

effectively process multiple faces simultaneously from an input image. To our best knowledge, no methods have been proposed for face forgery detection and segmentation officially. We attribute this partially to the lack of a large-scale dataset for training and testing. To encourage more studies in this field, we present four contributions in this paper.

First, we present a comprehensive study on tasks related to massive face forgery in-the-wild. Particularly, we introduce two new tasks: *multi-face forgery detection and segmentation in-the-wild*. This is the first formal exploration of these tasks to the best of our knowledge. Previous work has explored only single-face forgery recognition.

Second, we propose generating an infinite number of fake individual identities using GAN models for non-target face-swapping without repeatedly training a deepfake AE. Our proposed forgery workflow reduces the cost of synthesizing fake data.

Third, using the proposed forgery workflow, we introduce a novel image dataset to support the development of multi-face forgery detection and segmentation tasks. Our newly constructed *OpenForensics dataset is the first large-scale dataset designed for these tasks*. It consists of 115K unrestricted images with 334K human faces. Unlike existing datasets, ours contains various backgrounds and multiple people of various ages, genders, poses, positions, and face occlusions. All images have *face-wise rich annotations* supporting multiple tasks, such as forgery category, bounding box, segmentation mask, forgery boundary, and general facial landmarks (see Figs. 1 and 2). The dataset can thus support not only multi-face forgery detection and segmentation tasks but also conventional tasks involving the general human face.

Fourth, we present a benchmark suite to facilitate the evaluation and advancement of these tasks. We conducted an extensive evaluation and in-depth analysis of state-of-the-art instance detection and segmentation models in various scenarios.

The whole dataset, evaluation toolkit, and trained models will be freely available on our project page[1].

2. Related Work

2.1. Existing Forensic Datasets

Table 1 summarizes basic information about existing forensic datasets. The DF-TIMIT dataset [31] has 640 fake videos crafted from Vid-TIMIT dataset [63] using Faceswap-GAN [64]. The UADFV dataset [78] consists of 98 videos, half of which are fake, created using FakeAPP [18]. The FaceForensics++ dataset [61] contains 1000 pristine videos from YouTube and 4000 synthetic videos manipulated using deepfake methods [1, 68, 32, 67]. The Google DFD dataset [16] includes 3068 fake videos. The Facebook DFDC dataset [14] contains 128K original and manipulated videos created using various deepfake and augmentation methods [59, 24, 79, 56, 28]. The Celeb-DF dataset [46] comprises YouTube celebrity videos and 5,639 fake videos. The DeeperForensics dataset [27] consists of 10K manipulated videos using a deepfake VAE and augmentations on 1000 original videos in the FaceForensics++ dataset. The WildDeepfake dataset [84] contains face sequences extracted from 707 deepfake videos collected from the Internet. As shown in Table 1, our OpenForensics is the first dataset designed for face forgery detection and segmentation.

Existing forensic datasets were created by dividing long videos into short ones, leading to that even pristine videos have the same background. Subsequent synthesizing many fake videos from one pristine video resulted in lots of similar backgrounds. Deep models trained on the existing datasets may not generalize well to the real world due to the repeated background. In contrast, our large-scale image dataset contains diverse backgrounds. Inspired by the work of Dolhansky *et al.* [14] and Jiang *et al.* [27], we systematically applied a mixture of perturbations to raw manipulated images to imitate real-world scenarios. With the existing datasets, a deepfake model needs to be trained on each pair of videos to swap human identities, yielding a considerable number of models requiring training. In contrast, a massive number of fake faces in our dataset are synthesized by GAN without repeatedly re-training deepfake models. While existing datasets were developed for only the single-face forgery classification task, our dataset is the first one designed for multi-face forgery detection and segmentation

[1] https://sites.google.com/view/ltnghia/research/openforensics

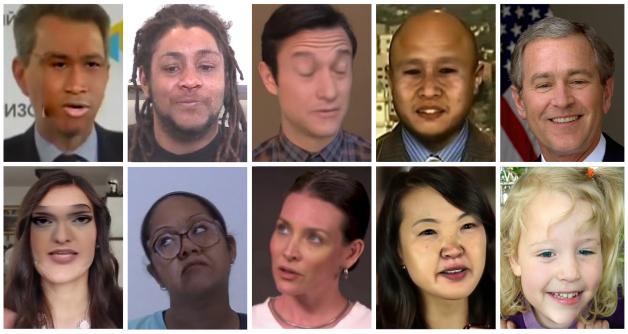


Figure 3. Visual artifacts of forged faces in datasets. From left to right, FaceForensics++ [61], DFDC [14], DeeperForensics [27], Celeb-DF [46], and our OpenForensics. Faces generated in our dataset have the highest resolution and best quality.

tasks, which require more annotation than the classification task. Our dataset can also be utilized for various general face-related tasks.

2.2. Face Manipulation and Generation

A number of deepfake open-source techniques for swapping human faces have been released [32, 1, 64]. These techniques have gradually evolved from using hand-crafted features [32] to using deep learning by training AE architectures [1] and GAN models [64] [42] to achieve realism. Facial reenactment techniques have been developed for transferring expressions [68, 67, 56]. Different techniques such as 3D reconstruction [68] and neural textures [67] were used to preserve the target skin color and lighting conditions. Boundary latent space [75] and disentangle shape [66] were combined with AE models to morph expressions. In addition to transferring expressions, the head pose can be controlled by using a recurrent neural network to enhance naturalness [56] by using different modalities [74] and by using human interpretable attributes and actions [70].

Subsequently proposed techniques for face synthesis use deep learning. They generally use GAN for facial attribute translation [8, 9, 28, 29], for identity-attribute combination [3], for identified characteristics removal [51], and for interactive semantic manipulation [40, 83]. Facial disentangled features are being interpreted in different latent spaces, resulting in more precise control of attribute manipulation in face editing [28, 29, 65, 60].

Existing deepfake methods require face pairs for specific training, meaning that the cost of training is very high. Training requires sequences of images; thus, these methods are practical only for videos, and the generated faces usually have low-resolution. Although existing face synthesis methods can generate high-quality faces, the synthesized faces are oriented to the front and are not consistent with the original faces if the original faces are not close to the distribution of the training data. We combine these two approaches to generate an infinite number of fake human identities without repeatedly training the AEs. We achieve this by transforming GAN-based high-quality synthesized faces into original poses.

Table 2. Scale of object detection/segmentation datasets.

Dataset	Year	Object Type	#Annotated Images	Ground-Truth Type
COCO [48]	2014	General object	200,000	Coarse mask
CityScapes [11]	2016	Road object	25,000	Coarse&Fine mask
WiderFace [77]	2016	Human face	32,200	Bounding box
SESIV [37]	2019	Salient object	5,700	Fine mask
ADV [38]	2020	Accident object	10,000	Fine mask
CAMO++ [36]	2021	Camouflaged object	5,500	Fine mask
OpenForensics	2021	Forged face	115,325	Fine mask

Table 3. Image distribution in OpenForensics dataset.

Subset	#Images	#Faces	#Real Faces	#Forged Faces
Training	44,122	151,364	85,392	65,972
Validation	7,308	15,352	4,786	10,566
Test-Development	18,895	49,750	21,071	28,670
Test-Challenge	45,000	117,670	49,218	68,452
Total	115,325	334,136	160,67	173,660

2.3. Face Forgery Classification

Researchers have been investigating the problem of face forgery classification, which is generally regarded as merely a binary classification problem (real/fake). The research task is also called 'deepfake detection,' but the term 'detection' may lead to a misunderstanding of the fundamental task of *object detection*. Early methods exploited inconsistencies created by visual artifacts in deepfake images and videos by analyzing biological clues such as eye blinking [44], head pose [78], skin texture [49], and iris and teeth color [50]. A few works investigated artifacts in affine face warping [45] or in the blending boundary [43] to distinguish real and fake faces. Most current methods are data-driven, directly training deep networks on real and fake images and videos [2, 54, 61, 53, 82, 71]. They do not rely on specific artifacts.

Existing face forgery classification approaches do not have a face localization ability. They can work only on a single cropped face; thus, their performance relies heavily on independent face detection performed as pre-processing. To the best of our knowledge, ours is the first work addressing multi-face detection and segmentation in-the-wild.

3. Large-Scale OpenForensics Dataset

The emergence of new tasks and datasets has led to rapid progress in human research areas [77, 13, 55, 20, 19]. However, research on human forgery prevention is only now beginning, and the field is still immature with work only on the face forgery classification task. With this in mind, our goal is to study and develop a dataset that will support challenging new forgery research tasks in both the computer vision and forensic communities.

3.1. Dataset Construction

As shown in Fig. 4, the dataset construction workflow includes three main steps: real human image collection, forged face image synthesis, and multi-task annotation.

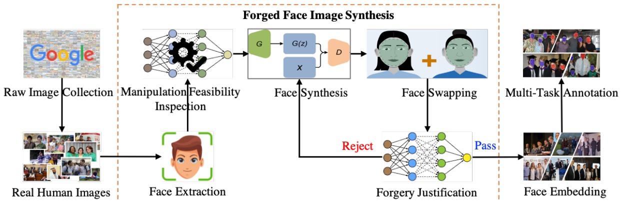


Figure 4. Dataset construction workflow: 1) collect raw images and manually select real face images; 2) synthesize forged face images (for each original extracted face, new identities are repeatedly generated until swapped faces can spoof our simple classifier); 3) perform face-wise multi-task annotation.

3.1.1 Real Human Image Collection

We collected raw images from Google Open Images [34] and removed images without people. Images consisting of unreal human faces (*e.g.*, images on money and in books, magazines, cartoons, and sketches) or human-like objects (*e.g.*, dolls, robots, and sculptures) were also removed. We ended up with 45,473 images, which were used as pristine data.

3.1.2 Forged Face Image Synthesis

Figure 4 shows an overview of the process used to synthesize forged face images. First, all faces in the real human images are extracted and checked in the manipulation feasibility inspection module to see whether they can be manipulated. This is done using various conditions (*e.g.*, face size, image quality, and blurring) and a random manipulation probability. If manipulation is feasible, the image undergoes a cyclical process. Inspired by GAN-based face synthesis [9, 29], we first extract the facial identity latent vector and modify it using random values. The modified latent vector is then fed into GAN models [65, 60] to generate a new face. The synthesized face is subsequently transformed into an original pose. Feasible manipulation regions in the synthesized face (*e.g.*, regions inside facial landmarks or the entire face) are extracted and blended into the original face using Poisson blending [58] and a color adaptation algorithm in the face-swapping module, with the final result being a new identity. The new identity image is then tested to determine whether it can spoof a simple classifier (*i.e.*, XceptionNet [10]) in the forgery justification module, which is trained to distinguish real and fake identities. Those for which spoofing is successful are overlaid onto the original image. The others are discarded, and new faces are generated. We provide detailed implementation and training of networks in the supplementary material.

Our synthesis workflow features the ability to synthesize an unlimited number of fake identities at low cost for non-target face-swapping without paired training. Meanwhile, other deepfake methods use a limited number of fake identities extracted from videos and perform paired training using deep models for target face-swapping. They thus require much time and resources to synsthesize datasets. Our synthesis approach also overcomes the limitations of existing approaches. Existing approaches [61, 14, 27] generate low-resolution faces (typically less than 256 × 256 pixels) while our approach generates faces with *higher resolution (*i.e., 512×512 *pixels) and better visual quality* (cf. Fig. 3). Our use of Poisson blending [58] and a color adaptation algorithm to reduce the color mismatch between the synthesized and original face (Fig. 3) *enhances the naturalness of the forged faces*. We also *improve the smoothness of the blending mask* by extracting 68 facial landmark points and training face segmentation models, resulting in fine boundaries and complete facial coverage (see Fig. 2 for different blending masks). The blending masks used to create existing datasets were either rectangular or rough convex hulls between the eyebrows and lower lip, resulting in incomplete facial coverage or visible boundaries (cf. Fig 3).

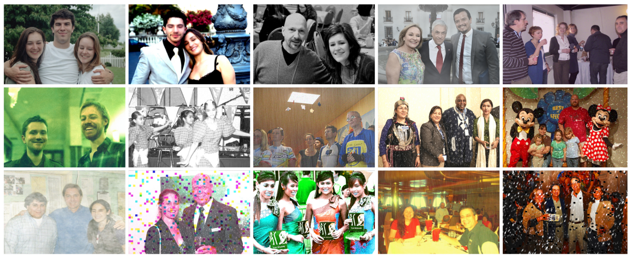


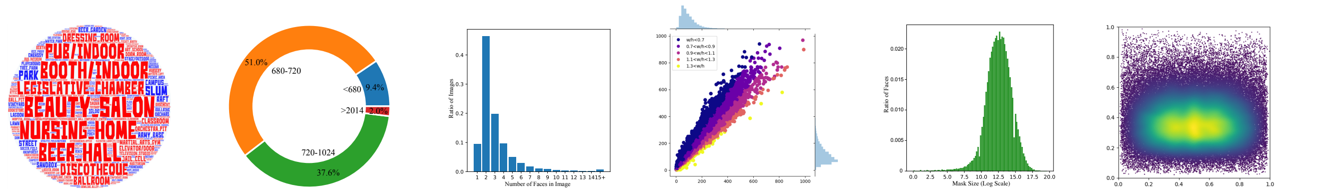
Figure 5. Example images in test-challenge set (three levels: easy, medium, and hard from top to bottom). Each image contains at least one forged face. See supplementary material for overlaid ground truth.

Finally, we randomly split the accepted images into separate training, validation, and test-development sets (ratio of 60:10:30). Table 3 shows the distribution of images and faces in our newly constructed OpenForensics dataset.

3.1.3 Challenging Scenario Augmentation

To enhance the challenges posed by our OpenForensics dataset for real-world face forgery detection and segmentation, we applied various perturbations to better simulate contexts in natural scenes, resulting in a test-challenge subset. Various augmented operators are divided into overarching groups.

- Color manipulation: Hue change, saturation change, brightness change, histogram adjustment, contrast addition, grayscale conversion.
- Edge manipulation: edge detection and alteration.
- Block-wise distortion: color grouping, color pooling, color quantization, and pixelation.
- Image corruption: elastic deformation, jigsaw distortion, JPEG compression, noise addition, and dropout.
- Convolution mask transformation: Gaussian blurring, motion blurring, sharpening, and embossing.
- External effect: fog, cloud, sun, frost, snow, and rain.



a) Scene word cloud b) Image resolution c) Faces per image d) Bounding box size e) Mask size f) Face centroid

Figure 6. Distributions in OpenForensics dataset (best viewed online in color with zoom-in). In image scene distribution, red represents indoor scenes and blue represents outdoor scenes (percent of indoor scenes is 63.7%). There are 2.9 faces per image on average.

These augmentations are divided into three intensity levels (*i.e.*, easy, medium, and hard) to ensure diverse scenarios. For each level, random-type augmentation is applied separately or as a mixture, resulting in 45,000 images. Example images in the test-challenge set are shown in Fig. 5.

3.2. Dataset Description

Task Diversity. Existing deepfake datasets [61, 14, 27, 46] focus exclusively on video-wise labels for classification. In contrast, we aim to exploit the face-wise ground truth, which requires much more annotation effort, to advance further forgery analysis. Each face was labeled with various ground-truths such as forgery category (real/fake), bounding box, segmentation mask, forgery boundary, and facial landmarks (cf. Fig. 2). Our rich annotation can be utilized for various tasks and even multi-task learning.

Dataset Size. OpenForensics is one of the largest detection and segmentation datasets (cf. Table 2) and is large enough to train and evaluate deep networks. This should encourage more research in this field.

Diverse Scenarios. Existing datasets [61, 14, 27, 46] were released as short videos. Although they contain a vast number of images, frames in a short video are similar and do not contribute much to the training of deep networks. With these datasets, data sampling is usually used for training deep networks to avoid overfitting and to reduce training time. We define similar frames in a short video as a 'scenario' and assert that training using a diversity of scenarios helps to make deep networks more effective. Table 1 shows that the OpenForensics dataset is an order of magnitude larger than existing datasets in terms of the number of scenarios, with only slightly fewer than in the DFDC dataset.

Image Scene. Existing deepfake datasets [61, 46] contain limited types of image scenes, such as indoor scenes and television scenes. In contrast, the OpenForensics dataset contains various types of scenes. We computed scenes using a pre-trained model on the large-scale Places2 dataset [81]. Figure 6(a) shows the distribution as a word cloud, with the various outdoor scenes accounting for 36.3% of the images.

Image Resolution. Figure 6(b) shows the distribution of image resolutions in the OpenForensics dataset. The large number of high-resolution images, which provide more face boundary details for model training, results in better performance.

Multiple Faces Per Image. Existing deepfake datasets [61, 14, 27, 46] mostly have only one face per image. In contrast, the OpenForensics dataset has multiple faces per image (2.9 on average). Figure 6(c) shows the distribution.

Face Characteristics. Figures 6(d and e) show the distribution of faces in the OpenForensics dataset by bounding box size and mask size (*i.e.*, number of pixels covering face). OpenForensics contains faces of various sizes, from tiny to large. The distribution of face centroids in Fig. 6(f) shows that the faces tend to be near the image center. In addition, the ratio of male and female faces is 50:50, and there is a diversity of ages. More details are provided in the supplementary material.

Data Augmentation. Deep models trained on existing deepfake datasets may not perform well in the real world due to overfitting caused by image similarity in the training data. Although strong deep models have obtained very high accuracy [54, 43], even near 100%, they may easily fail in the real world if they do not share a close distribution with the training dataset. To simulate real-world contexts in the OpenForensics dataset, diverse perturbations were used to improve scenario diversity so as to better imitate real-world data distributions. Improvements have been made to a couple of existing datasets by using simple perturbations, which have increased their size. For instance, the DFDC dataset [14] and DeeperForensics dataset [27] have been improved by applying geometric and color transforms, adding noise, blurring, and overlaying objects.

3.3. User Study

To evaluate the visual quality of the images in the OpenForensics dataset and human performance in face forgery detection, we conducted a user study with 200 participants, 80 of whom are experts, who can provide knowledgeable opinions due to their researching deepfakes. The study results can fairly reflect the performance of both experts and non-experts.

The study was conducted on the OpenForensics dataset and four existing deepfake datasets: FaceForensics++ [61], DFDC [14], Celeb-DF [46] and DeeperForensics [27]. For each dataset, we randomly selected 600 images and pre-

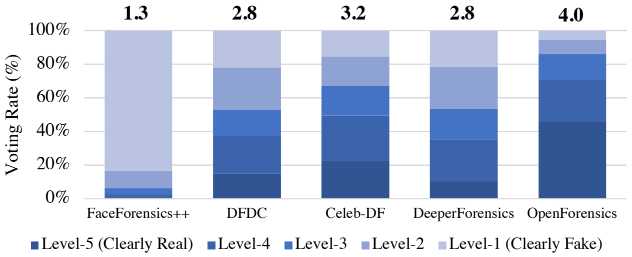


Figure 7. Distributions of image realism scores for five compared datasets. Mean opinion scores (MOS) are shown at top of bars. OpenForensics dataset achieved highest MOS and had highest percentage of level-5 scores.

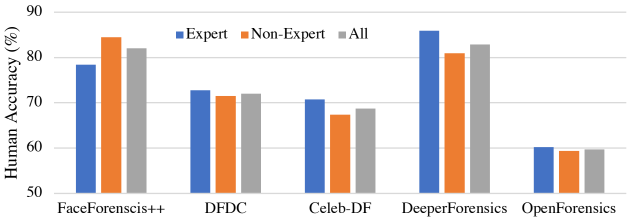


Figure 8. Human accuracy in face forgery classification. Images in OpenForensics dataset were most effective in spoofing both experts and non-experts.

pared a virtual platform for the participants.

We argue that participants can quickly see that a face is fake if they see two similar images but different people, leading to unfair comparison with existing datasets. In addition, the forgery identification may becomes difficult if forged faces are mixed with real faces. To investigate these hypothesises, our user study focused on both two cases: cropped faces to eliminate surrounding contexts and full images with multi-face.

Evaluation of Image Realism. We cropped the forged heads, which had been doubly extended from the faces, to ensure that the upper-half of each person was completely extracted. The participants were asked to view 200 forged head images and then provide feedback on each image's realism in the form of a score 1 to 5, corresponding to 'clearly fake,' 'weakly unreal,' 'borderline,' 'almost real,' and 'clearly real.' As shown by the results in Fig. 7, the visual quality of the images in the OpenForensics dataset was highly evaluated by most of the participants. That is, the forged faces in the OpenForensics dataset were judged to be the most realistic. Our dataset achieved the highest mean opinion score (MOS) 4.0, much higher than that of the second-best dataset Celeb-DF (3.2). The DeeperForensics and DFDC datasets had medium-quality images (MOS of 2.8). The FaceForensics++ dataset had the most unrealistic images (MOS of only 1.3).

Human Performance on Face Forgery Classification. We again cropped the heads similar to the cropping done for the evaluation of image realism. The participants were asked to view a mixture of 400 images randomly composed of pristine and forged heads with a ratio of 50:50. After viewing each image, the participants were asked whether the image was 'real' or 'fake.' As shown in Fig. 8, the participants had the most trouble distinguishing between the real and fake images in the OpenForensics dataset. This is evidenced by the OpenForensics dataset having the lowest overall accuracy (59.7%), followed by Celeb-DF (68.7%), DFDC (72.0%), FaceForensics++ (82.0%), and DeeperForensics (82.9%,). The graph also shows that both experts and non-experts had difficulty distinguishing between the real and fake images in our dataset. It is interesting that although experts could recognize fake faces better than non-experts, they incorrectly identified real faces with low quality, low resolution, or low contrast (*i.e.*, FaceForensics++ dataset). We attribute this to their overconfidence and their belief that GANs might generate such faces, leading to misidentification.

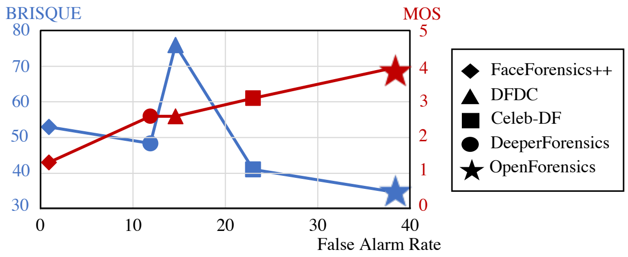


Figure 9. Correlation between visual properties and human ability to recognize forged faces. The ability to recognize forged faces depends on image realism (higher MOS is better) and visual quality (lower BRISQUE is better). False alarm rate is higher for images with higher quality and more realism, meaning that OpenForensics is the best dataset in terms of having realistic images.

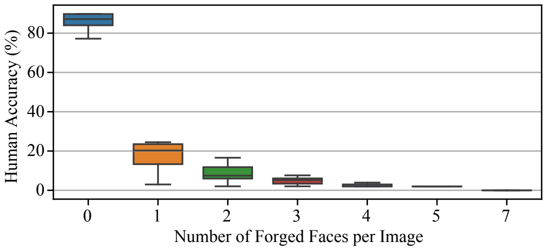


Figure 10. Human performance on multi-face forgery detection. Accuracy deceased as number of forged faces increased.

Figure 9 illustrates the correlation between the visual properties and the human ability to recognize forged faces. The ability to recognize forged faces depends on image realism, resulting in an increased false alarm rate as realism improves (*i.e.*, as the MOS increases). The graph shows that a large number of participants misclassified forged faces in the OpenForensics dataset as real faces. The OpenForensics dataset had the highest MOS (4.0) and the highest false alarm rate (34.6%). The figure also shows that the BRISQUE score [52] of the OpenForensics dataset was the lowest (35.2), which indicates that the images in our dataset have the best visual quality. Reducing image quality (*i.e.*, increasing the BRISQUE score) would affect human observation, resulting in a lower false alarm rate.

Human Performance on Multi-Face Forgery Detection. The participants were asked to view a set of 160 images, each with multiple persons and each consisting of both pristine and forged faces randomly selected, of only pristine faces, or of only forged faces. They were asked to identify the number of forged faces in each image. Figure 10 shows that detection accuracy was the highest (86%) when there were no forged faces in the images and tended to drop as the number of forged faces increased. This can be explained that when there are many faces in an image, participants tend to less carefully check each face and guess that all the faces are real. That explains why the accuracy is high when all faces are real while it significantly reduces when forged faces exist. Indeed, when the number exceeded 7, accuracy dropped to 0%. Even people find it extremely difficult to identify forged faces among mixture of pristine and forged faces on in-the-wild images, highlighting the challenge of our OpenForensics dataset.

4. Benchmark Suite

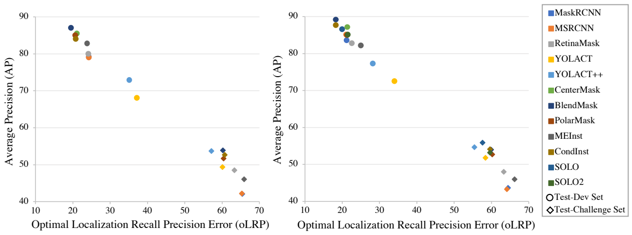
4.1. Baseline Methods

We conducted a competitive benchmark for multi-face forgery detection and segmentation. To this end, we trained and evaluated the latest instance detection and segmentation methods in various scenarios. The methods were MaskRCNN [22], MSRCNN [25], RetinaMask [17], YOLACT [4], YOLACT++ [5], CenterMask [41], BlendMask [7], PolarMask [76], MEInst [80], CondInst [69], SOLO [72], and SOLO2 [73]. MaskRCNN and MSRCNN are well-known two-stage models that perform detect-then-segment slowly. The YOLACT ones [4, 5] are early single-stage models aimed at real-time performance. The remaining methods are widely used modern single-stage models that overcome accuracy and processing time problems. Among them, the SOLO ones [72, 73] directly output masks without computing bounding boxes.

All the methods were used with the same backbone (FPN-ResNet50 [47, 23]) to make the comparison fair. We trained models on PCs with 32 GB of RAM and a Tesla P100 GPU. The models were initialized with ImageNet weights [33] and trained on our training set for 12 epochs. The base learning rate was decreased by $1/10$ at the 8^{th} and 11^{th} epochs. Other settings were in accordance with the default public configurations provided by the authors.

4.2. Evaluation Metrics

We evaluated the methods using standard COCO-style average precision (AP) [48]. We report the results for mean AP and AP on different scales (AP_S, AP_M, AP_L, where S, M, and L represent small, medium, and large objects). We also evaluated the methods using the localization recall precision (LRP) error [57]. We report the results for mean optimal LRP (oLRP) and its error components including localization (oLRP$_{Loc}$), the false positive rate (oLRP$_{FP}$), and the false negative rate (oLRP$_{FN}$).



a) Forgery Detection

b) Forgery Segmentation

Figure 11. Benchmark results achieved by baseline methods for multi-face forgery multi-task on OpenForensics dataset (best viewed online in color with zoom-in). Test-dev set results reflect benchmark performance on standard images while test-challenge set results reflect robustness for unseen images. Lower oLRP error is better while higher AP is better. BlendMask had the best performance, and YOLACT++ was the most robust. Result for CenterMask on test-challenge set is out of the range and is shown in Table 5.

4.3. Overall Evaluation

As shown in Fig. 11, BlendMask had the best performance, with the highest AP and lowest oLRP error for both the detection and segmentation tasks on standard images. The other modern single-stage methods also had high performance, and the two-stage methods had medium performance. The YOLACT methods had the worst performance on both tasks because they are mainly focused on real-time processing. YOLACT++ and BlendMask were the most robust for unseen images.

4.4. Multi-Face Forgery Detection Benchmark

Table 4 shows detailed results for the multi-face forgery detection task broken down by metric. They show that BlendMask had the best performance, achieving the highest AP (87.0) and the lowest oLRP error (19.5). BlendMask also achieved the highest AP for all object scales. The modern single-stage methods (*i.e.*, BlendMask, PolarMask, and CondInst) had minor location errors and false positive rates while the two-stage methods (*i.e.*, MaskRCNN and MSRCNN) had low false negative rates.

4.5. Multi-Face Forgery Segmentation Benchmark

With the emergence of explainable AI (XAI) technology [15, 21, 35, 38], it is useful to identify manipulated areas in detected faces. Therefore, we also evaluated segmentation performance. As shown in Table 4, for the multi-face forgery segmentation task, the trends in the ranking of method performance are similar to those for the detection task. BlendMask had the best segmentation performance,

Table 4. Benchmark results for multi-face forgery detection and segmentation on test-dev set. Higher AP is better while lower oLRP error is better. Best and second-best results are shown in blue and red, respectively.

Method	Year	Multi-Face Forgery Detection								Multi-Face Forgery Segmentation							
		AP↑	AP$_S$↑	AP$_M$↑	AP$_L$↑	oLRP↓	oLRP$_{Loc}$↓	oLRP$_{FP}$↓	oLRP$_{FN}$↓	AP↑	AP$_S$↑	AP$_M$↑	AP$_L$↑	oLRP↓	oLRP$_{Loc}$↓	oLRP$_{FP}$↓	oLRP$_{FN}$↓
MaskRCNN [22]	ICCV 2017	79.2	29.9	80.2	79.5	24.3	9.5	2.7	4.0	83.6	16.1	82.1	85.8	21.2	7.6	3.0	4.2
MSRCNN [25]	CVPR 2019	79.0	29.5	80.1	79.5	24.3	9.6	2.7	3.8	85.1	16.8	84.2	86.8	21.1	7.7	2.6	4.4
RetinaMask [17]	arXiv 2019	80.0	30.9	80.2	80.7	24.2	9.0	3.0	4.6	82.8	16.4	80.6	85.1	22.6	8.1	2.9	4.9
YOLACT [4]	ICCV 2019	68.1	12.5	67.1	69.3	37.2	13.4	6.3	8.7	72.5	3.1	67.0	75.7	34.0	11.4	6.4	8.7
YOLACT++ [5]	TPAMI 2020	72.9	20.9	73.4	73.6	31.5	12.1	4.0	5.8	77.3	6.5	73.9	80.0	28.2	10.0	3.9	6.5
CenterMask [41]	CVPR 2020	85.5	32.0	85.2	86.2	21.1	6.8	3.3	5.9	87.2	16.5	85.0	89.4	21.4	6.1	3.2	7.8
BlendMask [7]	CVPR 2020	87.0	32.7	86.3	88.0	19.5	6.2	2.4	6.2	89.2	19.8	87.3	91.0	18.3	5.4	2.5	6.3
PolarMask [76]	CVPR 2020	85.0	27.4	85.4	85.7	20.7	6.6	2.5	6.6	85.0	15.3	83.3	87.0	21.3	6.9	2.5	6.6
MEInst [80]	CVPR 2020	82.8	26.0	82.7	83.4	23.8	7.6	4.1	6.8	82.2	13.9	81.5	83.3	25.0	8.1	4.0	7.2
CondInst [69]	ECCV 2020	84.0	29.4	83.6	84.8	20.8	7.4	2.3	5.2	87.7	18.1	85.1	89.8	18.3	5.9	2.4	5.3
SOLO [72]	ECCV 2020	-	-	-	-	-	-	-	-	86.6	15.4	85.6	88.4	20.0	6.6	2.1	6.0
SOLO2 [73]	NeurIPS 2020	-	-	-	-	-	-	-	-	85.1	13.7	83.7	87.1	21.5	7.1	3.1	5.8

Table 5. Benchmark results for multi-face forgery detection and segmentation on test-challenge set. Higher AP is better while lower oLRP error is better. Best and second-best results are shown in blue and red, respectively.

Method	Year	Multi-Face Forgery Detection								Multi-Face Forgery Segmentation							
		AP↑	AP$_S$↑	AP$_M$↑	AP$_L$↑	oLRP↓	oLRP$_{Loc}$↓	oLRP$_{FP}$↓	oLRP$_{FN}$↓	AP↑	AP$_S$↑	AP$_M$↑	AP$_L$↑	oLRP↓	oLRP$_{Loc}$↓	oLRP$_{FP}$↓	oLRP$_{FN}$↓
MaskRCNN [22]	ICCV 2017	42.1	11.8	46.2	40.5	65.4	13.6	29.3	40.0	43.7	4.7	44.3	44.0	64.4	11.8	29.4	41.2
MSRCNN [25]	CVPR 2019	42.2	11.8	45.9	40.8	65.3	13.7	29.6	39.9	43.3	5.2	44.6	43.5	64.1	11.8	30.4	39.6
RetinaMask [17]	arXiv 2019	48.5	12.8	51.0	48.1	63.3	12.6	33.2	34.6	48.0	4.7	46.5	49.7	63.3	11.8	30.9	38.0
YOLACT [4]	ICCV 2019	49.4	5.6	49.6	50.3	60.1	15.3	23.2	29.9	51.8	1.4	47.2	54.6	58.4	13.5	23.4	30.1
YOLACT++ [5]	TPAMI 2020	53.7	11.1	54.0	54.8	57.1	14.1	19.7	29.3	54.7	2.4	50.7	57.9	55.4	12.2	20.0	30.0
CenterMask [41]	CVPR 2020	0.03	0.4	0.0	0.0	99.5	29.7	97.7	97.9	0.02	0.0	0.0	0.0	99.6	28.3	97.9	98.4
BlendMask [7]	CVPR 2020	53.9	13.5	56.6	53.5	60.2	10.6	26.5	37.4	54.0	7.1	54.5	54.5	59.9	9.8	26.4	38.4
PolarMask [76]	CVPR 2020	51.7	12.3	53.2	51.5	60.4	10.7	24.6	39.5	52.7	5.3	54.1	37.6	60.2	10.4	24.7	39.5
MEInst [80]	CVPR 2020	46.1	8.6	49.9	44.9	65.9	12.4	34.6	39.7	46.0	3.8	49.0	45.2	66.2	12.6	34.8	39.8
CondInst [69]	ECCV 2020	52.7	12.6	55.3	51.8	60.7	11.5	28.3	35.3	54.1	6.5	55.2	53.8	59.6	10.0	26.7	37.3
SOLO [72]	ECCV 2020	-	-	-	-	-	-	-	-	55.9	3.9	53.3	57.3	57.6	11.3	24.6	33.0
SOLO2 [73]	NeurIPS 2020	-	-	-	-	-	-	-	-	53.2	3.6	52.1	54.0	59.6	11.0	24.5	37.2

with AP of almost 90 and an oLRP error of approximately 18 for the test-dev set.

Images in the real world obviously contain human faces of various sizes. It is thus essential to investigate detection and segmentation abilities on different scales. Table 4 shows that all the baseline methods achieved high performance for only medium-size and large faces. Performance decreased with the face size, resulting in a marginal difference between small faces and medium/large faces in both detection and segmentation. These results illustrate the challenges of our OpenForensics dataset, which consists of enormous face sizes.

Similar to the detection task, we found that single-stage methods, which are based on dense detection, have fewer FP errors while the two-stage ones, which are based on sparse detection, have fewer FN errors. Therefore, the development of post-processing using NMS and the improvement of RPN, respectively, can help to improve forgery detectors.

4.6. Robustness Evaluation

We conducted experiments to evaluate the robustness of the methods on our test-challenge set, which simulates scenarios in the real world. Table 5 shows that YOLACT++ and BlendMask were the most robust methods for unseen images. CenterMask was the least robust method, which is attributed to its results containing a lot of noise, resulting in extremely high false positive and false negative rates.

Tables 4 and 5 show a substantial drop in performance for all methods for unseen images, which are beyond the distribution of the training set. Although existing methods can work well on standard images, their robustness is weak for unseen images. Even leading forgery-identification methods in the deep learning era remain limited and cannot yet effectively address real-world challenges (Top-1: $AP < 60$ on test-challenge set). Hence, *multi-face forgery detection and segmentation problems in-the-wild are still far from being solved, leaving much room for improvement*. These results also illustrate the challenges of our OpenForensics dataset.

5. Conclusion and Outlook

As part of our comprehensive study on multi-face forgery detection and segmentation in-the-wild, we created a large-scale dataset. In-depth analysis of our OpenForensics dataset demonstrated its diversity and complexity. We also conducted an extensive benchmark by evaluating state-of-the-art instance segmentation methods in various experimental settings. We expect that our OpenForensics dataset will boost research activities in deepfake prevention. We intend to continue enlarging this dataset to accompany future developments in deepfake technology.

Thanks to the rich annotations in our OpenForensics dataset, there are a number of foreseeable research directions that will provide a solid basis for forgery and general face studies, including fundamental research (*e.g.*, weak/semi-supervised/self-supervised detection/segmentation, universal network for multiple tasks) and specific research (*e.g.*, anti-forgery robustness detection, forgery boundary detection, forgery ranking, face anonymization, face detection/segmentation, facial landmark prediction).

Acknowledgements. This work was partially supported by JSPS KAKENHI Grants JP16H06302, JP18H04120, JP21H04907, JP20K23355, and JP21K18023, and by JST CREST Grants JPMJCR18A6 and JPMJCR20D3, including the AIP challenge program, Japan.

References

[1] Deepfakes software for all. https://github.com/deepfakes/faceswap, 2017. [Online; accessed 18-Feb-2021]. 1, 2, 3

[2] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *International Workshop on Information Forensics and Security*, pages 1–7, 2018. 1, 3

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *Conference on Computer Vision and Pattern Recognition*, pages 6713–6722, 2018. 3

[4] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *International Conference on Computer Vision*, 2019. 7, 8

[5] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact++: Better real-time instance segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 2020. 7, 8

[6] John Brandon. Terrifying high-tech porn: Creepy 'deepfake' videos are on the rise. https://www.foxnews.com/tech/terrifying-high-tech-porn-creepy-deepfake-videos-are-on-the-rise, 2018. [Online; accessed 18-Feb-2021]. 1

[7] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2020. 7, 8

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Conference on Computer Vision and Pattern Recognition*, 2018. 3

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Conference on Computer Vision and Pattern Recognition*, 2020. 3, 4

[10] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Conference on Computer Vision and Pattern Recognition*, pages 1800–1807, 2017. 4

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 3

[12] Jesse Damiani. A voice deepfake was used to scam a ceo out of $243,000. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/, 2019. [Online; accessed 18-Feb-2021]. 1

[13] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Conference on Computer Vision and Pattern Recognition*, June 2020. 3

[14] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020. 2, 3, 4, 5

[15] Derek Doran, Sarah Schulz, and Tarek R Besold. What does explainable ai really mean? a new conceptualization of perspectives. *arXiv preprint:1710.00794*, 2017. 7

[16] Nicholas Dufour, Andrew Gully, Per Karlsson, Alexey Victor Vorbyov, Thomas Leung, Jeremiah Childs, and Christoph Bregler. Deepfakes detection dataset by google & jigsaw, 2019. 2

[17] Cheng-Yang Fu, Mykhailo Shvets, and Alexander C. Berg. RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free. In *arXiv preprint arXiv:1901.03353*, 2019. 7, 8

[18] Yaroslav Goncharov. Faceapp - face editor, makeover and beauty app. https://www.faceapp.com/, 2016. [Online; accessed 18-Feb-2021]. 1, 2

[19] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, 2020. 3

[20] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B. Gotway, and Jianming Liang. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In *Medical Image Computing and Computer Assisted Intervention*, pages 137–147, 2020. 3

[21] H. Hagras. Toward human-understandable, explainable ai. *Computer*, 51(9):28–36, 2018. 7

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision*, pages 2980–2988, 2017. 7, 8

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778, June 2016. 7

[24] Dong Huang and Fernando De la Torre. Facial action transfer with personalized bilinear regression. In *European Conference on Computer Vision*, 2012. 2

[25] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask Scoring R-CNN. In *Conference on Computer Vision and Pattern Recognition*, 2019. 7, 8

[26] Neocortext Inc. Reface. https://hey.reface.ai/, 2020. [Online; accessed 18-Feb-2021]. 1

[27] 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 *Conference on Computer Vision and Pattern Recognition*, June 2020. 2, 3, 4, 5

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2, 3

[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Conference on Computer Vision and Pattern Recognition*, 2020. 3, 4

[30] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep

video portraits. *ACM Transactions on Graphics*, 37(4), 2018. 1

[31] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 1, 2

[32] Marek Kowalski. 3d face swapping. https://github.com/MarekKowalski/FaceSwap, 2016. [Online; accessed 18-Feb-2021]. 1, 2, 3

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 7

[34] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 2020. 4

[35] Rodney LaLonde, Drew Torigian, and Ulas Bagci. X-caps: Diagnosis capsule network for interpretable medical image diagnosis. In *Medical Image Computing and Computer Assisted Intervention*, 2020. 7

[36] Trung-Nghia Le, Yubo Cao, Tan-Cong Nguyen, Minh-Quan Le, Khanh-Duy Nguyen, Thanh-Toan Do, Minh-Triet Tran, and Tam V Nguyen. Camouflaged instance segmentation in-the-wild: Dataset and benchmark suite. *arXiv preprint arXiv:2103.17123*, 2021. 3

[37] Trung-Nghia Le and Akihiro Sugimoto. Semantic instance meets salient object: Study on video semantic salient instance segmentation. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1779–1788, Jan 2019. 3

[38] Trung-Nghia Le, Akihiro Sugimoto, Shintaro Ono, and Hiroshi Kawasaki. Attention r-cnn for accident detection. In *IEEE Intelligent Vehicles Symposium*, 2020. 3, 7

[39] MIT Open Learning. Tackling the misinformation epidemic with "in event of moon disaster". https://news.mit.edu/2020/mit-tackles-misinformation-in-event-of-moon-disaster-0720, 2020. [Online; accessed 18-Feb-2021]. 1

[40] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Conference on Computer Vision and Pattern Recognition*, 2020. 3

[41] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2020. 7, 8

[42] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1, 3

[43] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Conference on Computer Vision and Pattern Recognition*, June 2020. 3, 5

[44] Y. Li, M. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *International Workshop on Information Forensics and Security*, pages 1–7, 2018. 3

[45] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 3

[46] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Conference on Computer Vision and Pattern Recognition*, June 2020. 2, 3, 5

[47] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Conference on Computer Vision and Pattern Recognition*, 2017. 7

[48] 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, 2014. 3, 7

[49] Zhengzhe Liu, Xiaojuan Qi, and Philip H.S. Torr. Global texture enhancement for fake face detection in the wild. In *Conference on Computer Vision and Pattern Recognition*, June 2020. 3

[50] F. Matern, C. Riess, and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *Winter Applications of Computer Vision Workshops*, pages 83–92, 2019. 3

[51] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixe. Ciagan: Conditional identity anonymization generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition*, June 2020. 3

[52] Anish Mittal, Anush K Moorthy, and Alan C Bovik. Blind/referenceless image spatial quality evaluator. In *ASILOMAR Conference on Signals, Systems and Computers*, pages 723–727, 2011. 6

[53] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *International Conference on Biometrics: Theory, Applications and Systems*, 2019. 1, 3

[54] H. H. Nguyen, J. Yamagishi, and I. Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *International Conference on Acoustics, Speech and Signal Processing*, 2019. 1, 3, 5

[55] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Arxiv Pre-print: 2012.15029*, 2020. 3

[56] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *International Conference on Computer Vision*, pages 7184–7193, 2019. 1, 2, 3

[57] Kemal Oksuz, Baris Cam, Emre Akbas, and Sinan Kalkan. Localization recall precision (lrp): A new performance met-

ric for object detection. In *European Conference on Computer Vision*, 2018. 7

[58] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *SIGGRAPH*, pages 313–318, 2003. 4

[59] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. Deepfacelab: A simple, flexible and extensible face swapping framework. *Arxiv Pre-print Arxiv:2005.05535*, 2020. 2

[60] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Conference on Computer Vision and Pattern Recognition*, 2020. 3, 4

[61] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner. Faceforensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision*, pages 1–11, Oct 2019. 2, 3, 4, 5

[62] Sigal Samuel. A guy made a deepfake app to turn photos of women into nudes. it didn't go well. https://www.vox.com/2019/6/27/18761639/ai-deepfake-deepnude-app-nude-women-porn, 2019. [Online; accessed 18-Feb-2021]. 1

[63] Conrad Sanderson and Brian C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *International Conference on Advances in Biometrics*, page 199–208, 2009. 2

[64] Shaoanlu. A denoising autoencoder, adversarial losses and attention mechanisms for face swapping. https://github.com/shaoanlu/faceswap-GAN, 2018. [Online; accessed 18-Feb-2021]. 1, 2, 3

[65] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Conference on Computer Vision and Pattern Recognition*, 2020. 3, 4

[66] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *European Conference on Computer Vision*, September 2018. 1, 3

[67] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics*, 38(4), 2019. 1, 2, 3

[68] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Computer Vision and Pattern Recognition*, 2016. 1, 2, 3

[69] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European Conference on Computer Vision*, 2020. 7, 8

[70] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Icface: Interpretable and controllable face reenactment using gans. In *Winter Conference on Applications of Computer Vision*, 2020. 3

[71] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. In *International Joint Conference on Artificial Intelligence*, pages 3444–3451, 2020. 3

[72] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: Segmenting objects by locations. In *European Conference on Computer Vision*, 2020. 7, 8

[73] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic, faster and stronger. In *Conference on Neural Information Processing Systems*, 2020. 7, 8

[74] O. Wiles, A.S. Koepke, and A. Zisserman. X2face: A network for controlling face generation by using images, audio, and pose codes. In *European Conference on Computer Vision*, 2018. 3

[75] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *European Conference on Computer Vision*, 2018. 3

[76] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *Conference on Computer Vision and Pattern Recognition*, 2020. 7, 8

[77] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Conference on Computer Vision and Pattern Recognition*, 2016. 3

[78] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *International Conference on Acoustics, Speech and Signal Processing*, pages 8261–8265, 2019. 1, 2, 3

[79] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *International Conference on Computer Vision*, 2019. 1, 2

[80] Rufeng Zhang, Zhi Tian, Chunhua Shen, Mingyu You, and Youliang Yan. Mask encoding for single shot instance segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2020. 7, 8

[81] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 2017. 5

[82] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. Two-stream neural networks for tampered face detection. In *Conference on Computer Vision and Pattern Recognition Workshops*, pages 1831–1839, 2017. 3

[83] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Conference on Computer Vision and Pattern Recognition*, June 2020. 3

[84] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *International Conference on Multimedia*, page 2382–2390, 2020. 2