Shallow-FakeFaceNet: A CNN-Based Detection Framework for GAN-Generated and Handcrafted Facial Forgeries

Furong Li
Hunan Mechanical electrical polytechnic, Changsha, 410000, Hunan, China
E-mail: yushaoo2@163.com

Keywords: fake face detection, GAN-generated images, data augmentation for forgery detection, AI-based image authentication, image processing

Received: March 17, 2025

With the rapid advancement of digital media technologies, facial image manipulation has become increasingly sophisticated. Both handcrafted editing tools and deep generative models such as Generative Adversarial Networks (GANs) can produce convincingly fake facial images, posing significant threats like misinformation and identity fraud. In this study, we introduce a novel Handcrafted Facial Manipulation (HFM) dataset, containing 1,527 manually edited images across multiple modification types and complexity levels. To detect these fakes along with GAN-generated images, we propose a lightweight neural network called Shallow-FakeFaceNet (SFFN), optimized for low-resolution images (64×64 and 128×128). The detection pipeline includes MTCNN-based face cropping, noise filtering, GAN-based facial super-resolution for enhancing small images, and extensive image augmentation using both Keras and ImgAug. Unlike prior works that rely on fragile metadata, our model operates solely on RGB image data, making it robust against common forgery tactics. Experimental results show that SFFN achieves an AUROC of 72.52% on handcrafted fakes and 93.99% on GAN-generated faces, outperforming several state-of-the-art models. This approach offers a practical, real-world solution for fake media detection in social platforms and biometric verification systems.

Povzetek: Raziskava uvaja Shallow-FakeFaceNet in ročno ustvarjen HFM nabor, ki omogočata zaznavanje ročnih ter GAN-generiranih ponaredkov obrazov, s čimer izboljšata robustnost metapodatkovno neodvisne forenzične detekcije.

1 Introduction

With developments in computer software for image manipulation such as Adobe Photoshop [1], facial photographs can now easily be manipulated, and users can create highly realistic forged images. Tools such as automatic area selection and foreground-aware inpainting enhance the quality of editing, and it is hard to detect forged images. Simultaneously, social media sites and online guides provide step-by-step guides, and it is simple for even beginners to create sophisticated fake material for the media. These manipulated images are employed for disinformation, defamation, and identity theft, which cause serious security concerns in interactive media and digital communications.

In addition to standard editing processes, deep learning-based generative models, particularly Generative Adversarial Networks (GANs) [2], have revolutionized artificial content creation with photorealistic images, video, and audio. The advancements are for creative sectors but pose hazards as well because GAN-synthesized fake faces can deceive humans as well as machine learning algorithms. Abusive use includes generated false identities, deepfake adult content, and election misinformation campaigns [3]. Further, deepfake-basis impersonation attacks infiltrate biometric authentication systems, increasing the threat of heightened privacy, security, and digital trust issues.

To mitigate these challenges, a number of detection techniques have been proposed. Previous research, e.g., Huh et al. [4], has analyzed metadata and image compression for authentication. Metadata can be manipulated, though, so these methods are unreliable. Similarly, Adobe's deep learning-based detection networks are ineffective against composite facial manipulations and entirely synthesized GAN-generated faces [5]. As both handcrafted and GAN-generated fake faces are challenging to detect, there remains a strong demand for successful neural network-based solutions.

To bridge this gap, we present Shallow-FakeFaceNet (SFFN), a novel neural network-based classifier capable of detecting handcrafted and GAN-generated spoof facial images from RGB image data alone, making it resistant to metadata manipulation. We also present Handcrafted Facial Manipulation (HFM) dataset, a collection of 1,527 spoof facial images handcrafted with Adobe Photoshop, carefully crafted to improve fake face detection performance. Experimental results validate that SFFN performs better than current forgery detection models, establishing a new benchmark for classifying fake images in interactive media.

Our contribution may be stated as follows:

- Construction of the Handcrafted Facial Manipulation (HFM) dataset, an open-source repository of 1,527 handcrafted forged facial images, enriching datasets for digital forgery detection.
- Introduction of Shallow-FakeFaceNet (SFFN), a light-weight neural network-based method that distinguishes both human-edited and GAN-generated forged faces while shattering dependence on easily manipulated metadata.
- Demonstration of SFFN's performance via comprehensive evaluations, surpassing state-of-the-art forgery detection techniques in various testing scenarios.
- Design of a real-world end-to-end detection pipeline, specific for social media security, biometric authentication, and multimedia forensics to counter deepfakes.

2 Literature review

Various research studies have focused on digital image forensics to detect tampered images with techniques such as compression artifact analysis, metadata verification, and noise pattern detection [6-7]. JPEG compression history and frequency-domain based traditional forensic approaches have proven successful in detecting straightforward image manipulation. As machine learning-based synthesis models evolved, the effectiveness of traditional approaches decreased. Deep learning- and neural networks-based classifiers have now become state-of-the-art in detecting forgery with higher accuracy and the ability to fight against handcrafted and AI-aided forgeries.

Researchers have prioritized GAN-generated image detection, given the prevalent use of deepfake technology today. Goodfellow et al. [8] first introduced Generative Adversarial Networks (GANs) as a strong tool for synthetic media generation, and subsequent advances by Karras et al. [5] enabled the generation of ultra-realistic human faces. While GANs have positive uses in art, gaming, and medical imaging, they have also been exploited to commit identity fraud, biometric security breaches, and disinformation operations. To address these threats, various research works have proposed deep learning-based detection models, e.g., Tariq et al. [9], who constructed models that are particularly designed to identify between GAN-generated images, and Zhou et al. [10], who examined CNN-based feature extraction approaches to identify inconsistencies in synthesized faces. Despite these initiatives, GAN-generated spoofed faces continue to be a hard problem due to their high realism and adaptive generative models.

To aid deepfake detection, various benchmark datasets have been developed. Rossler et al. [11] created FaceForensics++, a widely used dataset consisting of deepfake videos gathered from YouTube, while Li et al. [12] created CelebDF(v2) to remove bias and quality from existing deepfake datasets. However, handcrafted face forgeries—those produced manually by the application of photo-editing software such as Adobe Photoshop [1]—are underresearched. Contrary to GAN-created fakes, which come with novel generative patterns that are specifically handcrafted, carefully crafted forgeries tend to imply pixel-level specific alterations, thereby becoming more challenging to identify through conventional forensic devices. This justifies the acute need for human-edited as well as artificially created forgery datasets to facilitate detection models working well across multiple manipulation methods.

In addition to dataset limitations, deep learning-based forgery detection methods have also gained widespread popularity, with CNN-based models such as VGG16, ResNet, and Xception [13] being widely applied in image classification and anomaly detection. The models have demonstrated good performance in classifying forged images, but detecting highly realistic forgeries is still an evolving approach. Frequency-domain analysis techniques have also been researched by authors to detect forgery artifacts [14], as well as EXIF metadata self-consistency verification. These are simple to manipulate or delete, and thus such techniques are not highly robust when applied against high-fidelity forgeries. With these problems, therefore, there is a requirement for more robust, data-driven detection models that can identify both GAN-generated and handcrafted higher-fidelity forged faces.

3 Methodology

A) High-Quality Handcrafted Facial Manipulation Dataset

To encourage facial forgery detection research, we have established a manually created dataset of 1,527 forged images and 621 original images, all produced with varying editing complexities. Our university's skilled digital artists painstakingly created these forgeries manually using Adobe Photoshop CS6, with high realism and variety. Since hand-crafted forgery face datasets are not prevalent, our dataset addresses this gap by employing edits close to real-world forgeries found on social media platforms like Twitter, Instagram, and Facebook (Fig. 1). The source images were gathered using Google Image Search with the usage rights parameter set to Labeled for reuse with modification to prevent copyright infringement. To enhance diversity in the datasets, we added images of individuals of different ages, genders, and ethnicity as well as attributes that are difficult to handle, including heavy makeup, glasses, beards, and headwear. The fake images were then separated into three levels of complexity to simulate multiple levels of manipulation one can find on the internet. Lv.1 images symbolize obvious cut-and-paste manipulations without smoothing. Lv.2 images improve Lv.1 manipulations by blurring the edge of the region pasted. Lv.3 images further refine Lv.2 modifications even better by adjusting color

and lighting more to resemble actuality, increasing the level of difficulty to notice (Fig. 2).

We also utilized six different types of modifications in order to resemble different forgery techniques. These changes include changes to one's facial features such as the eyes, nose, or mouth (Modification (a)), changes that include more than one facial landmark (Modification (b)), half-face exchanges (Modification (c)), full-face substitutions (Modification (d)), additions like sunglasses or mustaches on the face (Modification (e)), and manipulations across more than one face in a single image (Modification (f)), as illustrated in Fig. 1. These variations cause the dataset to capture a broad set of manually created facial manipulations, providing a valuable resource for evaluating deepfake detection models.

Figure 1. Examples of manually crafted forged images showcasing different types of facial modifications, including partial feature swaps, multi-feature alterations, half-face and full-face swaps, accessory additions, and manipulations involving multiple faces.

B) Evaluation of HFM Dataset Quality

To assess the quality of our HFM dataset, we leverage the Inception Score (IS), an unbiased metric introduced by Salimans et al. [56] for synthetic image quality assessment. Originally developed for estimating GAN-generated images based on a crowd-sourcing service, IS bypasses the intrinsic human judgment using the Inception model [15] to derive the conditional label distribution. A dataset with a score of about 2.0 for a two-class problem is simpler to classify, while a score of around 1.0 indicates more difficulty. Our HFM dataset has an average IS score of 1.0046 with a standard deviation of 0.00078, proving its complexity and difficulty in classification.

Apart from IS, we also assess image quality in relation to frequency transformation, a method introduced by Durall et al. [58] to identify real vs. fake images based on spatial frequencies. Fig. 3 displays the single-dimensional power spectrum metrics of our HFM dataset against the PGGAN dataset [8]. As observed from Fig. 3(a), the PGGAN dataset has large frequency gaps, indicating a strong real-generated image difference. On the contrary, Fig. 3(b) shows how our HFM dataset has the minimum gap, meaning that fake images created manually in our dataset are more difficult to distinguish from real ones. This also makes our dataset more robust in mimicking real-world manipulation, and the detection is that much more difficult.

Algorithm 1: MTCNN Noise Filtering

Input: Input image x
Output: Filtered & cropped faces D_{faces}

$\mathcal{M} \leftarrow \text{MTCNN}(x)$ /* where $\mathcal{M} = \{m_1, m_2, m_3, ..., m_n\}$ is a set of detected faces from the MTCNN face detector. */
$r_i \leftarrow m_i^w + m_i^h$ /* where $i \in \{1, 2, 3 \ldots, n\}$ and each m_i has the attribute of width m_i^w and height m_i^h of the cropped face. */
$\mathcal{R} \leftarrow \{r_1, r_2, r_3, ..., r_n\}$
$r_{max} \leftarrow max(\mathcal{R})$
for $i \in \{1, 2, 3 \ldots, n\}$ **do**
  if $r_i \geq \frac{r_{max}}{\tau}$ **then**
    Append $D_{faces} \leftarrow m_i$
  else
    Discard m_i as Noise
  end
end

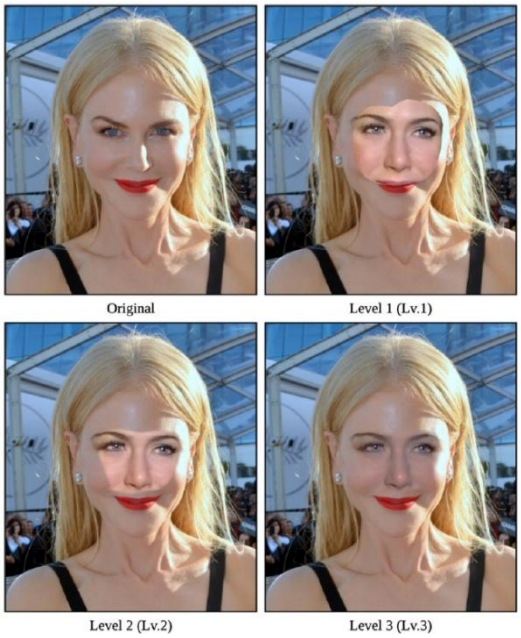
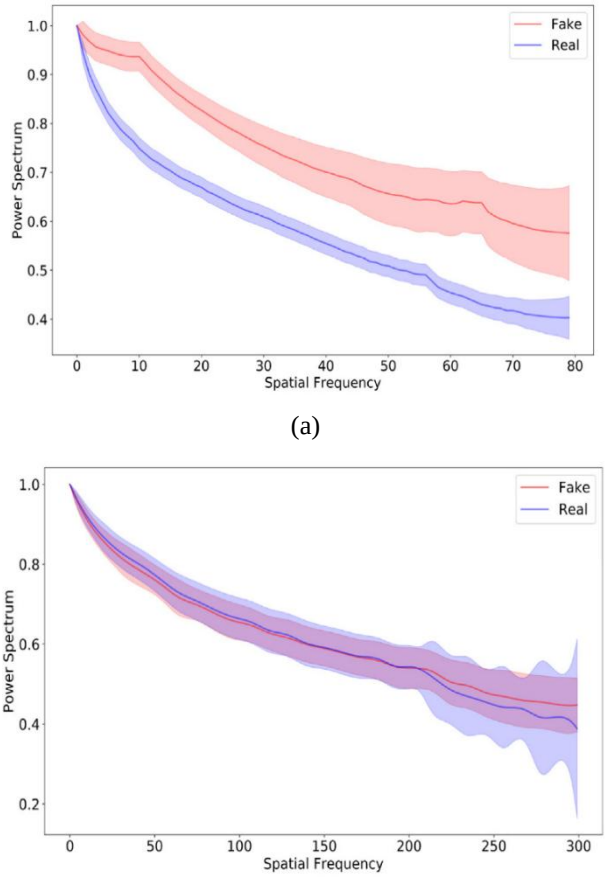


Figure 2: Examples of forged images categorized by complexity levels: Lv.1 (basic cut-and-paste edits with visible rough edges), Lv.2 (smoothed edges for improved blending), and Lv.3 (enhanced realism with adjusted color and lighting).

C) Facial Forgery Detection Framework

Our face forgery detection system is designed to differentiate between authentic and forged images according to a systematic pipeline. The process involves facial preprocessing and classification according to our Shallow-FakeFaceNet (SFFN) model, in which only RGB data and not metadata is used to prevent forgery exploitation.

As illustrated in Fig. 4, there are two significant phases included in the pipeline: preprocessing and classification. Preprocessing consists of (1) face region extraction, (2) false positive removal, (3) resolution refinement, and (4) data augmentation. The refined images are input to the classifier for training and manipulation detection using SFFN.



(a)

(b)

Figure 3: Comparison of 1D Power Spectrum statistics between (a) PGGAN dataset and (b) HFM dataset.

1) Face cropping and noise filtering

To crop the face region, we utilize MTCNN [16], which is a computational face detection method. The process, however, risks misclassifying non-facial objects (e.g., accessories, hands) or detecting occluded and small faces and thus resulting in false positives (Fig. 5). In multi-face conditions, erroneous detections would also make classification challenging.

In order to minimize such errors, we have a noise removal strategy, which is organized into Algorithm 1. On given

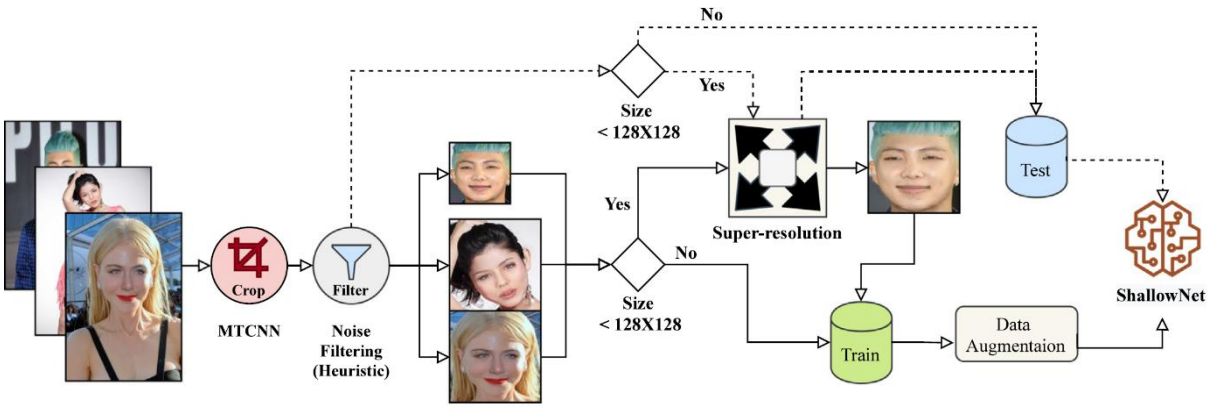


Figure 4: Overview of the fake face detection pipeline, consisting of two main stages: (1) Preprocessing, which includes face cropping, noise filtering, upscaling, and augmentation, and (2) Classification, where the preprocessed images are analyzed using the Shallow-FakeFaceNet (SFFN) model to detect facial manipulations.

Figure 5: Examples of false positives detected by the MTCNN face detector, including occluded faces and relatively small facial regions (highlighted in red boxes).

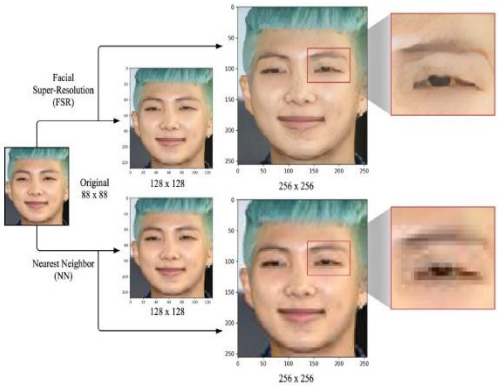


Figure 6: Comparison of image upscaling techniques. (Top row): Results from Facial Super-Resolution (FSR), which progressively enhances image quality while preserving facial details. (Bottom row): Results from Nearest Neighbor Upscaling (NNU), which

image x, we locate the faces using MTCNN with the detected face set denoted as M and each of which has width m_i^w and height m_i^h. We obtain the size ratio r_i using the following expression:

$$r_i = m_i^w + m_i^h \quad (1)$$

The maximum value from the set of objects detected, R, is used as a reference, and any object detected with size smaller than a certain threshold τ is eliminated as noise. For our experiments, $\tau = 1.732$ was used after empirical selection of false positives. If a detected object passes the filtering step but still gets mislabeled, it counts as training noise instead of a good detection.

2) Image upscaling for small faces

Approximately 8% of images in our HFM dataset are below 128×128 pixels, distinguishing real and fake faces becomes challenging due to limited pixel information. To enhance resolution, we compare Nearest Neighbor Upscaling (NNU) and Facial Super-Resolution (FSR) [17]. NNU replaces pixels with adjacent values, often causing pixelation artifacts, while FSR utilizes a GAN-based neural network with Facial Attention Loss to progressively upscale images in multiple steps, preserving facial details. As illustrated in Fig. 6, FSR significantly reduces pixelation, making it our preferred method for resolution enhancement.

The model is trained using the Adam optimizer, which optimizes learning rates adaptively for rapid convergence. This carefully crafted deep learning pipeline makes the framework scalable, adaptable, and capable of optimizing energy usage in real-time. By integrating AI, IoT, and predictive analytics, this approach is in line with the tenets of Industry 4.0, enabling intelligent buildings to reduce energy costs, enhance sustainability, and enhance operational efficiency.

3) Image-based data augmentation

Handcrafting fake images is a laborious and time-intensive task. Our collection has 2148 images (1528 fake

and 620 real), complemented by the Real and Fake Face (RFF) dataset [18], adding 960 fake and 1081 real images, making it 4188 images. Yet, this number is still too small to train deep learning models efficiently. Earlier work [19] has noted data augmentation as an important method of dataset enlargement and model generalization. In order to address sparsity in the data, we use two methods of augmentation: (1) Keras image preprocessing, which performs real-time adjustments such as shifting, shearing, zooming, and flipping, and (2) ImgAug [20], which generates a significantly larger dataset by performing diverse changes, including geometric shifts, color alteration, blurring, and blending. Fig. 7(a) displays a test image augmented by six different Keras transformations, while Fig. 7(b) displays 64 ImgAug transformations for an image.

4) Shallow-FakeFaceNet (SFFN) for facial forgery detection

To efficiently identify face manipulations, we designed Shallow-FakeFaceNet (SFFN), a light CNN model specifically for low-resolution images. In our experiments, we noticed that deeper models such as DenseNet and Xception [21] were not efficient on small images (64 by 64 and 128 by 128) since they are designed for large-scale, high-resolution datasets with large parameters and data demands. As Xception was first trained on 350 million images with 17,000 classes, it was not efficient for our dataset. We then introduced a shallow but efficient CNN for fake face detection, as presented in Fig. 8. SFFN employs L2 regularization (0.0001) and dropout layers (0.25) to prevent overfitting and enhance generalization, especially with the presence of few training samples. SFFNV1 (Fig. 8a), three variants were designed: consists of three convolutional layers (kernels: 3×3,3×3,1×1) with max pooling and repeated in six stages, with final dense layers of 3933 and 2 neurons. SFFNV1, however, had decreased detection accuracy in small images. For better performance, we implemented SFFNV2 (Fig. 8 b) with eight convolutional layers, all 3×3 except the last 1×1, and set the final dense layer size to 1024 and 2, resulting in improved classification on small images. To minimize further computational cost and training time, we implemented SFFNV3 (Fig. 8c), reinstating max pooling layers in SFFNV2 and modifying the sizes of the convolutional kernels to maximize efficiency. The performance of fake detection of all SFFN models is compared to that of other CNN-based classifiers based on preprocessed images in Section 5. We also compare our model with Adobe's Photoshopped Face Detector [67] to provide a benchmark comparison of performance among different fake detection architectures.

4 Experimentation and results

We have undertaken rigorous evaluations to assess and contrast the efficiency of our detection framework in two specific cases: GAN-generated and hand-crafted spoof face detection).

A) Handcrafted Facial Manipulation (HFM) Image Detection

1) Dataset characteristics

In order to detect facial forgeries, we employ three datasets. HFM consists of 1527 forged images manually created using Adobe Photoshop [1] and 621 authentic images downloaded from Google. The forged images include three levels of complexity and six types of modification, as shown in Figs. 1 and 2. We also add the Real and Fake Face (RFF) dataset [18] consisting of 960 fakes and 1081 reals collected from Kaggle. Compared to HFM with multi-face forgeries and complex manipulations, RFF consists of single-face, simpler manipulations, so HFM is a more challenging benchmark. To have a balanced set, we add 1177 real celebrity faces from the CelebFaces Attributes (CelebA)

(a)

(b)

Figure 7: Examples of image augmentation. (a) Keras transformations (shifting, shearing, zooming, flipping). (b) ImgAug enhancements (geometric,

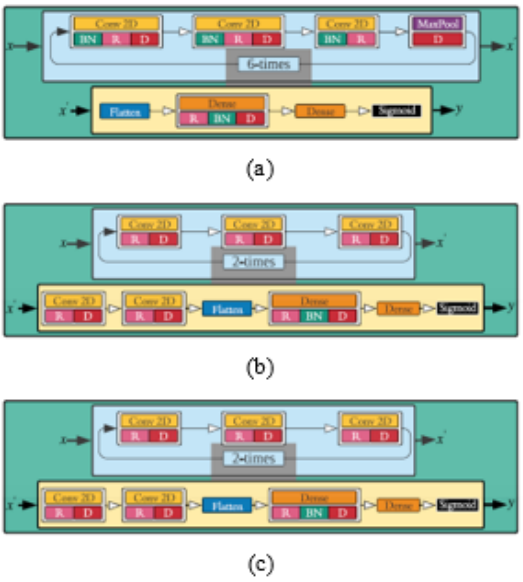


Figure 8: Architecture of Shallow-FakeFaceNet (SFFN) models. (a) SFFNV1 uses three convolutional layers with max pooling. (b) SFFNV2 has eight convolutional layers for better small-image detection. (c) SFFNV3 optimizes efficiency with

dataset [17]. These additional samples add diversity to real faces, making the model more robust. When we mix HFM and RFF together, we now have 4189 images (2487 false, 1702 true). The dataset gets bigger to 4645 samples (2911 false, 1734 true) after preprocessing (cropping and filtering) because images contain more than a single face. To achieve a well-balanced real-to-fake ratio, we supplement 1177 additional real images from CelebA, and the overall dataset is 5822 preprocessed images.

2) Experimental setup and baseline models

We contrast a number of CNN-based classifiers implemented in Keras [68] for detecting fake face images. Test models include Xception, Xception pre-trained on ImageNet (Xception-IN), ResNext, VGG19, Inception ResNet, MesoNet, NasNet, and our novel Shallow-FakeFaceNetV3 (SFFNV3). We also contrast findings with Adobe Photoshopped Face Detector [22], a benchmark. Since Xception has already demonstrated good detection performance on manipulated face datasets [11], we retrain it with ImageNet weights to study its performance on our dataset. The Adobe detection model is evaluated with pre-trained weights published by the authors [. For apples-to-apples comparisons, we standardize the input resolution to 128×128 pixels and upscale smaller images using the FSR method. Keras real-time data augmentation is also applied in training for improved model generalization. All the models are trained using the ADAM optimizer for 200 epochs, batch size 32, and learning rate 0.00005. The binary cross-entropy (BCE) loss function is used, as provided in Eq. (1):

Table 1: Performance comparison of CNN models on the HFM dataset for 128×128 and 256×256 image resolutions.

	128 by 128				*256 by 256*			
Model	Precision	Recall	F1-score	AUC	Precision	Recall	F1-score	AUC
SFFNV3 (Ours)	64.71	62.8	61.55	69.47	70.26	70.2	70.18	72.52
Xception	61.51	61.4	61.31	63.3	63.97	63.9	63.86	69.83
Adobe	51.99	50.7	46.79	48.47	50.99	50.7	47.79	47.17
Xception (IN)	65.22	66.2	66.19	69.61	66.27	67.0	65.86	66.74
Inception ResnetV2	62.41	63.4	64.39	66.28	63.8	62.8	63.8	66.04
MesoNet	24.0	50.0	34.33	50.0	25.0	50.0	33.33	51.0
ResNext	63.68	63.2	63.91	68.23	63.81	64.7	63.63	68.88

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log\left(p(y_i)\right) + (1-y_i) \cdot \log\left(1-p(y_i)\right)\right]$$

where y represents the ground truth label (0 for real, 1 for fake), and $p(y)$ is the model's predicted probability of an image being fake.

3) Performance on the HFM dataset

For analysis, we process 1000 test images (500 altered, 500 authentic) major performance metrics: Precision, F1-score, Recall, and AUROC [23]. AUROC, in particular, is used to gauge model performance, with greater values indicating better performance. Detection results are listed in Table 1, with the bold values representing the top-

performing models. Our findings indicate that the Adobe detector and the MesoNet perform poorly, with 50.00% and 47.47% AUROC detection rates, respectively, equivalent to random guess. In contrast, the Xception with ImageNet weights leads among the CNN-based models with a detection rate of 69.61% AUROC, followed by Inception ResNet (63.30%), ResNext (66.28%), and SFFNV3 (68.23%). While Xception performs well, our finding is that SFFNV3 is the best model overall with higher accuracy across different configurations. As illustrated in Figure 6, visualization of historical energy data plots patterns of energy consumption in terms of different meters (electricity, hot water, chilled water, and steam) in kilowatt-hours (kWh) hourly. This structured information facilitates better decision-making since building managers are able to identify inefficiencies, optimize usage, and implement effective energy-conserving strategies. Smart buildings are then able to adjust energy delivery in real-time, eventually maximizing efficiency and sustainability, based on a daily understanding of energy consumption.

4) Effect of input image sizes

We explore the impact of image resolution on classification performance, considering prior research [10,11] indicating GAN-synthesized images exhibit resolution-dependent differences in performance. To validate this, we compare models trained on 128×128 and 256×256 cropped faces. The results in Table 1 indicate most models show little improvement with higher image size. However, Xception and SFFNV3 are more F1-score and AUROC, in which Xception improves by 5.52% (AUROC), 2.52% (F1-score) and SFFNV3 improves by 3.15% (AUROC) and 8.64% (F1-score). The error distribution matrices for Table 1 are illustrated in Fig. 9.

5) Effect of super-resolution methods

As 8% of cropped faces in our dataset are of sizes less than 128×128 pixels, we compare two image upscaling techniques: NNU and FSR, as shown in Table 2. While NNU is cheap in terms of computation, it creates pixelation artifacts by duplicating pixels, degrading visual quality (Fig. 6). FSR, however, uses a GAN-based approach, preserving facial features during upscaling. Benchmarking the highest-performing two models (SFFNV3 and Xception-IN) on images that benefited from FSR, we can see that Xception-IN provides a small AUROC increase (+3.4%), whereas **SFFNV3 improves significantly from 57.98% to 72.52% AUROC, demonstrating FSR to be better than NNU. These facts suggest that better super-resolution methods enhance model accuracy, particularly with regard to small face images.

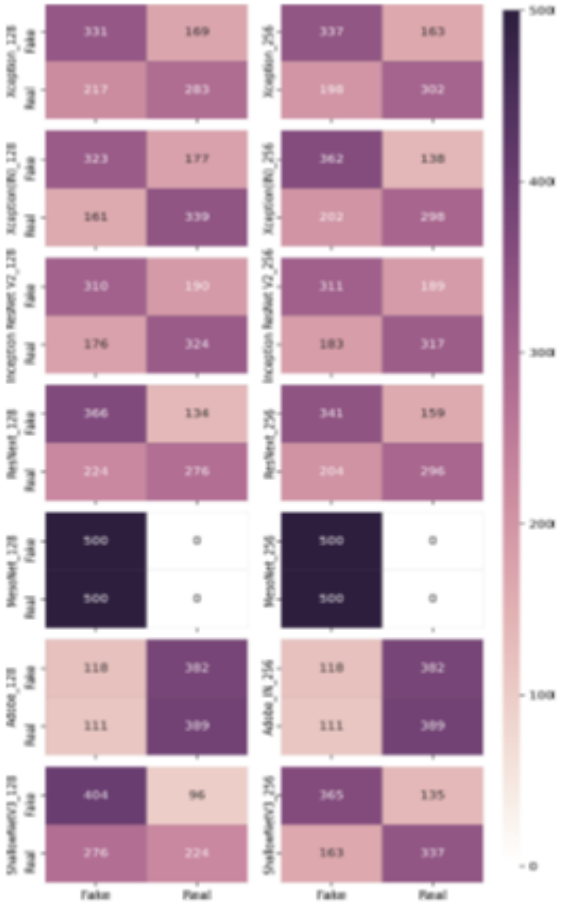


Figure 9: Confusion matrices for different models on the HFM dataset, illustrating classification performance for real and fake images across various training configurations.

Table 2: Performance comparison of FSR and NNU upscaling

	Nearest Neighbor Upscaling				Facial Super-Resolution			
Model	Precision	Recall	F1-score	AUC	Precision	Recall	F1-score	AUC
SFFNV3 (Ours)	56.77	56.1	54.99	57.98	70.26	70.2	70.18	72.52
Xception (IN)	65.52	65.2	65.02	66.21	66.22	66.2	66.19	69.61

Table 3: Performance comparison of different augmentation methods

	Without Augmentation (%)				Keras (%)				ImgAug (%)			
Model	Precision	Recall	F1-score	AUC	Precision	Recall	F1-score	AUC	Precision	Recall	F1-score	AUC
SFFNV3 (Ours)	54.99	54.7	54.84	54.84	70.26	70.2	70.18	72.52	55.67	55.1	53.58	55.67
Xception (IN)	63.9	63.2	62.73	58.85	66.22	66.2	66.19	69.61	60.49	60.3	60.12	57.36

Figure 10: Comparative visuals of natural (CelebA) and GAN-produced (PGGAN) images at varying sizes (a) show real celebrity images from CelebA, while (b) present synthetic faces from PGGAN.

Figure 11: Class Activation Maps (CAM) for fake face detection.

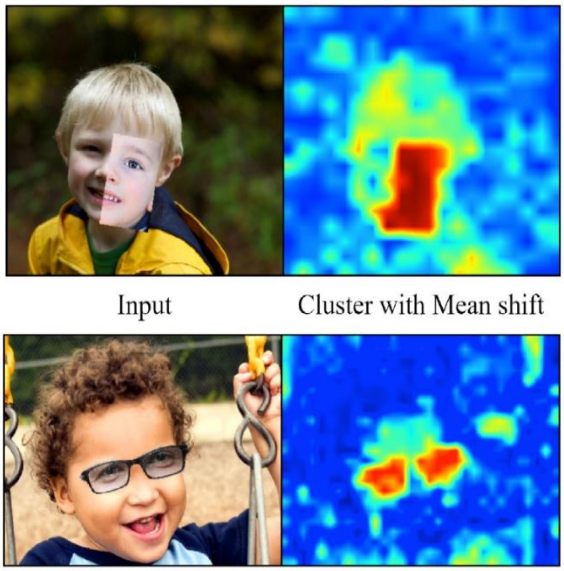


Figure 12: Comparison of image splice detection and our proposed method.

6) Effect of data augmentation

To improve detection performance, we explore two methods of augmentation: Keras real-time augmentation and ImgAug [65]. The Keras library performs six transformations: random shifting, shearing, zooming, and flipping (Fig. 7a), and ImgAug introduces 64 types of augmentation, such as geometric transformations, blurring, color changes, and blending (Fig. 7b). For the comparison of efficacy of augmentations, we train high-performance two models (Xception-IN and SFFNV3) using FSR-augmented images and indicate performance across varied augmentation strategies (Table 3). Performance examination reveals that Keras augmentation increases
detection accuracy while Xception-IN increases AUROC by 10.76% and SFFNV3 by 17.68%. It is in opposition to ImgAug having a negative impact on performance where F1-score decreases by Xception-IN (62.73% → 60.12%) and SFFNV3 (54.03% → 53.58%).

B) GAN-Generated Image Detection

Fake faces created by GAN are detected through the comprehensive detection framework illustrated in Fig. 4. Unlike handcrafted fakes, high-res images1024×1024 are rather reduced to 64×64 and 128×128 for training purposes without making use of any sort of upscaling. In terms of improving detection, ImageNet-pretrained weights are used for model initialization with minimal requirements for large-labeled datasets.

1) Dataset descriptions

We use two datasets to detect GAN-generated images: CelebA [17], a collection of 200K real celebrity faces, and PGGAN [8], a collection of 100K fake faces. Some example comparisons between real faces and GAN-generated faces are presented in Figs. 10(a) and 10(b). The PGGAN images are labeled as "fake", with CelebA being real face data. These images are passed through CNN models, as illustrated in Fig. 4.

2) Performance on GAN-Generated faces

We train the models on 200K images, validating on 20% of the dataset and testing on 18K samples. The performance is checked at multiple resolutions (64×64, 128×128, 256×256, and 1024×1024) using performance matrices like Recall, Precision, F1-score, and AUROC metrics.Results indicate that Shallow-FakeFaceNet (SFFN) outperforms all the models even at the lowest resolution (64×64), with an ensemble of SFFNV1 and SFFNV3 producing the best accuracy (93.99%–99.99. Surprisingly, deeper networks like Xception and NASNet struggle to detect images at low resolutions, whereas SFFN models are resilient with robust detection capabilities for all sizes. These findings confirm that shallow models are better for detection in GAN-generated images, particularly for low-resolution inputs.

3) HFM detection model analysis

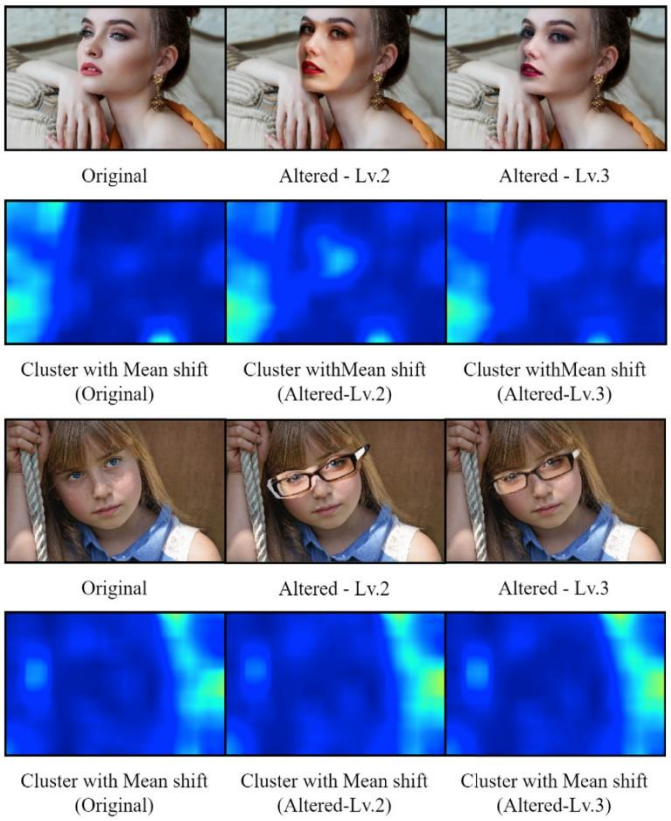


Figure 13: Forged images and splice detection results

We also test our HFM detection model with Class Activation Maps (CAM), which denote areas accountable for classification. Grad-CAM analysis in Fig. 11 demonstrates that Xception-IN focuses at the center of the face, while SFFNV3 pays attention to the manipulated landmarks and boosts fake detection. Fig. 11 demonstrates that HFM images contain minimal forgeries in the form of altered noses and mouths, and SFFNV3 can catch them well. This suggests that facial landmark focus enhances fake detection accuracy. We compare our approach with Huh et al.'s forgery detection algorithm based on image metadata [16]. From Figs. 12 and 13, their method identifies forged regions when metadata is present but fails to do so in the absence of metadata. Our model, however, identifies over 70% of the forged regions, as certified in Tables 1–3. This proves that our approach is better suited for metadata-independent fake face detection, particularly on HFM images.

5 Discussion and limitation

Our shallow-FakeFaceNet (SFFN) outperforms deep CNNs in detecting handcrafted fake faces, especially on low-res images. Contrary to the belief that deeper networks are better, our findings agree with He et al. [24], which shows that shallow models retain important

features, whereas deep models like Xception and InceptionResNetV2 fail due to excessive pooling. One of the main issues in HFM dataset detection is its limited dataset size [19]. To this end, we introduce the HFM dataset of 1527 simulated and 621 real images, which we are releasing publicly upon paper acceptance. But our approach is not flawless. The data do not have diversified medical conditions (e.g., Bell's palsy, vitiligo, facial burns) and therefore the detection may be biased. The tiny faces on images of multiple people are usually removed by presetting filtering thresholds and thus provide false negatives. As a countermeasure, we propose utilizing multiple facial detectors in combination and ensemble decision strategies (e.g., majority voting), which would be useful to improve overall detection rate.

While the proposed approach demonstrates significant improvements over existing models, several limitations remain that warrant further investigation. First, the detection performance on handcrafted fake faces—though notably better than prior models—remains moderate (AUROC ~72.52%), indicating the intrinsic difficulty of such subtle manipulations. Second, the HFM dataset, although novel and diverse in editing types, is still relatively small and lacks edge-case conditions such as facial asymmetry, scars, or medical anomalies, which may affect generalizability. Third, the fixed-size filtering strategy in the face preprocessing step may inadvertently remove valid small or occluded faces in multi-person images. Furthermore, certain augmentation methods (e.g., ImgAug) were found to degrade model performance, possibly due to their distortion of fine-grained facial cues. Finally, while SFFN is designed to be computationally efficient, a comparative analysis of runtime performance was not conducted, and real-world deployment scenarios such as adversarial robustness or social media-scale evaluations were beyond the scope of this study. These limitations present important directions for future work, including dataset expansion, ensemble-based preprocessing, and broader validation across operational contexts.

6 Conclusion

Manipulated and machine-generated fake photography present serious dangers such as identity falsification and slander. To combat this, we introduce Shallow-FakeFaceNet (SFFN) and the Handcrafted Facial Manipulation (HFM) dataset to enhance the identification of fake images. The HFM dataset provides a huge set of handcrafted fake images with different levels of editing difficulty and facial changes, filling the gap in the existing datasets. Our pipeline-based detection system, which utilizes super-resolution and data augmentation, enables SFFN to be 72.52% AUROC with less than 2500 synthesized false images and reach 93.99% accuracy on GAN-generated images, particularly low-resolution cases. The proposed model has potential applications in social network pre-check systems (Twitter, Facebook, Instagram) and verification media tools in an effort to mark manipulated media content. In the future, we plan to expand the HFM dataset with more diverse facial conditions and examine other editing tools such as Pixlr, GIMP, and Photoscape. Another direction is employing GAN-based augmentation to improve training data or using transfer learning from DeepFakes and FaceSwap detection models to enhance performance in handcrafted fake image detection.

References

[1] Adobe (2020) *Adobe Photoshop: Best photo, image, and design editing software*. Available at: https://www.adobe.com/products/photoshop.html (Accessed: 22 April 2024).

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) 'Generative adversarial nets', *Advances in Neural Information Processing Systems*, pp. 2672–2680. DOI: 10.48550/arXiv.1406.2661

[3] Venkateswarulu, S. and Srinagesh, A., 2024. DeepExplain: Enhancing DeepFake Detection Through Transparent and Explainable AI model. *Informatica*, *48*(8). DOI: 10.31449/inf.v48i8.5792

[4] Huh, M., Liu, A., Owens, A. and Efros, A.A. (2018) 'Fighting fake news: Image splice detection via learned self-consistency', *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 101–117.

[5] Karras, T., 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv preprint arXiv:1710.10196*. DOI: 10.48550/arXiv.1710.10196

[6] Farid, H., 2009. Exposing digital forgeries from JPEG ghosts. *IEEE transactions on information forensics and security*, *4*(1), pp.154-160. DOI: 10.1109/TIFS.2008.2012215

[7] Zhou, P., Han, X., Morariu, V.I. and Davis, L.S., 2018. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1053-1061).

[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. *Advances in neural information processing systems*, *27*. DOI: 10.48550/arXiv.1406.2661

[9] Tariq, S., Lee, S., Kim, H., Shin, Y. and Woo, S.S., 2019, April. Gan is a friend or foe? a framework to detect various fake face images. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing* (pp. 1296-1303).

[10] Zhou, P., Han, X., Morariu, V.I. and Davis, L.S., 2018. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1053-1061).

[11] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Nießner, M. (2018) 'FaceForensics: A large-scale video dataset for forgery detection in

human faces', *arXiv preprint*, arXiv:1803.09179. DOI: 10.48550/arXiv.1803.09179

[12] Li, Y., Yang, X., Sun, P., Qi, H. and Lyu, S. (2019) 'Celeb-DF: A new dataset for DeepFake forensics', *arXiv preprint*, arXiv:1909.12962. DOI: 10.48550/arXiv.1909.12962

[13] Simonyan, K. and Zisserman, A. (2014) 'Very deep convolutional networks for large-scale image recognition', *arXiv preprint*, arXiv:1409.1556, pp. 1–14.

[14] Murali, S., Chittapur, G.B., Anami, B.S. *et al.* (2013) 'Comparison and analysis of photo image forgery detection techniques', *arXiv preprint*, arXiv:1302.3119. DOI: : 10.48550/arXiv.1302.3119

[15] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) 'Rethinking the inception architecture for computer vision', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826.DOI: 10.1109/CVPR.2016.308

[16] de Paz Centeno, I. (2016) *MTCNN face detection implementation for TensorFlow*. Available at: https://github.com/ipazc/mtcnn (Accessed: 31 December 2020).

[17] Kim, D., Kim, M., Kwon, G. and Kim, D.-S. (2019) 'Progressive face super-resolution via attention to facial landmark', *arXiv preprint*, arXiv:1908.08239. DOI: 10.48550/arXiv.1908.08239

[18] Computational Intelligence and Photography Lab Yonsei University (2019) *Real and fake face detection*. Available at: https://www.kaggle.com/ciplab/real-and-fake-face-detection (Accessed: 31 December 2020).

[19] Hussain, Z., Gimenez, F., Yi, D. and Rubin, D. (2017) 'Differential data augmentation techniques for medical imaging classification tasks', *AMIA Annual Symposium Proceedings*, *American Medical Informatics Association*, p. 979.

[20] Jung, A.B., Wada, K., Crall, J., Tanaka, S., Graving, J., Reinders, C., Yadav, S., Banerjee, J., Vecsei, G., Kraft, A., Rui, Z., Borovec, J., Vallentin, C., Zhydenko, S., Pfeiffer, K., Cook, B., Fernández, I., De Rainville, F.C.-M., Weng, C.-H., Ayala-Acevedo, A., Meudec, R., Laporte, M. *et al.* (2020) *ImgAug*. Available at: https://github.com/aleju/imgaug (Accessed: 31 December 2020).

[21] Chollet, F. (2017) 'Xception: Deep learning with depthwise separable convolutions', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258. DOI: 10.1109/CVPR.2017.195

[22] Venkateswarulu, S. and Srinagesh, A., 2024. DeepExplain: Enhancing DeepFake Detection Through Transparent and Explainable AI model. *Informatica*, *48*(8). DOI: 10.31449/inf.v48i8.5792

[23] Chu, H. (2017) *GitHub: AUROC (Area Under the Receiver Operating Characteristic)*. Available at: https://github.com/hyoungseokchu/AUROC (Accessed: 31 December 2020).

[24] He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep residual learning for image recognition', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. DOI: 10.1109/CVPR.2016.90