An Investigation into the Impact of AI-Powered Image Enhancement on Forensic Facial Recognition

Justin Norman and Hany Farid
University of California, Berkeley
Berkeley CA, USA
justin.norman@berkeley.edu and hfarid@berkeley.edu

Abstract

Advances in machine learning and computer vision have led to significant improvements in automated facial recognition. Many real-world forensic settings, however, are confronted with challenging low-quality and low-resolution images that often confound even state-of-the art facial recognition. We investigate if and when advances in neural-based image enhancement and restoration can be used to restore degraded images while preserving facial identity for use in forensic facial recognition.

1. Introduction

Although automatic facial recognition has its roots in the mid 1960s [2, 3], it wasn't until fairly recently that the accuracy of facial recognition has achieved levels allowing it to be credibly deployed in real-world forensic settings [22]; albeit, not without concerns regarding human-rights violations [25], privacy [10, 24], and bias [4, 8, 20]. It has been argued that automatic facial recognition is as or more accurate than human-level recognition [18] (see [19, 27] for some opposing views). These advances in automatic facial recognition have been largely fueled by advances in machine learning along with access to increasingly larger and more diverse datasets.

Parallel advances in machine learning have also fueled a revolution in image enhancement in which noisy, low-resolution, or blurry images can be seemingly miraculously restored to their high-resolution and high-quality originals [6, 11, 21, 30]. Because automatic facial recognition can struggle with low-quality images [9], and because low-resolution and blurry images are not uncommon in real-world scenarios, we wondered if these image enhancement tools would improve facial recognition accuracy in the face of low-quality images.

On the one hand, as shown in the top portion of Figure 1, a super-resolution image enhancement [14] appears to be able to recover many facial features from a $4\times$ downsampled version of the original. On the other hand, as shown in the bottom portion of this figure, the same image enhancement appears to hallucinate facial features not present in the original.

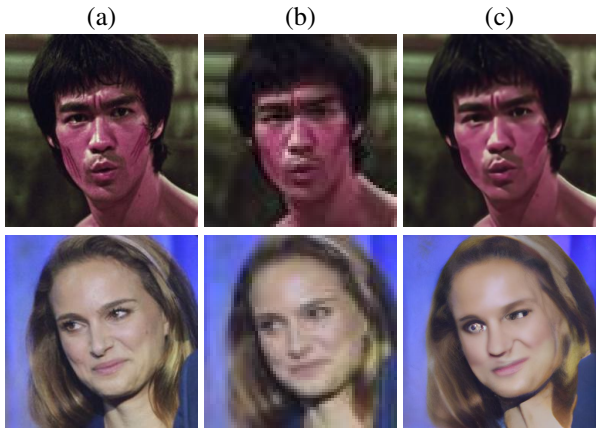

Figure 1. An example of (a) an original image, (b) a $4\times$ low-resolution version of this image, and (c) the result of enhancing the degraded image in panel (b). In one case (top) the image enhancement appears to respect the original facial features, while in another case (bottom), the enhancement introduces or distorts facial features relative to the original.

To this end, we examine how image enhancement in the form of super-resolution and de-blurring impacts facial recognition. This study makes use of two large and diverse facial sets, two popular deep-learning facial recognition systems and 12 different GAN- and diffusion-based image enhancement techniques. We conclude with recommendations for the use (and not) of image enhancement in forensic facial recognition.

2. Face Enhancement

We begin by describing two typical image enhancements that a forensic analyst might utilize: super-resolution in

which a low-resolution image is up-sampled to a higher resolution while restoring original image details; and deblurring in which optical or motion blur is removed from an image. Each of these problems has received considerable academic and industry attention [30, 31]. Recent neural-based approaches, however, have made significant progress in terms of recovering the original content from even highly degraded inputs. We will begin by reviewing several of these state-of-the-art neural approaches to image enhancement.

2.1. Super-resolution

We explore three distinct neural-based techniques for super-resolution. These techniques span a range of different underlying mechanisms from generative adversarial networks (GANs) to convolutional neural networks (CNNs), Transformers, and combinations of all three.

2.1.1 EDSR

The enhanced deep residual neural network for single image super-resolution (EDSR) [15] achieves impressive results through the use of several clever model architecture optimizations. The authors of EDSR build on Ledig et al.'s [13] successful application of the ResNet architecture to super-resolution tasks. Specifically, EDSR is a modified version of Ledig et al.'s model in which the batch normalization layers are removed, leading to a significant improvement in the fidelity of the super-resolution images. This improvement is due to the tendency of batch normalization layers to restrict the full range of network representations as a by-product of feature normalization. Perhaps more importantly, removing batch normalization layers tends to yield a significant reduction in GPU memory utilization. This optimization, in turn, allows for a larger overall model which generally correlates with improved performance.

Additionally, the EDSR model architecture addresses the common problem of increasing training instability as feature map size increases by both adding constant scaling layers after the final convolutional layers, and removing activation layers after residual block layers. The EDSR model also takes advantage of interim model training by initializing the final model with pre-trained parameters.

2.1.2 Swin-IR

Vision Transformers have recently been introduced as an alternative to CNN-based image restoration methods. These Transformers address the problems of CNN's content-independent relationship between input images and the convolution kernel, and CNN's struggle with capturing long-range dependencies [16]. Transformer-based models, due to their use of self-attention layers, are able to represent global relationships between different contexts. This technique, however, still relies on a fixed-size patching strategy, each of which is processed independently. This leads to limitations in that neighboring patches cannot share context and, as a result, patch-border artifacts tend to appear in the restored images.

By combining both architectures, the Swin Transformer [16] addresses both of these challenges as well as the inefficiencies of CNN-based approaches, while also improving upon the basic Transformer technique. These new architectures are able to accept large image resolutions as inputs to the pipeline as a result of the local attention property of CNNs. Simultaneously, Swin Transformers benefit from the long-range dependency modeling capabilities of the shifted window process.

Swin-IR [14] improves upon the Swin Transformer by applying a three-phased process: shallow feature extraction, deep feature extraction, and a high-quality image restoration method. The shallow feature extraction utilizes a convolution layer which is then applied directly to the image reconstruction as the primary low-frequency input. The deep feature extraction is accomplished through the introduction of Residual Swin Transformer blocks, which apply the shifted window method for local attention. Both of these features are then combined via a convolution layer for feature optimization and aggregation. The reconstruction is completed as deep and shallow features are combined to form a high-quality image.

2.1.3 LDM

Diffusion models (DM) have also proven to be surprisingly effective on a variety of image restoration tasks. The general approach of diffusion image models is to leverage denoising autoencoders in order to segment image formation into sequential and progressive steps [21]. This process, however, relies on processing images directly in pixel space, which is computationally expensive and requires massive computing infrastructure, typically available to only a few well-resourced organizations.

Latent diffusion models (LDM) were introduced to address this shortcoming. As the name suggests, LDMs operate in a lower-dimensional latent space which supports the ability to train the image restoration models on more standard and accessible computing resources [21]. Beginning with a pretrained (in pixel space) diffusion model, an LDM follows the standard, two-phase DM process of perceptual compression for low-frequency learning, followed by semantic learning and composition in the semantic compression phase. An autoencoder is trained in the latent space to provide a low-dimension representation which is a perceptual twin of the image space. LDMs then leverage the lower model complexity to restore an image in a single net-

Figure 2. A representative set of real (top) and synthetically-generated (bottom) faces.

work pass, still in the latent space.

2.1.4 CodeFormer

The blind face restoration technique CodeFormer has been employed for two primary tasks: reduce or remove perceptible image degradation and match degraded image features to a desired image quality and style [32]. This technique employs a Transformer-based architecture to create a representation for low-quality images that is specifically contextualized for human faces. The CodeFormer architecture is defined by three distinct components: (1) a quantized autoencoder trained for the purpose of creating a contextual codebook for the face reconstruction task; (2) to correct for the tendency of algorithmic feature matching to fail when processing corrupted textures, a Transformer is utilized to provide references to more global representations; and (3) a controllable feature transformation is leveraged to flexibly manage context flow from the low-quality sample processed by the encoder to the decoder feature set.

2.2. De-blurring

We explore three distinct neural-based techniques for deblurring. These techniques span a range of different underlying mechanisms from generative adversarial networks (GANs) to CNNs, Transformers, and combinations of all three.

2.2.1 MPRNet

Image restoration has historically been highly dependant on labor intensive hand-crafted and explicitly curated training data [31]. As was the case with many of the super-resolution models described above, researchers have more recently turned to CNNs as a method for organically learning image representations from large datasets [31]. Most of these CNN-based approaches, however, employ a single-stage design, which is suboptimal for many complex computer vision problems.

MPRNet leverages a multi-stage approach, employing an encoder-decoder for multi-scale learning, while applying a final stage that functions directly on the original image resolution in order to capture fine-grained spatial detail [31]. MPRNet also leverages a supervised attention module (SAM) sandwiched between stage pairs in order to facilitate continuous learning of features, additionally utilizing ground-truth to fine-tune progressive stages based on the previous stages. Finally cross-stage feature fusion (CSFF) is applied, which true to name, combines the learned features from early stages (of different scales) to progressively later stages.

2.2.2 HINet

This second image deblurring technique builds on MPRNet described in the previous section. A half instance normalization network (HINet) approach [6] was originally proposed as an effort to avoid the computational cost of a high number of multiplier-accumulator operations within other multi-stage image restoration architectures such as MPRNet. Since the sampled small image patches within training batches are highly variant, batch normalization is not a popular technique for low-level image formation tasks. Instance Normalization (IN) instead is able to calculate and balance this variance of image features without leveraging the batch process which makes the resulting networks more tolerant to changes in scale.

The improved performance observed by HINet was primarily achieved through two innovations: the addition of half-instance normalization (HIN) blocks and the implementation of a multi-stage network architecture that applies these HIN blocks as stacked layers in each encoder subnetwork stage. The HINet model also employs the use of CSFF and applies a SAM between these stages. These steps have the result of enriching features at differing scales (which MPRNet originally advanced) while preserving the performance gain of the HIN blocks.

2.2.3 Restormer

As discussed earlier, though efficient, CNNs struggle when it comes to capturing long-range, complex correlations from

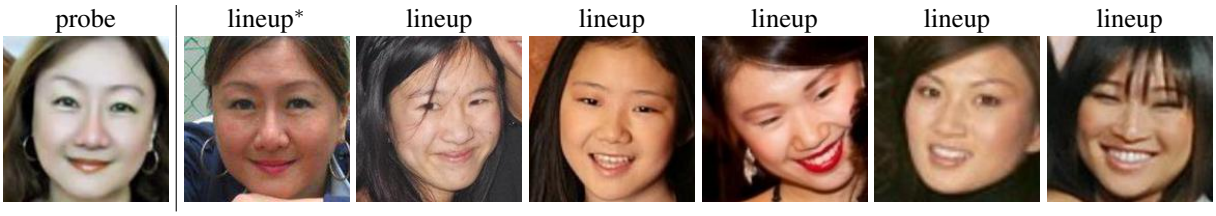

Figure 3. Example of a facial forensic lineup consisting of a probe image (left) and six standardized images, one of which (*) matches the identity in the probe image, and the rest of which are decoys.

inputs. This is because CNNs usually have smaller receptive fields than a typical Transformer [30]. Although Transformer-based models largely mitigate these limitations, they do so at the cost of network complexity and computational efficiency, particularly with respect to high-resolution images. The restoration Transformer model Restormer addresses these complexity issues.

Restormer's mitigation is accomplished through the introduction of a new representation-learning approach applied both locally and globally on high-resolution image samples [30]. This approach avoids segmentation of the inputs into smaller local patches, thus preserving global image context. This approach also involves the application of a multi-deconvolution self-attention layer (Dconv), which is an efficient upsampling process capable of fusing both local and global context between pixels. A gated-DConv feed-forward network approach is utilized to perform a highly controllable transformation of features, biasing the inclusion of high-information features into the subsequent representation. The improvements introduced by Restormer allow for Transformer-based models to be practically utilized for image restoration tasks.

3. Datasets

We make use of two datasets for our evaluations. The first real-world dataset is derived from the CASIA-Webface dataset, consisting of 491,414 images derived from 10,575 identities. These images are of various size, quality, pose, subject clothing, and environment, Figure 2 (top row). Due to the initial quality of the dataset, some manual curation was necessary including the removal of duplicate images and the removal of incorrectly labeled images.

A second, synthetically-generated dataset, is also used as it affords more fine-grained control over differences in each subject's appearance within and across identities. Specifically, we employ Synthesis AI's commercial software (https://synthesis.ai) which uses a combination of classic rendering and generative synthesis to create photorealistic human faces. All images are rendered at a resolution of 512×512 pixels. A total of 200,000 images were rendered consisting of 8,000 unique identities with varying head poses, expressions, head wear, facial hair, hairstyles, glasses (opaque and clear), masks, backgrounds, and environmental lighting, Figure 2 (bottom row).

4. facial recognition

4.1. Forensic Lineup

We evaluate two popular facial recognition systems, FaceNet [23] and ArcFace [9]. FaceNet utilizes an inception ResnetV1-based model architecture, trained and evaluated on either the CASIA-WebFace [28] or Visual Geometry Group Face Dataset 2 (VGGFace2) [5] datasets. For our analysis, we utilized the VGGFace2-trained version. The network yields a 128D embedding from each input image. This results in an output such that the squared L2 distances in embedding space represent face similarity, where similar faces have small distances and dissimilar faces have large distances.

ArcFace utilizes a 512D normalized embedding feature, organized into distinct clusters representing individual identities. The model architecture then employs an additive angular margin loss, yielding better identity separability and, in turn, recognition accuracy. ArcFace is trained and evaluated on CASIA-Webface, VGGFace2, and a curated and tightly-cropped-to-faces version of MS1MV0 [1].

Over the past decade, the improvement of facial recognition models for forensic identification tasks has been dramatic. However, much of the evaluation of the performance of such models has been conducted in controlled lab settings that do not necessarily replicate the data diversity and task difficulty inherent in real-world forensic settings.

In order to address this shortcoming, a new forensic lineup methodology was proposed [17]. In this task, a single image (the probe) is compared against a lineup of six perceptually similar faces (as measured by the latent representation of each face). The face in the lineup that is most similar to the probe, as measured by any standard distance metric in any latent representation is considered a presumptive match. This lineup approach ensures that the comparison group across a large database is always similar.

Evaluation against both the synthetic and real-world datasets (Section 3) reveals that previously reported facial recognition accuracy for these two face-recognition models exceeding 95% fall to as low as 65% in this more controlled

model	resolution	FaceNet accuracy (%)	ArcFace accuracy (%)
Original	$1\times$	78.2	83.8
Baseline	$\Downarrow 4\times$	74.0	82.7
TorchSR edsr	$\Downarrow 4\times + \Uparrow 4\times$	80.7	87.3
TorchSR ninasr_b2	$\Downarrow 4\times + \Uparrow 4\times$	81.6	87.2
Swin-IR	$\Downarrow 4\times + \Uparrow 4\times$	81.5	87.5
LDM	$\Downarrow 4\times + \Uparrow 4\times$	58.7	69.6
CodeFormer	$\Downarrow 4\times + \Uparrow 4\times$	73.5	82.0
Baseline	$\Downarrow 8\times$	47.1	56.2
TorchSR edsr	$\Downarrow 8\times + \Uparrow 8\times$	69.7	80.4
TorchSR ninasr_b2	$\Downarrow 8\times + \Uparrow 8\times$	69.1	80.2
Swin-IR	$\Downarrow 8\times + \Uparrow 8\times$	69.4	79.4
LDM	$\Downarrow 8\times + \Uparrow 8\times$	58.4	68.3
CodeFormer	$\Downarrow 8\times + \Uparrow 8\times$	51.6	58.2

Table 1. Facial recognition accuracy for images at their original resolution ($1\times$), at reduced resolution ($\Downarrow 4\times$ and $\Downarrow 8\times$), and these reduced resolution super-resolved to the original resolution ($\Downarrow 4\times + \Uparrow 4\times$ and $\Downarrow 8\times + \Uparrow 8\times$).

and challenging forensic lineup task.

We employ this same forensic lineup task in evaluating the impact of super-resolution and de-blurring on facial recognition.

4.2. Super-Resolution

Operating on the real-world CASIA-Webface dataset (Section 3), the accuracy on the forensic lineup task for FaceNet and ArcFace is 78.2% and 83.8%, (top row of Table 1). With six images in the lineup, chance performance is $1/6 = 16.7\%$.

As shown in Table 1, the average accuracy on this lineup task for FaceNet reduces to 74.0% and 47.1% for probe images reduced in resolution by $4\times$ and $8\times$. For ArcFace, accuracy reduces to 82.7% and 56.2%.

Shown in the upper portion of Table 1 are the accuracies after down-sizing each probe image by $4\times$ and then applying different super-resolution enhancements to return each image to its original resolution (Section 2.1). For FaceNet, the average accuracy on the super-resolved probe images ranges from 81.6% (an improvement of 7 percentage points as compared to the baseline of operating on the $4\times$ lower-resolution image) to 58.7% (a degradation of 15 percentage points compared to baseline). For ArcFace, the pattern is similar where the average accuracy ranges from 87.5% (an improvement of 5 percentage points over baseline) to 69.6% (a degradation of 13 percentage points compared to baseline).

Shown in the lower portion of Table 1 are the accuracies after down-sizing each probe image by $8\times$ followed by super-resolution. For FaceNet, the average accuracy on the super-resolved probe images is consistently higher than baseline of operating on the $8\times$ lower-resolution image, ranging from an improvement between 4 and 22 percentage points. For ArcFace, the average accuracy on the super-resolved probe images is consistently higher than baseline ranging from an improvement between 2 and 24 percentage points.

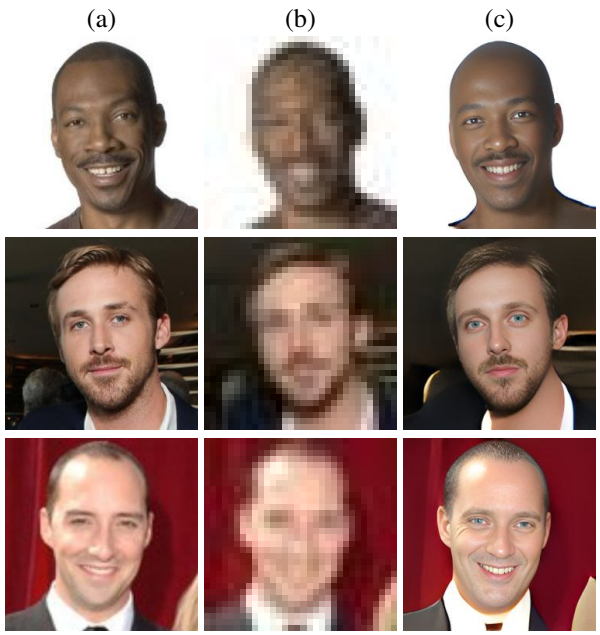

Figure 4. Examples of (a) an original image, (b) a $8\times$ low-resolution version of this image, and (c) the result of up-sampling using CodeFormer the low-resolution image back to its original resolution. Note that although the super-resolution image restores high-resolution features, the enhancement has introduced or distorted facial features relative to the original leading to an apparent different identity.

In most cases super-resolution improves facial recognition accuracy as compared to operating on the lower-resolution images. Somewhat surprisingly for the $4\times$ resolution, in most cases super-resolution yields slightly improved accuracy relative to the *original* resolution (first row of Table 1). For the $8\times$ resolution, however, results are more mixed with some models (TorchSR) able to achieve accuracies close to the original resolution while other models (CodeFormer) significantly under perform.

At $4\times$ up-sampling, some of super-resolution algorithms perform better than the original image and better than operating on the baseline lower-resolution image. On the other hand, one super-resolution algorithm (LDM) yields lower accuracy as compared to the original image and to operating on the baseline lower-resolution image. At $8\times$ up-sampling, all of the algorithms perform worse than original resolution but much better than baseline. The appropriate super-resolution algorithm can, we conclude, be an asset to forensic facial recognition.

These results, however, hide a potentially dangerous aspect of some super-resolution algorithms. Shown in Figure 4 are three original images, these images down-sampled

by $8\times$, and these down-sampled images up-sampled to the original resolution using CodeFormer (Section 2.1.4). In these examples, the super-resolved images appear to be of high quality with clear and undistorted features, but the identity of the person is not the same as in the original. In practice it can be difficult to identify this type of failure case.

Although a more in depth analysis is required, we do not find a systematic bias in the hallucinating of facial features in which, for example, a single identity is consistently generated or in which hallucination is dependent on race or gender. A more in-depth analysis will be required to fully explore any other systematic biases.

While, for the most part, the super-resolved images appear of high quality, they will occasionally hallucinate facial features. We have not found a way of determining when this will happen. Care, therefore, must be taken to consider this potential pitfall when using image super-resolution in a forensic setting.

4.3. De-blurring

Operating on the original resolution images from the real-world CASIA-Webface dataset (Section 3), the baseline accuracy on the forensic lineup task for FaceNet and ArcFace is 78.2% and 83.8% (chance performance is $1/6 = 16.7\%$).

We induce optical blur by blurring the probe image with a symmetric box-car kernel ranging in size from 3×3 to 23×23. We induce motion blur by blurring the probe image with an asymmetric Gaussian kernel in the same size range with an aspect ratio of $1:3$ in terms of the horizontal and vertical Gaussian variance, and random orientation.

With no de-blurring intervention, the average accuracy on the lineup task for FaceNet in the presence of optical blur steadily reduces from 77.6% for a 3×3 kernel to 22.3% for a 23×23 kernel (see top row (Baseline) of Table 2). A similar pattern emerges with ArcFace where accuracy ranges from 83.6% to 24.7%.

Motion blur impacts facial recognition less than optical blur. With no de-blurring intervention, the average accuracy on the lineup task for FaceNet in the presence of motion blur steadily reduces from 77.6% for a 3×3 kernel to 55.6% for a 23×23 kernel (see top row (Baseline) of Table 2). A similar pattern emerges with ArcFace where accuracy ranges from 83.5% to 70.7%.

As seen in Table 2, generally speaking, for both motion and optical blur, de-blurring from most models affords an improvement in facial recognition accuracy as compared to operating on the baseline blurred probe images. For the best-performing model (RestormerLocal-HIDE) de-blurring is able to surpass baseline accuracy across all blur types and amounts. There is, however, one exception where CodeFormer consistently yields results worse than baseline.

Perhaps not surprisingly, overall accuracy is worse for optical blur than for motion blur, most likely because the asymmetric motion blur is simply less severe than the symmetric optical blur. Generally speaking, we did not find the same type of facial feature hallucination seen with some super-resolution examples (Figure 4). In particular, for larger amounts of blur, the de-blurring algorithm just fails to completely remove the blur. This is probably a more desirable failure case as it can be clear from the de-blurred image that the image has not been fully enhanced.

We next wondered if we could improve on these results by re-training the best-performing model architecture (RestormerLocal-GoPro) on a more representative dataset. Because this model was trained on GoPro video and not necessarily on faces, we replaced 10% or 50% of the Restormer dataset with images from the VGGFace dataset [5]. Adapting the same method detailed in the original Restormer model [29], we utilized a four-GPU cluster to train four iterations of the Restormer model. These two new models were then used to evaluate accuracy on the same task. As shown in the last two rows of Table 2, accuracy from these retrained models has little impact on accuracy across all blur types and amounts.

Combined, we find that even with large amounts of blur, the appropriate de-blurring model can be an asset to forensic facial recognition.

5. Synthetic Faces

Because they afford more control and diversity, synthetic faces are often used to evaluate and train facial recognition systems. Though not our central focus, we explored the impact of de-blurring and super resolution on synthetically-generated faces. The synthetic faces were generated using Synthesis AI's commercially available software (https://synthesis.ai) [26]. This rendering engine leverages a hybrid of generative AI and traditional 3D modeling to generate photorealistic human faces across a variety of demographics, clothing, scenes and environmental conditions (see Section 3 and bottom row of Figure 2).

With respect to resolution, accuracy for FaceNet on the original resolution synthetic faces is 75.4%, similar to real faces (78.2%). For images downsized by $4\times$ and $8\times$, accuracy for FaceNet drops to 60.8% and 25.4% (as compared to 74.0% and 47.1% for real images). Accuracy for ArcFace on the original resolution synthetic faces is 96.8%, 13 points higher than real faces (83.8%). For images downsized by $4\times$ and $8\times$, accuracy for ArcFace drops to 90.9% and 39.4% (as compared to 82.7% and 56.2% for real images).

Using one of the best performing super-resolution models (Swin-IR), and using the same processing and analysis as with real faces, accuracy for FaceNet is 61.9% and 54.8% for images downsampled and upsampled by $4\times$ and

		motion blur						optical blur					
model	model-dataset	3	7	11	15	19	23	3	7	11	15	19	23
FaceNet	Baseline	77.6	76.7	73.8	69.1	63.1	55.6	77.6	72.8	58.1	42.6	30.3	22.3
FaceNet	HINetLocal-GoPro	77.3	77.7	75.9	72.8	68.0	61.3	78.0	74.6	63.3	47.5	35.6	26.7
FaceNet	MPRNetLocal-GoPro	77.1	76.8	75.7	73.4	70.1	63.7	77.5	74.2	65.2	50.2	37.2	26.7
FaceNet	HINetLocal-HIDE	77.7	77.8	76.0	72.7	67.9	61.4	77.2	74.4	63.6	47.6	35.0	26.1
FaceNet	MPRNetLocal-HIDE	77.6	77.0	76.0	73.6	69.9	64.2	76.9	73.6	63.9	48.8	36.0	26.9
FaceNet	RestormerLocal-HIDE	78.1	77.8	77.5	75.5	71.9	66.1	77.4	74.9	64.3	49.1	37.8	29.8
FaceNet	HINetLocal-REDS	76.4	75.7	73.4	70.2	65.0	57.6	76.7	71.7	57.5	41.7	31.0	24.0
FaceNet	CodeFormer	60.5	55.2	46.7	39.1	32.6	26.8	56.9	40.4	28.2	20.4	16.5	13.8
FaceNet	RestormerLocal-GoPro	77.7	77.3	77.4	75.2	71.3	65.6	78.3	75.2	64.7	49.7	37.8	29.2
ArcFace	Baseline	83.5	83.5	82.6	80.5	76.7	70.7	83.6	82.0	70.8	49.3	32.3	24.7
ArcFace	RestormerLocal-GoPro	84.4	84.0	84.0	83.2	82.0	79.1	84.7	83.8	77.9	65.8	49.7	35.0
FaceNet	RestormerLocal-GoPro (10%)	78.1	77.6	76.0	72.3	67.4	60.7	77.3	73.5	62.0	45.4	34.4	25.2
FaceNet	RestormerLocal-GoPro (50%)	77.0	77.3	75.1	71.3	64.3	57.9	77.9	77.4	76.3	48.2	37.4	28.3

Table 2. Facial recognition accuracy for images that have been motion (left) and optically (right) blurred with kernels in size ranging from 3×3 to 23×23, and then de-blurred. The baseline accuracy corresponds to performing facial recognition directly on the blurred images. The last two rows correspond to the accuracy after the de-blurring model was retrained on facial images.

			motion blur						optical blur					
model	model-dataset	data set	3	7	11	15	19	23	3	7	11	15	19	23
FaceNet	Baseline	synth	75.4	71.8	62.5	51.6	41.9	32.5	74.3	58.1	34.2	22.3	15.3	13.1
FaceNet	RestormerLocal-GoPro	synth	75.6	74.7	71.8	64.4	54.4	44.3	74.9	66.1	44.7	27.7	19.3	14.7
FaceNet	Baseline	real	77.6	76.7	73.8	69.1	63.1	55.6	77.6	72.8	58.1	42.6	30.3	22.3
FaceNet	RestormerLocal-GoPro	real	77.7	77.3	77.4	75.2	71.3	65.6	78.3	75.2	64.7	49.7	37.8	29.2
ArcFace	Baseline	synth	96.9	95.9	93.5	85.9	74.3	60.7	96.6	88.9	55.5	32.0	22.3	18.3
ArcFace	RestormerLocal-GoPro	synth	96.8	96.8	95.9	93.4	87.3	76.1	96.8	93.5	74.2	48.8	31.5	23.5
ArcFace	Baseline	real	83.5	83.5	82.6	80.5	76.7	70.7	83.6	82.0	70.8	49.3	32.3	24.7
ArcFace	RestormerLocal-GoPro	real	84.4	84.0	84.0	83.2	82.0	79.1	84.7	83.8	77.9	65.8	49.7	35.0

Table 3. Facial recognition accuracy for synthetic (top) and real (bottom) images that have been motion (left) and optically (right) blurred with kernels in size ranging from 3×3 to 23×23, and then de-blurred. The baseline accuracy corresponds to performing facial recognition directly on the blurred images.

$8\times$, as compared to 61.5% and 29.8% on the downsampled images. For ArcFace, accuracy is 89.2% and 82.2% for images downsampled and upsampled by $4\times$ and $8\times$, as compared to 90.9% and 39.4% on the downsampled images.

With respect to de-blurring, generally, with the exception of CodeFormer, all the de-blurring models performed slightly better than baseline when the images were optically or motion blurred. For small kernel sizes, the improvement was relatively slight over baseline, but for larger kernel sizes, the improvement was more significant. Shown in top portion of Table 3 are the accuracies for one of the best performing de-blurring models, RestormerLocal-GoPro; for comparison, the accuracies for real images from Table 2 are reproduced here.

For FaceNet, as we saw above, accuracy on original resolution and quality images is similar for synthetic and real images. The impact of motion blurring, however, is different for synthetic and real images, with the impact on synthetic images being more significant for blur kernels larger than 11×11 (top/left portion (Baseline) of Table 3). When de-blurring is applied, accuracy is improved and is similar to real images for kernel sizes between 3×3 and 11×11; for larger kernels, the improvement is less pronounced as compared to real images (see rows 2 and 4 of Table 3). A similar pattern emerges for optical blur (top/right portion of Table 3) but the impact of blurring is even more severe with a difference between synthetic and real emerging after a blur kernel size of 7×7.

The story for ArcFace is different. First, as we saw above, accuracy on original resolution and quality images is significantly higher for synthetic images than real images. For both motion and optical blur, accuracy for synthetic images degrades slightly more for larger kernels as compared to real images. When de-blurred, accuracy for synthetic images recovers similar to real images.

Combined, these results suggest that although the use of synthetic images can be desirable, and holds promise for training and evaluating facial recognition systems, its use in real-world applications is complex. Our results demonstrate that there is a significant, and at times difficult to interpret, interplay between the facial recognition model, image enhancement model, and image quality. This interplay

yields significantly different facial recognition performance between synthetic and real images. We conclude that synthetic images are not a simple proxy for real images. As described in [17], however, the quality and resolution of synthetic images can be calibrated to more closely match real images. Nevertheless, care should be taken when incorporating synthetic images into the training or evaluation of facial recognition tasks.

6. Discussion

Having explored the impact of super-resolution and motion/optical de-blurring on forensic facial recognition, we find that under certain conditions, and with the appropriate choice of enhancement model, these tools can be an asset. At the same time, this type of image enhancement is not a panacea, and care must be taken when deploying these techniques to carefully understand their efficacy in the presence of different levels of image degradation, the type of degradation, the nature of the desired enhancement and the underlying face-recognition model.

On the other hand, the failure cases we observed are concerning. We observed that at times, image enhancement can hallucinate facial features and facial identity (Figure 4). What is particularly worrying about these hallucinations is that there is no obvious way to determine that such a hallucination has occurred by only looking at the enhanced image.

Further analysis will be required to assess the efficacy of other forms of image enhancement in the form of, for example, de-noising and in-painting, and the interplay between different forms of image degradation.

Our initial attempt to retrain a generic de-blurring model on faces did not yield an improvement. Further analysis is also required to determine if facial recognition will be improved by more specialized image enhancement models [12] and/or explicitly training a facial recognition model on a wide range of degraded images [7]. Because there is an interplay between image enhancement and the underlying facial recognition model, it is important that any evaluation be performed holistically.

Acknowledgements

We are grateful to the team at Synthesis AI (https://synthesis.ai) for generously providing access to their image synthesis API.

References

[1] Gwangbin Bae, Martin de La Gorce, Tadas Baltrusaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. DigiFace-1M: 1 Million Digital Face Images for Face Recognition. arXiv:2210.02579, 2022. 4

[2] Woodrow Wilson Bledsoe. Man-machine facial recognition. Panoramic Research Inc., Palo Alto, CA, 1966. 1

[3] Woodrow Wilson Bledsoe. The model method in facial recognition. Panoramic Research Inc., Palo Alto, CA, 1966. 1

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018. 1

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018. 4, 6

[6] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *International Conference on Computer Vision and Pattern Recognition*, pages 182–192, 2021. 1, 3

[7] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Low-resolution face recognition. In *14th Asian Conference on Computer Vision*, pages 605–621. Springer, 2019. 8

[8] Alexandra Chouldechova, Siqi Deng, Yongxin Wang, Wei Xia, and Pietro Perona. Unsupervised and semi-supervised bias benchmarking in face recognition. In *European Conference on Computer Vision*, pages 289–306. Springer, 2022. 1

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1, 4

[10] Kashmir Hill. The secretive company that might end privacy as we know it. In *Ethics of Data and Analytics*, pages 170–177. Auerbach Publications, 2020. 1

[11] Junjun Jiang, Chenyang Wang, Xianming Liu, and Jiayi Ma. Deep learning-based face super-resolution: A survey. *ACM Computing Surveys*, 55(1):1–36, 2021. 1

[12] Yimei Kang and Wang Pan. A novel approach of low-light image denoising for face recognition. *Advances in Mechanical Engineering*, 6:256790, 2014. 8

[13] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In

International Conference on Computer Vision and Pattern Recognition, pages 4681–4690, 2017. 2

[14] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1, 2

[15] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *International Conference on Computer Vision and Pattern Recognition Workshop*, pages 136–144, 2017. 2

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision and Pattern Recognition*, pages 10012–10022, 2021. 2

[17] Justin Norman, Shruti Agarwal, and Hany Farid. An evaluation of forensic facial recognition. arXiv:2311.06145, 2023. 4, 8

[18] Alice J O'Toole and Carlos D Castillo. Face recognition by humans and machines: Three fundamental advances from deep learning. *Annual Review of Vision Science*, 7:543–570, 2021. 1

[19] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018. 1

[20] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151, 2020. 1

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *International Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2

[22] Antoaneta Roussi. Resisting the rise of facial recognition. *Nature*, 587(7834):350–354, 2020. 1

[23] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *International Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 4

[24] Marcus Smith and Seumas Miller. The ethical application of biometric facial recognition technology. *AI & Society*, 37(1):167–175, 2022. 1

[25] John Sudworth. The faces from China's Uyghur detention campus. *BBC*, 2022. 1

[26] Synthesis.AI. https://synthesis.ai/synthesis-humans/, 2022. 6

[27] Alice Towler, James D Dunn, Sergio Castro Martínez, Reuben Moreton, Fredrick Eklöf, Arnout Ruifrok, Richard I Kemp, and David White. Diverse types of expertise in facial recognition. *Scientific reports*, 13(1):11396, 2023. 1

[28] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. arXiv:1411.7923, 2014. 4

[29] Syed Waqas Zamir. Restormer. https://github.com/swz30/Restormer, 2023. 6

[30] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *International Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 1, 2, 4

[31] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *International Conference on Computer Vision and Pattern Recognition*, pages 14821–14831, 2021. 2, 3

[32] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *Advances in Neural Information Processing Systems*, volume 35, pages 30599–30611, 2022. 3