Face detection and recognition under real-world scenarios - dealing with deepfake incidents and malicious data distortions

Ewelina Bartuzi-Trokielewicz[1], **Alicja Martinek**[1,2], **Adrian Kordas**[1]
{ewelina.bartuzi-trokielewicz, alicja.martinek, adrian.kordas}@nask.pl

[1]NASK National Research Institute
ul. Kolska 12
01-045 Warsaw, Poland

[2]AGH University of Kraków
al. Mickiewicza 30
30-059 Kraków, Poland

Abstract

The growing use of deepfake technology and synthetic facial images presents significant challenges to biometric verification systems, particularly when identity-specific features are obscured by textured masks, halftones, and watermarks. This paper evaluates six face detection methods and four state-of-the-art facial recognition engines on two types of data: unaltered facial images and synthetic faces masked with diverse textures. We proposed a custom Python library for generating image distortions and obfuscations to simulate textural noise and assessed their impact on face verification. Results revealed vulnerabilities of current systems to textured masks and watermarks, highlighting challenges posed by occluded features in manipulated media. Additionally, we introduced a novel face detection method utilizing this augmentation library, achieving up to 99% detection accuracy under noisy conditions. Our analysis identified statistically significant obfuscation techniques that affect model performance the most, providing insights for improving robustness against real-world distortions.

1. Introduction

The increasing deployment of facial recognition technology across various digital platforms, including social media, e-commerce, and security applications, highlights both the potential and the challenges associated with its use in unregulated digital spaces. As recently reported by BBC News, companies like Meta are implementing facial recognition to counteract fraudulent advertisements featuring celebrity endorsements [12]. Based on findings presented in the report [11], the use of deepfake technology has escalated significantly, with incidents involving video deepfakes tripling and those using voice deepfakes growing eightfold from 2022 to 2023. Projections for 2024 indicate a continued rise, with an estimated 50-60% increase in deepfake incidents, potentially reaching between 140,000 and 150,000 cases worldwide. In this context, the need for reliable verification is evident, especially given the rising complexity of manipulative media, where celebrity images are digitally altered or deepfake-generated to promote false narratives and exploit public trust. Despite advancements in biometric verification, contemporary systems face significant limitations when confronted with synthetic manipulations and obfuscations that mimic identity-specific features in digital media.

The challenge is further exacerbated by the advent of sophisticated noise techniques, such as textured masks (known as halftones) and watermarks, which can obscure distinct facial attributes. These obfuscations, often used in deepfake-enhanced advertisements, diminish the accuracy of traditional facial recognition systems. Textured masks, including halftone and watermark effects, have proven effective in concealing key identity cues, thereby impairing the ability of these systems to reliably differentiate between genuine and synthetic content. As facial recognition systems are increasingly relied upon for authentication and identity verification in noisy media environments, understanding and mitigating the impact of such obfuscations is critical.

This paper addresses the limitations by evaluating the performance of six face detection methods and four state-of-the-art facial recognition engines when confronted with noisy, obfuscated facial images. Our evaluation leverages two types of data: one with unaltered facial images, and another featuring synthetically generated faces obscured with various textured patterns. To improve resilience under these challenging conditions, we introduce a novel data augmentation method that simulates textural noise, enhancing the system's capacity to accurately verify faces even when identity-specific features are obscured. Although this study focuses primarily on assessing the impact of these obfuscations, it also proposes a novel face detection method integrated with the augmentation library, achieving a significant improvement in detection under noisy conditions. Furthermore, this augmentation library can be applied in fine-tuning and developing facial recognition systems, providing a foundation for improving robustness and reducing error rates in noise-rich environments.

This study makes two primary contributions. First, it identifies the vulnerabilities of existing face detection and recognition systems in handling synthetic obfuscations, such as textured masks and watermarks. Second, it demonstrates the potential of targeted data augmentation techniques to enhance detection accuracy and improve the overall robustness of facial recognition systems. These findings highlight the need for evolving biometric technologies to adapt to the complexities of modern digital manipulations, aligning with ongoing industry efforts to bolster digital security and trust in biometric verification systems.

While the topic of attacks on facial recognition systems has been extensively studied, and deepfake detectors embedded in social media platforms are under constant development, the ability of synthetic manipulations to bypass these systems underscores a significant gap in the current research. This study addresses this gap by providing a detailed analysis of the impact of obfuscations on facial detection and recognition, offering a foundation for the development of more resilient biometric systems.

2. Related work

Research on facial identity verification has progressed significantly, especially with the advent of deep learning techniques that have boosted accuracy and reliability in face recognition systems. However, the presence of occlusions, intentional noise, and synthetic alterations still poses critical challenges, particularly in scenarios where identity-specific features are masked or distorted.

Face-based identity verification methods

Face-based identity verification has been widely adopted due to its non-intrusive nature and high accuracy in controlled environments. Traditional methods rely on holistic or appearance-based approaches, which treat the face region as a whole and map it into a lower-dimensional subspace for recognition. Prominent examples include Eigenfaces [4, 34], which use Principal Component Analysis (PCA) for dimensionality reduction, and other approaches based on linear subspaces [16] and manifold learning [26, 33]. Techniques leveraging sparse representations have also emerged, with algorithms such as SRC [37] demonstrating robustness to occlusions and noise. Then, local feature-based methods gained popularity, focusing on hand-crafted descriptors that analyze specific facial regions. Techniques such as Gabor features [18, 25, 29] and Local Binary Patterns (LBP) [1] were particularly effective in capturing texture details and invariance to lighting conditions. Further advancements introduced Binarized Statistical Image Features (BSIF) [15], which extend LBP by learning data-driven filters for feature extraction, and other LBP variants [23, 32].

Recent advances in convolutional neural networks (CNNs) and deep learning have set new standards for facial recognition. Techniques based on deep neural networks have introduced powerful embeddings that map face images into feature space, making them effective for both verification and recognition tasks. However, despite their impressive performance, these methods are often tested on datasets where facial attributes are fully visible, leading to decreased accuracy when applied to noisy or masked images. Early breakthroughs like DeepFace [7, 31] utilized deep architectures to map face images into compact feature spaces, achieving high accuracy in controlled scenarios. FaceNet [27] introduced the triplet loss function, optimizing embeddings to minimize intra-class variance and maximize inter-class separability, setting a foundation for modern verification techniques. Further innovations, such as ArcFace [10], utilized additive angular margin loss to improve the discriminative power of embeddings, making them robust to variations in pose, illumination, and expression. Similarly, SphereFace [19] and CosFace [35] enhanced performance by introducing angular and cosine margin-based losses, improving feature separability in challenging conditions. More recent methods, like MagFace [22], focused on embedding quality by dynamically scaling features based on their confidence, making systems adaptable to occlusions and low-quality inputs

Data augmentation techniques for improved model performance

Data augmentation has emerged as a critical strategy to improve model resilience, particularly in the presence of distortions. Techniques such as random cropping, rotation, and synthetic noise addition have been shown to improve model generalization by exposing it to a wider variety of conditions during training. Studies have demonstrated that augmenting datasets with occlusions, blurring, and artificial lighting variations can enhance model robustness in real-world settings. Early techniques, such as flipping, cropping, and rotation, proved effective in boosting performance, as demonstrated by Parkhi et al. [24]. Masi et al. [21] extended this by introducing pose-based augmentations to handle head orientation changes. More advanced methods, like Random Erasing by Zhong et al. [42], addressed occlusions by masking parts of the image, while Shi and Jain [30] employed uncertainty modeling to handle noisy inputs. The survey [36] extensively reviews existing techniques for face data augmentation, categorizing them into distinct transformation types and methods to improve model robustness under diverse real-world conditions. Geometric transformations, such as translation, rotation, scaling, and flipping, simulate pose and spatial variations, while photometric transformations adjust brightness, contrast, noise, and sharpness to handle diverse lighting and imaging conditions. Facial attribute transformations use methods like 3D Morphable Models and GANs to generate variations in pose, expressions, and age, maintaining identity features. Facial component transformations, including makeup transfer, accessory manipulations, and texture alterations, simulate occlusions and aesthetic changes. Lastly, domain-specific augmentations, such as low-light simulations and synthetic data blending, enhance resilience in challenging environments, making models more robust to real-world conditions.

Impact of noise and obfuscation on face recognition accuracy

The influence of noise and occlusions on facial recognition accuracy has been widely studied, with findings indicating that even minor obstructions can significantly reduce model performance. Research has shown that noise patterns like Gaussian blur, partial occlusions, and textured overlays can disrupt feature extraction and impact model accuracy. In particular, masks that cover key facial landmarks - such as the mouth and eyes - impair verification models, as these features contribute heavily to identity representation. Recent studies have examined specific cases, such as the impact of COVID-19 face masks, but there is a lack of research focusing on synthetic obfuscations common in fake advertisements and deepfake videos. Our work extends the current body of research by systematically investigating the effects of various textured masks and watermark noise on face recognition accuracy, which are frequently encountered in fraudulent media.

In summary, this article builds upon previous research in face-based identity verification, data augmentation for noise resilience, and the impact of obfuscation on recognition accuracy. By focusing on the challenges posed by synthetic noise, we aim to provide insights into how these distortions affect facial recognition systems.

3. Masking techniques in fake materials

Fake materials, particularly advertisements, present a growing challenge for facial recognition and biometric technologies, which are now facing digital manipulations beyond standard image or audio analysis. This phenomenon is intensifying with the increased accessibility of advanced editing tools that enable the creation of highly realistic fraudulent content. These materials still exhibit numerous detectable errors that can be identified through technical analysis, including flaws resulting from the imperfections in deepfake generation techniques, as well as logical inconsistencies and various social engineering tactics. However, fraudsters use deliberate information masking techniques to hide these distortions, thereby complicating identification and recognition.

Common audiovisual errors in fake materials include incongruities in lip-sync, blurring around the mouth, unnatural facial shadows, inconsistencies in body elements, language changes, mismatches between the emotional tone of the voice and gestures, audio interruptions, digital artifacts, robotic-sounding voices, a reading-like speech tempo, and unnatural pitch changes. While these errors are often visible to the human eye, cybercriminals employ various techniques to obscure them, enhancing the credibility of fake materials.

Information masking in such materials is a deliberate tactic aimed at complicating their analysis and detection. Common masking techniques include image blurring, and texture-based interference in video, such as halftones and watermarks. These manipulations obscure key facial features, significantly challenging the accuracy of biometric systems in face detection and verification. Examples of these obfuscation methods are shown in Figure 1.

4. Methodology

4.1. Albutortion

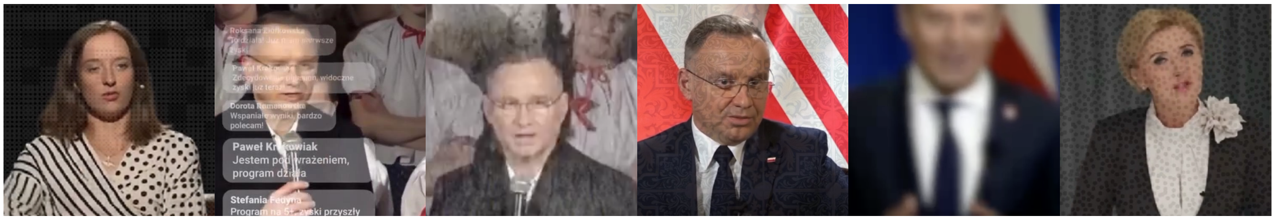
To simulate real-world obfuscations seen in fake audiovisual materials, we developed a custom image augmentation library - Albutortion[1]. This library is designed to add various types of noise and distortions that are commonly observed in manipulated media - in deepfake materials. This library extends the capabilities of the Albumentations framework [5] by incorporating additional 25 functionalities tailored specifically to the challenges of deepfake detection. The library introduces various types of noise, distortions, manipulations and augmentation techniques. Each of them can be applied with adjustable parameters such as intensity or density, enabling precise customization to simulate diverse real-world conditions.

The core functionalities of Albutortion, can be divided into the following categories:

- **texture overlays** - adds elements like dots, stripes, grids, checkerboards, and textures from 50 predefined patterns, simulating distortions, tampering effects, and watermarks,
- **environment simulation** - replicates conditions like rain, dirt, and reflections, mimicking weather-related and temporal distortions in visual data,
- **compression and color adjustment** - simulates image degradation from lossy compression and reduced color depth, testing performance under low-quality media conditions,
- **focus and sharpness modification** - introduces blur to mimic out-of-focus conditions and sharpness adjustments to simulate overprocessed visuals, testing robustness to clarity variations.

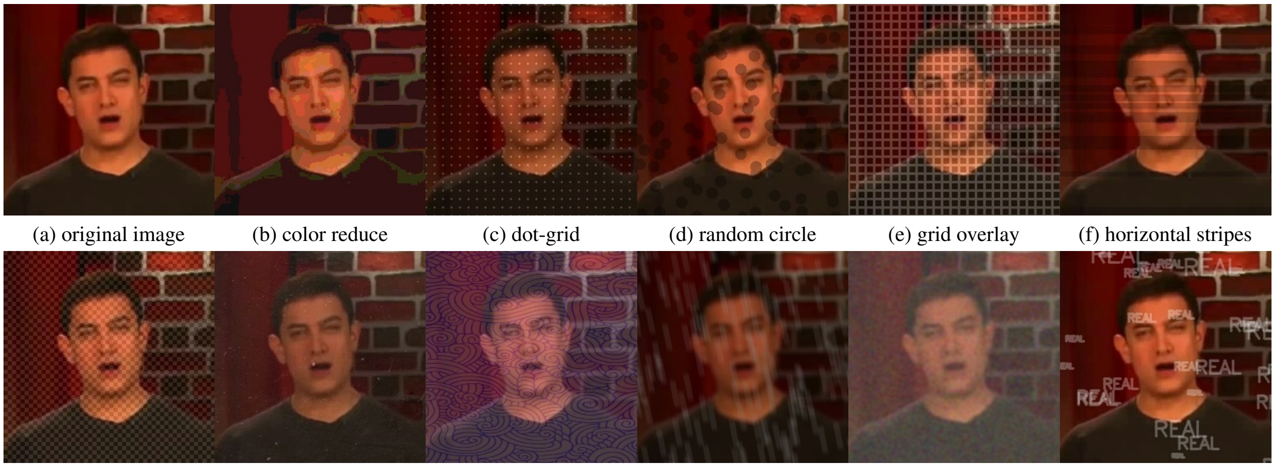
Examples of the described augmentations with selected parameters can be seen in Figure 2.

[1] https://github.com/NASK-Biometria/albutortion


Paweł Krakowiak
Jestem pod wrażeniem, program dzisia
Dorota
Wspaniałe wyniki, bardzo polecam!
Stefania

(a) random mesh fractal (b) social media (c) rain (d) fancy texture (e) blur (f) circles

Figure 1. Examples of textures added in deepfake materials found on social media platforms.



(a) original image (b) color reduce (c) dot-grid (d) random circle (e) grid overlay (f) horizontal stripes

(g) chess board (h) aged photo texture (i) oriental pattern texture (j) heavy rain (k) vintage effect (l) watermark effect

Figure 2. Sample image from Celeb-DF-v2 dataset with various distortions generated using the Albutortion library.

4.2. Experimental dataset

To evaluate the impact of different obfuscation techniques on facial detection and recognition accuracy, we utilized database composed of the following datasets: **Celeb-DV-v2** [17] - a large-scale dataset containing 590 original YouTube videos featuring 59 individuals of diverse ages, ethnicities, and genders. It also includes 5639 high-quality deepfake videos, designed to closely replicate the realism of manipulated media commonly found online. **Incident-Based Dataset** - gathered from real-life incident reports involving manipulated media by deepfake technology, from 20 public figures. It includes two categories:

- deepfake materials, collected from online sources or documented incident reports, includes 43 videos of clear deepfake content (synthetic), and 50 videos of deepfake manipulation with additional obfuscations (synthetic-masked)
- reference real materials - 54 videos of the same individuals from trusted sources as a baseline for comparison.

Both datasets were processed with the same methodology. In the first step, for each video, an image per second (with respect to FPS) was extracted and saved. These images were later subjects to face detection and face pose estimation algorithms. The output of face pose calculations, namely yaw, pitch and roll, allowed to select images representing most frontal faces. The number of data samples processed from each dataset is presented in Table 1.

Dataset	Type	Videos	Photos
Celeb-DF-v2	authentic	590	1686
	synthetic	5639	16901
Incident-Based	authentic	54	233
	synthetic	43	128
	synthetic-masked	50	149

Table 1. Sample counts across experimental datasets.

For the evaluation of algorithms, we applied 138 modifications to the images, both real and deepfake, using the Albutortion library. These modifications included varying parameters such as frequencies, texture sizes, and opacity levels ranging from 0.1 to 0.3 to simulate different levels of obfuscation and distortions. However, for the fake-masked category, no additional modifications were implemented, as these images already contain natural noise and obfuscations derived from real-world incidents.

4.3. Face detection

We evaluated the performance of six widely used face detection algorithms on both unaltered and obfuscated experimental data. The detectors included MediaPipe [3], RetinaFace [9], SSD [39], OpenCV's DNN module [38], Haar cascades [2], and MTCNN [41]. These methods were tested under varying conditions to assess their robustness in detecting faces with and without obfuscations, such as textured masks and watermarks.

Experimental data type	**Authentic**							**Synthetic**						
	MediaPipe	SSD	DNN	Haar	MTCNN	Retina	ours	MediaPipe	SSD	DNN	Haar	MTCNN	Retina	ours
Incident-Based Dataset	24.66	79.37	97.76	99.10	**100.0**	**100.0**	99.10	21.09	82.03	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Incident-Based - synthetic-masked subset	-	-	-	-	-	-	-	9.34	72.89	93.45	92.52	97.16	**99.07**	**99.07**
Incident-Based Dataset with Albutortion-based distortions	15.17	61.81	75.74	92.93	95.80	85.19	**98.83**	10.17	58.97	69.79	91.00	91.56	85.20	**99.47**
Detection accuracy for the most challenging obfuscations in detection process														
Checkerboard - 25, 0.3	0.89	0.0	4.93	70.85	21.52	17.93	**96.86**	0.0	0.0	2.34	60.93	10.15	23.43	**99.21**
GridOverlay - 20, 5, 0.2	4.03	15.69	14.79	47.08	77.13	10.31	**96.86**	0.0	6.25	0.78	39.06	42.18	17.18	**100.0**
WindowTexture (all)	15.93	61.26	74.96	**97.51**	96.97	81.09	97.14	12.12	57.90	72.95	96.0	95.43	81.51	**99.37**

Table 2. Face detection accuracy presented as percentages of successful detections for authentic and synthetic samples across all models. Best performing models are printed in bold.

The results on the original RetinaFace model were promising, demonstrating high detection accuracy in the original dataset, yet a noticeable drop in performance for the augmented one. Based on results of all models it was decided to fine-tune RetinaFace model. Our augmentation library served as a source for creation of a new training dataset. Using the same augmentation pipeline as in the Incident-Based Dataset, Celeb-DFv2 [17] with selected frames was utilized. The final dataset consisted of 189,450 real and fake images. To generate PASCAL-VOC labels for the new training set, the MTCNN face detection model was applied to the original images, and the coordinates of the bounding boxes were transferred to the corresponding fake images. A MobileNet0.25 model [14] pretrained on the ImageNet [8] dataset was used as the backbone. The fine-tuning process was scheduled for 50 epochs, including 5 epochs with a warm-up phase. As a result, we observed a significant improvement in RetinaFace performance in detection tasks involving augmented images. For the most challenging obfuscation methods, an average improvement of 61.21% was achieved. Furthermore, we noticed that the fine-tuned model exhibited better stability in its results, showing reduced susceptibility to specific augmentation methods compared to its original counterpart. The tests, conducted with various face detection methods, were performed on a separate dataset derived from real-world incidents and processed using our obfuscation algorithms. The confidence threshold for detections was set to 0.6.

4.4. Face verification

For face verification, we employed four top-performing, open-source CNN matchers: ArcFace [10], FaceNet [28], MagFace [22], and VGGFace2 [20]. All models generate feature embeddings for face images, enabling identity verification by measuring the similarity between these embeddings using cosine distance. Each model outputs high-dimensional embeddings: 512-dimensional features for ArcFace, FaceNet, and MagFace, and 2048-dimensional features for VGGFace2. Input images are prealigned and resized to fixed dimensions (e.g., 112x112 for ArcFace and MagFace, and 224x224 for VGGFace2). The models were trained on large-scale datasets, including MS1M [13], CASIA-WebFace [40] and VGGFace2 [6], comprising millions of face images, ensuring robust feature representation for verification tasks. Each model was evaluated on both real and deepfake data, with and without masking, to assess robustness against textured masks, watermarks, and other synthetic noise introduced during the augmentation process. Testing under various conditions allowed analysis of their ability to handle obfuscations and maintain accurate face verification in complex real-world scenarios.

5. Experimental results

5.1. Impact of noise on face detection

Detection accuracy was evaluated based on the ability of algorithms to correctly localize faces across datasets and different obfuscation levels.

Table 2 presents face detection accuracy across various models evaluated on different subsets of the dataset.

Across authentic samples, traditional models like RetinaNet, Haar, and MTCNN consistently achieved 100% detection rates, demonstrating robustness to clear, unaltered images. Synthetic samples, even without additional distortions, posed a challenge for most models, except for DNN, Haar, MTCNN, RetinaNet, and OURS, which consistently maintained 100% accuracy. Masked synthetic samples presented one of the most significant challenges. Detection rates for MediaPipe and SSD dropped to 9.40% and 76.51%, respectively. However, Retina and OURS outperformed most models, achieving 99.07%, indicating its resilience to more complex noise and masking scenarios.

The Albutortion-based distortions, particularly those mimicking environmental factors (e.g., rain, focus drift), further tested the models' resilience. Here (third row), OURS demonstrated robustness, achieving nearly perfect detection rates — 98.83% on authentic data and 99.47% on synthetic ones. In contrast, models such as SSD and MediaPipe performed poorly, with detection rates dropping to 61.81% and

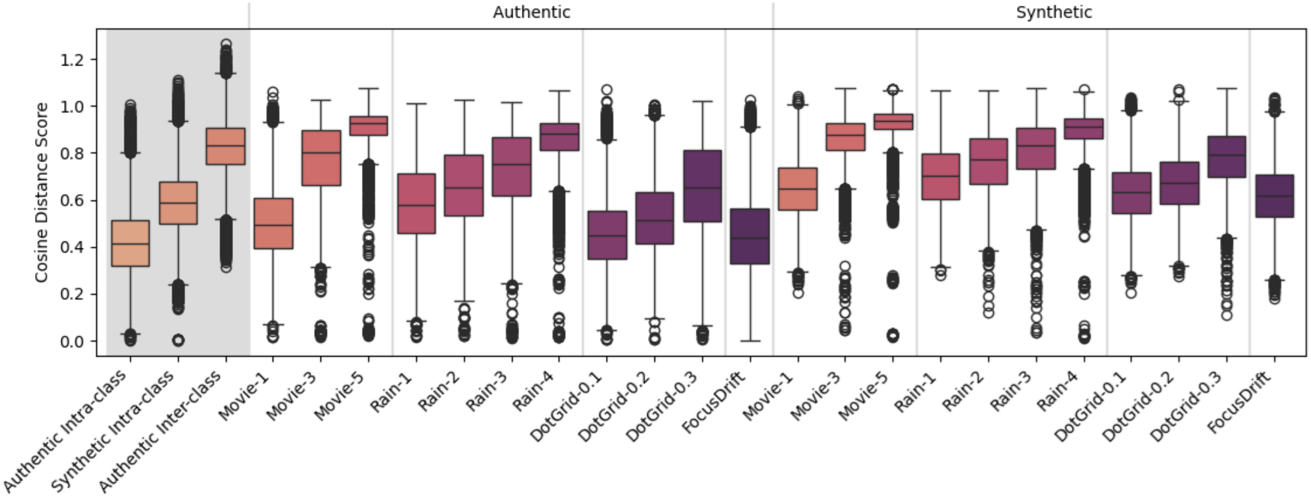


Figure 3. Impact of noise on face verification for Celeb-DF-v2 dataset using ArcFace model.

15.17% on authentic samples, and on distorted data 10.17% and 58.97% for synthetic data, respectively. On the other hand, MTCNN and Haar Cascade handled the distortions relatively well, achieving detection rates of 95.80% and 92.93% on authentic data, and 91.56% and 91.00% on synthetic. The analysis of most frequent detection failures highlights that certain types of obfuscations pose significant challenges for face detection systems. Specifically, Checkerboard with width of 25 px, GridOverlay, and WindowTexture distortions resulted in the most substantial drops in detection accuracy across models.

Checkerboard-25 involves a checkerboard pattern with a grid width of 25 pixels applied to the image with an opacity over 0.2. This obfuscation heavily impacts detection, with models like MediaPipe achieving as low as 0.00- 0.89% accuracy. Traditional models like SSD and Haar also struggled, while OURS maintained a high detection rate of 97.14 - 99.21%, demonstrating strong robustness. GridOverlay-20,5,0.2 is a grid overlay (grid cell size 20 px, line thickness 5 px) with an opacity of 0.2 and applied to the images. This obfuscation significantly hindered detection for MediaPipe (0.00- 4.03%) and SSD (6-25-15.69%). Despite the difficulty, OURS achieved almost perfect detection rate of 96.86 - 100%, showcasing its ability to handle this specific type of distortion effectively. WindowTexture refers to various types of synthetic textures applied to the image, mimicking distortions such as reflections or occlusions caused by windows. Detection rates ranged widely, with MediaPipe performing poorly at 12.12 - 15.93% and MTCNN achieving 95.43 - 97.51%. OURS showed high resilience, achieving a detection rate of 97.14 - 99.37%, further validating its robustness against diverse texture-based obfuscations.

5.2. Impact of noise on face recognition

The next experiment aimed to investigate how various types of noise affect face recognition performance. To achieve this, we conducted a series of comparisons across different conditions, examining both intra-class and inter-class matching, as well as the influence of synthetic face data on recognition accuracy. Figure 3 presents boxplots for each type of comparison within the dataset and for selected obfuscation techniques, whereas Table 3 presents the comparison results in terms of the Equal Error Rate, providing a measure of the model's performance under different noise conditions.

As a baseline, we utilized intra-class and inter-class comparisons of authentic samples. Intra-class comparisons involved matching authentic samples with each other under ideal conditions, while inter-class comparisons assessed the ability of the models to differentiate between samples from different individuals. Additionally, authentic samples were compared to their synthetically generated counterparts corresponding to the same class to evaluate the models' performance in distinguishing real from manipulated data.

To analyze the impact of noise, each authentic sample was compared to its noisy counterpart, where various types of noise and distortions were applied with different parameters (e.g., intensity, opacity). These intra-class comparisons (authentic samples vs. authentic samples with noise) allowed us to evaluate how specific noise types affect biometric face recognition and whether they effectively mask individual-specific features across different verification methods.

Similarly, authentic samples were compared to synthetic samples with added noise to asses how specific obfuscation methods influence recognition performance in real-world usecases, and whether noise techniques effectively conceal features associated with manipulated images, such as deepfakes, while maintaining a realistic appearance.

The goal of this experiment was to simulate a real-world scenario where reference data for known individuals (authentic samples) are used to identify fake materials, often deepfake videos or images, that are deliberately obfuscated to evade detection in online environments. By examining

Comparison type	Effect	Celeb-DF-2								Incidents							
		ArcFace		Facenet512		MagFace		VGG-Face		ArcFace		Facenet512		MagFace		VGG-Face	
		Auth	Synth	Auth	Synth	Auth	Synth	Auth	Synth	Auth	Synth	Auth	Synth	Auth	Synth	Auth	Synth
Impostors (base)		0,076	0,264	0,028	0,170	0,048	0,229	0,055	0,227	0,126	0,412	0,073	0,422	0,068	0,319	0,085	0,447
Masked samples		-	-	-	-	-	-	-	-	0,296	0,296	0,288	0,284	0,236	0,234	0,323	0,325
Comparisons between unmasked and masked samples																	
FocusDrift - 5		0,468	0,234	0,484	0,154	0,478	0,230	0,485	0,212	0,568	0,513	0,498	0,373	0,476	0,285	0,526	0,364
WatermarkEffect		0,242	0,122	0,272	0,090	0,369	0,177	0,257	0,119	0,369	0,254	0,361	0,259	0,373	0,280	0,388	0,269
ColorReduce - 4	increasing amount of colors	0,232	0,113	0,288	0,094	0,261	0,120	0,283	0,139	0,239	0,149	0,297	0,252	0,291	0,211	0,313	0,213
ColorReduce - 8		0,420	0,200	0,467	0,154	0,421	0,178	0,453	0,219	0,435	0,306	0,443	0,333	0,418	0,317	0,459	0,351
ColorReduce - 16		0,489	0,241	0,517	0,168	0,489	0,222	0,511	0,232	0,463	0,431	0,465	0,442	0,453	0,262	0,460	0,345
Compression - 10	increasing quality of compression	0,391	0,196	0,446	0,152	0,415	0,193	0,449	0,213	0,419	0,356	0,438	0,414	0,415	0,222	0,474	0,453
Compression - 30		0,493	0,245	0,520	0,170	0,495	0,231	0,510	0,235	0,513	0,391	0,474	0,329	0,471	0,177	0,483	0,282
Compression - 50		0,500	0,253	0,525	0,168	0,505	0,230	0,520	0,230	0,445	0,464	0,435	0,507	0,352	0,318	0,431	0,331
Movie - 1	increasing intensity	0,395	0,203	0,465	0,157	0,452	0,219	0,461	0,227	0,422	0,278	0,438	0,342	0,389	0,269	0,452	0,386
Movie - 3		0,148	0,053	0,198	0,059	0,257	0,126	0,207	0,101	0,141	0,100	0,203	0,146	0,226	0,145	0,205	0,110
Movie - 5		0,034	0,020	0,020	0,012	0,084	0,049	0,028	0,014	0,076	0,067	0,071	0,074	0,130	0,091	0,089	0,063
Rain - 1	increasing intensity of rain	0,302	0,161	0,303	0,123	0,349	0,187	0,331	0,167	0,385	0,259	0,345	0,258	0,340	0,258	0,406	0,303
Rain - 2		0,233	0,121	0,246	0,100	0,285	0,159	0,252	0,126	0,346	0,224	0,334	0,183	0,366	0,251	0,365	0,321
Rain - 3		0,165	0,088	0,178	0,078	0,260	0,139	0,188	0,098	0,192	0,130	0,223	0,099	0,269	0,127	0,261	0,126
Rain - 4		0,074	0,035	0,081	0,039	0,154	0,089	0,084	0,047	0,116	0,049	0,109	0,102	0,216	0,162	0,129	0,117
DotGrid - 0.1	decreasing transparency of dots	0,905	0,440	0,968	0,324	0,955	0,455	0,928	0,453	0,842	0,723	0,880	0,820	0,911	0,765	0,870	0,882
DotGrid - 0.2		0,367	0,178	0,347	0,119	0,425	0,202	0,293	0,147	0,366	0,276	0,455	0,355	0,427	0,276	0,436	0,243
DotGrid - 0.3		0,254	0,105	0,192	0,070	0,370	0,174	0,138	0,061	0,291	0,116	0,324	0,139	0,307	0,151	0,262	0,118

Table 3. EER results for intra-class and inter-class comparisons (first row) and intra-class comparison between authentic and synthetic probes across different noise types and datasets calculated on 10 folds of data. 'Comparison Types' represent intra-class comparisons vs. inter-class comparisons (Impostors), Masked samples, and various types of noise-based augmentations based on Albutortion. Color scale is shared for all models and datasets, where yellow-ish tint represents the most similar samples relative to the reference probes, and purple indicates the most separated samples (highest diversity relative to the reference probes).

the distribution of similarity scores across these scenarios, we evaluated the robustness of biometric engines to various types of noise and their ability to detect deepfake content masked with obfuscation techniques.

For distinguishing between different individuals, Facenet512 demonstrates the best performance on the Celeb-DF-v2 dataset, achieving the lowest EER of 2.8%, indicating strong separation between identities. On the Incident-Based dataset, MagFace outperforms other models, with an EER of 6.8%, showcasing its robustness in more challenging, real-world scenarios. For matching real images to deepfake-manipulated faces, ArcFace exhibits the highest EER values, indicating increased difficulty in distinguishing these cases, for Celeb-DF-v2, EER was 26.4%, for Incidents - 41.2%. If we were to evaluate which systems is the most secure in rejecting manipulated data, FaceNet512 demonstrates the best performance. It achieves the lowest EER with value of 17.0% on Celeb-DF-v2 dataset, and 29.6% on the incidents.

For **Focus Drift**, the results indicate a significant overlap between the distributions of focus-distorted samples and authentic samples, with EER values around 50% across all models. This overlap reflects the tendency of focus distortions to slightly shift the scores toward impostors while maintaining a high degree of similarity to genuine distributions. In contrast, for synthetic samples, the focus drift causes a notable shift in the distributions toward genuine samples. This shift results in an increased EER, effectively masking the features that distinguish synthetic data as deepfakes. Consequently, this obfuscation reduces the models' ability to detect synthetic content, highlighting the challenge of distinguishing deepfake materials under focus drift distortions. Similar conclusions can be drawn for **Watermarks**, however, they have a less impact on shifting the distributions and are particularly noticeable in the Celeb-DF-v2 dataset.

Color reduction has the most significant impact in cases of drastic color reduction. For example, reducing the colors to just 4 levels practically makes face verification impossible. This effect is particularly noticeable in the Celeb-DF-v2 dataset, which contains faces of lower quality compared to the Incident-Based dataset. This difference can be attributed to the baseline resolution of face crops (mean face area in Celeb was 171 x 121px, Incidents - 412 x 303px). **Compression** significantly impacts face verification, especially at extreme levels (e.g., Compression-50). For authentic samples, it increases EER by obscuring identity-specific features, while for synthetic samples, it hides deepfake artifacts, making them resemble genuine data and complicating detection.

Noise introduced through the **Movie** augmentation with a high intensity index (indicating the strength or frequency of the applied distortion) renders identity verification nearly impossible, for both authentic data and deepfake data. The results suggest that such high-frequency noise disrupts identity-specific features, making it difficult to distinguish between identities. Moreover, this type of noise also obscures deepfake artifacts, effectively hiding evidence of manipulation and making detection and verification equally challenging. Dense **grids** (or more generally, fine and small perturbations) make verification nearly impossible, similar to the movie effect. Such patterns disrupt identity-specific features and obscure deepfake artifacts, leading to significant challenges in both detecting manipulations and verifying identities. This behavior underscores how small, repetitive obfuscations effectively mask both unique facial characteristics and evidence of synthetic manipulations, making them particularly problematic for biometric systems.

It is evident that low-opacity **DotGrid** obfuscations (e.g., 0.1) have a minimal impact on face verification accuracy for authentic samples. The distribution shift remains minor, as indicated by a modest rise in EER scores. However, as opacity increases, the overlap between distributions (genuine and impostor) decreases significantly. For example for Celeb-DF-v2 dataset, for opacity of 0.3, the overlap is reduced to approximately 25% for ArcFace, 19% for FaceNet512, 37% for VGG, and 14% for MagFace, indicating a significant degradation in the separability of authentic samples. In synthetic samples, the impact is even more pronounced. The addition of DotGrid obfuscations significantly increases the EER, demonstrating that synthetic data becomes much harder to differentiate from authentic data as the opacity of the noise increases. This suggests that DotGrid obfuscations effectively obscure key identity features in synthetic images. The same observations are evident for synthetic data, similarly across both datasets.

6. Conclusions

The experimental results underscore the critical need for advancing face detection models tailored to noisy and obfuscated data. Standard detection algorithms demonstrated significant limitations when confronted with challenging conditions, such as intense distortions, grid overlays, and environmental noise. Proposed model, utilizing the Albutortion-based augmentation approach, showed markedly improved performance, achieving high detection rates even under extreme scenarios. This highlights the potential of targeted augmentation and model optimization in enhancing the robustness of detection systems. Further development of detectors specifically designed for noisy datasets is imperative to ensure reliable operation in real-world applications, where such conditions are increasingly prevalent.

Across all analyzed biometric engines, certain obfuscations significantly degrade performance, highlighting the sensitivity of face recognition systems to specific distortions. Models like ArcFace and MagFace demonstrate relatively robust performance across various augmentations but remain vulnerable to occlusions and noise-heavy transformations. VGG-Face is the most affected by textural obfuscations and environmental effects, while Facenet512 struggles with occlusions like masked fakes. Adding noise frequently complicates the verification process; however, in certain cases, it can obscure the distinctive features that characterize deepfake materials, aligning the distribution of synthetic masked samples closer to genuine samples. This effect poses an additional challenge for identifying manipulated content.

The study highlights the need for augmentation-aware training and improved resilience to obfuscations, especially those mimicking real-world manipulations, for e.g. fake advertisements, watermarks, and environmental noise.

6.1. Discussion of potential negative societal impact

Data that includes unique information about person's identity is currently protected by law in many regions of the world. In European Union, in accordance with GDPR, each use of subject's image has to be consensual. However, this law does not stop scammers from using images of real people. The extent of misuse of individuals' portrayals surpasses estimations. The consequences of this phenomenon are twofold, as there are two victims when speaking of deepfake-based fraudulent scams and spreading disinformation. First victim, more personal, is the individual that can be seen in synthetic material. Depending on the content, such a person, can experience defamation, reputational damages as well as can lose societal credibility. These consequences are often followed by psychological damages that directly affect the victims. The other type of 'exploited' victims by misuses of imaged via deepfake technology is the society.

People naturally trust visual information, making them susceptible to disinformation. The proliferation of fraudulent content can lead to misguided decisions, the spread of false narratives, and a general erosion of public trust in authorities and reliable information sources. In extreme cases, manipulated media can incite societal unrest, deepen polarization, and undermine democracy by sowing confusion and doubt.

These "two sides of the coin" underscore the urgent need to develop advanced technologies to detect and counter synthetic content. While this paper focuses on enhancing the robustness of biometric systems against synthetic media, the broader challenge remains to address the societal implications. Beyond technological solutions, there is a pressing need for public education, transparent policies, and cross-sector collaboration to safeguard individuals and society from the growing threats posed by deepfake technology.

Acknowledgments

This work was supported by the "Excellence Initiative – Research University" program and the Faculty of Physics and Applied Computer Science at AGH University of Kraków. Additional funding came from the "Science4Business" program under "Inkubator Rozwoju," promoting science-industry collaboration.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006. 2

[2] Shahad Salh Ali, Jamila Harbi Al' Ameri, and Thekra Abbas. Face detection using haar cascade algorithm. In *2022 Fifth College of Science International Conference of Recent Trends in Information Technology (CSCTIT)*, pages 198–201, 2022. 4

[3] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus, 2019. 4

[4] Peter N Belhumeur, Joao P Hespanha, and David J Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. 2

[5] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. 3

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. *arXiv preprint arXiv:1710.08092*, 2018. 5

[7] Hardie Cate, Fahim Dalvi, and Zeshan Hussain. Deepface: Face generation using deep learning. *CoRR*, abs/1701.01876, 2017. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[9] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5202–5211, 2020. 4

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 2, 5

[11] Redline Digital. Fake news statistics facts. https://redline.digital/fake-news-statistics/, 2024. 1

[12] Tom Gerken and Chris Vallance. Facebook and Instagram launch celebrity scam ad crackdown. https://www.bbc.com/news/articles/cg565mrdz7zo. Accessed: 2024-11-05. 1

[13] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. 5

[14] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017. 5

[15] Juho Kannala and Esa Rahtu. Bsif: Binarized statistical image features. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1363–1366, 2012. 2

[16] Kuang-Chih Lee, J. Ho, and D.J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005. 2

[17] Yuezun Li, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, Seattle, WA, United States, 2020. 4, 5

[18] Chengjun Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, 11(4):467–476, 2002. 2

[19] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[20] R. C. Malli. VGGFace implementation with Keras framework, 2019. Last accessed on June 2019. 5

[21] Iacopo Masi, Sara Rawls, Gerard Medioni, and Premkumar Natarajan. Do we really need to collect millions of faces for effective face recognition? In *ECCV*, pages 579–596, 2016. 2

[22] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. *CoRR*, abs/2103.06627, 2021. 2, 5

[23] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 2

[24] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, 2015. 2

[25] Md. Tajmilur Rahman and Md. Alamin Bhuiyan. Face recognition using gabor filters. In *2008 11th International Conference on Computer and Information Technology*, pages 510–515, 2008. 2

[26] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. 2

[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 2

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 5

[29] Linlin Shen and Li Bai. A review on gabor wavelets for face recognition. *Pattern Anal. Appl.*, 9:273–292, 09 2006. 2

[30] Yichun Shi and Anil K. Jain. Probabilistic face embeddings. 2019. 2

[31] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. 2

[32] Xiaoyang Tan and Bill Triggs. Fusing gabor and lbp feature sets for kernel-based face recognition. In S.K. Zhou, W. Zhao, X. Tang, and S. Gong, editors, *Analysis and Modeling of Faces and Gestures. AMFG 2007*, volume 4778 of *Lecture Notes in Computer Science*, pages 235–249. Springer, Berlin, Heidelberg, 2007. 2

[33] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. 2

[34] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991. 2

[35] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. *CoRR*, abs/1801.09414, 2018. 2

[36] Xiang Wang, Kai Wang, and Shiguo Lian. A survey on face data augmentation for the training of deep neural networks. *Neural Computing and Applications*, 32(19):15503–15531, Mar. 2020. 2

[37] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009. 2

[38] Wei Wu, Hanyang Peng, and Shiqi Yu. Yunet: A tiny millisecond-level face detector. *Machine Intelligence Research*, pages 1–10, 2023.

[39] Bin Ye, Yunlin Shi, Huijun Li, Liuchuan Li, and Shuo Tong. Face ssd: A real-time face detector based on ssd. In *2021 40th Chinese Control Conference (CCC)*, pages 8445–8450, 2021. 4

[40] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. In *arXiv preprint arXiv:1411.7923*, 2014. 5

[41] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016. 4

[42] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020. 2