BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES,
DOI: 10.24425/bpasts.2027.7967

THIS IS AN EARLY ACCESS ARTICLE.
This article has been accepted for publication in a future issue of this journal,
but has not been fully edited. Content may change prior to final publication.

Enhancing Face Similarity Detection with GAN-Generated Synthetic Data

Wojciech Domski[1]*, **Adam Jankowiak**

[1] Department of Cybernetics and Robotics
Wrocław University of Science and Technology
ul.Janiszewskiego 11/17, 50-372 Wrocław, Poland

Abstract. Face similarity detection is a critical task in computer vision, particularly for applications like facial recognition and verification. Training robust models for this task often requires large, diverse datasets, which can be costly and time-consuming to acquire. This paper explores the potential of Generative Adversarial Networks (GANs) to augment training data with synthetic face images. We compare three GAN-based approaches-StyleGAN2 with labels, StyleGAN2 without labels, and custom GAN models-and evaluate their impact on CNN model performance across three scenarios: mixed datasets (real + generated), generated-only datasets, and fine-tuned models (first trained on real data, then fine-tuned with generated images). Our results show that while integrating GAN-generated data can enhance model performance in certain configurations, particularly when combined with real data, it may also lead to degradation. Notably, only two out of nine evaluated scenarios outperformed the baseline model trained solely on real data. StyleGAN2 without labels achieved the highest accuracy (87.90%) in the mixed-data setting and 87.06% with synthetic-only training, whereas fine-tuning the real-data model with these synthetic images produced the lowest accuracy (81.79%). These findings suggest that a balanced approach to integrating synthetic and real data is crucial for optimal performance.

Key words: face similarity, GAN, CNN, data augmentation

1. INTRODUCTION

In the previous work, [1] we systematically investigated how the structure and arrangement of input face pairs (e.g. horizontal/vertical concatenation) influence the performance of a CNN-based face similarity classifier. The objective of our CNN-based classifier was to determine whether a pair of input face images depicted the same individual (same-person class) or different individuals (different-person class). We proposed four different image concatenation schemas: horizontal and vertical concatenation, and intertwining by row and by column. As shown in [2], leveraging GAN-generated synthetic data can mitigate overfitting and improve generalization, which aligns with our motivation to extend the ORL dataset [3] beyond traditional augmentation techniques. The model was trained on a relatively small dataset, ORL [3]. Naturally, we wanted to investigate whether further extension of the data set will increase the performance of the model. There are a number of data augmentation techniques that can extend the original dataset; however, they do not create unique samples. We decided to investigate the possibility of using GAN network architecture in order to extend the initial dataset in addition to using common 2D image augmentation techniques. This approach mirrors studies like [4], where GAN-based synthetic data training (GAN-ST) demonstrated superior performance over conventional augmentation, particularly for lightweight CNNs. Systematic evaluations of augmentation for facial tasks [5] and contemporary GAN-based face-recognition models such as DSU-GAN [6] confirm that data scarcity remains the central bottleneck in this domain, while recent surveys trace the maturation of deep-learning face recognition and the emerging role of synthetic data [7, 8, 9]. The scope of the research is bimodal. The first goal was to create our own GAN model capable of producing new samples of human faces. The second goal of the research was to determine if synthetic samples can be successfully used during the training process and lead to an increase in performance.

A large, diverse, and representative dataset is essential for training robust neural networks. To improve model accuracy, extending or augmenting the training dataset is often necessary, particularly when real-world data collection is limited by constraints such as cost, time, or ethical considerations. The methods by which an initial dataset can be extended are limited. Some problems allow for the collection of new samples that expand the covered problem space. However, as highlighted in [10], GANs offer a scalable solution for medical and non-medical applications alike, addressing critical data scarcity issues where traditional augmentation fails. However, real-world data collection often faces practical challenges, including ethical constraints, high costs, or logistical difficulties, rendering it impractical or infeasible in many scenarios. If no new samples can be collected, usually, depending on the nature of the process, they can be generated through simulation, and depending on the scale, this can also be expensive. For instance, [11] demonstrated that GAN-generated synthetic cardiomyocyte images significantly improved classification accuracy compared to real data alone, underscoring the potential of synthetic augmentation in domains with limited sample diversity. When all options are exhausted, dataset augmentation techniques are commonly used to inflate the initial dataset. In turn, some techniques are not suitable for a given problem, rendering the number of available methods smaller. In opposition to the aforementioned approaches, it is possible to generate

*e-mail: wojciech.domski@edu.pl

synthetic samples that should cover the entire state space while providing samples that are indistinguishable from real ones, as shown by comprehensive studies of generative-adversarial augmentation [12, 13, 14, 15, 16]. That is why we have decided to use a GAN approach to generate synthetic face images to extend the initial dataset.

There is one more open question related to usage of the generated synthetic samples. Since the synthetic samples extend the initial dataset, they might influence the overall performance of the target model. As [17] found, GAN-synthesized data not only improved classification in imbalanced datasets (e.g., C-shaped root canals) but also maintained visual fidelity, suggesting that synthetic augmentation preserves real-world applicability. However, the generated samples might be used during different stages of working on a model. Namely, the generated samples can extend the training or testing subset. The injection of new samples might be perceived as contamination of the dataset. This debate is echoed in [4], where GAN-ST models improved accuracy by covering diverse data distributions, though careful validation remains essential to avoid negative transfer effects. That is why there is a debate on whether to extend the dataset in such ways, and if so, which subset (training or testing) dataset should be extended or even replaced.

The structure of the paper is provided in the following order. First we discuss the dataset based on which the training of models was performed. Then we discuss the common image augmentation techniques that were used during the process of training the classifier. Next, we describe three different approaches that led to new samples generation based on GANs. Finally, we use differently augmented datasets to determine if the performance of our classifier has increased or not. At the end, we provide conclusions how the performance of the initial classifier was affected.

2. INPUT DATA

We selected the ORL database [3], which contains 400 face images from 40 distinct individuals, with each person having 10 photographs captured under varying lighting conditions and expressions. Some subjects wore glasses in certain images, while others did not, adding significant variability to the dataset characteristics. All images are grayscale with a resolution of 92x112 pixels. Operating on a single luminance channel rather than three RGB channels reduces the input dimensionality and the associated storage and computational cost by roughly one third, while preserving the structural facial features on which similarity decisions rely; recent work shows that face-recognition accuracy is essentially unchanged when colour information is discarded [18]. The images feature frontal poses with minimal rotation (maximum $\pm 6^{\circ}$), ensuring full facial visibility without significant angle variations. The preprocessing pipeline involved two crucial steps: first, cropping the original images from their borders to create square images of 92x92 pixels that better capture complete facial proportions while preserving structural details; second, scaling these down to the final resolution of 64x64 pixels. This two-stage scaling approach was found to be optimal, as it maintained sufficient detail for GAN training while significantly reducing computational requirements.

2.1. Training set curation

The preparation of training sets required careful adaptation based on the specific model architecture being employed. For StyleGAN2 models, we implemented two distinct approaches: a single labeled model capable of generating images across all individuals through conditional generation based on input labels, and 24 separate unlabeled models where each was specialized for generating high-fidelity images of only one specific person. Additionally, we developed and trained 24 custom GAN models using our proprietary architecture, with one dedicated model per individual to capture their unique facial characteristics. Training parameters were carefully optimized across all models, including consistent batch sizes of 16 and gamma values of 3.2768, with training durations varying significantly depending on the approach - up to 1000 epochs for the labeled StyleGAN2 model compared to fewer iterations for specialized models. Before their use in classifier training, the generated images were screened manually under human supervision. Samples were rejected when faces were visibly deformed or when strong image noise made the facial structure unreliable; no automated face-quality threshold was used. The entire training process was conducted using NVIDIA Quadro P4000 GPUs to maximize computational efficiency.

3. DATA AUGMENTATION

The data augmentation process is a common tool that is used to enlarge the initial database with new samples based on the original samples. When 2D images are being considered, a number of techniques can be deployed in order to extend the dataset. Following can be included: noise addition, rotation, scaling, contract adjustment, and many more. Example can be seen in the Fig. 1 showing comparison between original and augmented images with scaling. However, the main goal of this work was to investigate how the original model for face similarity detection will behave when artificial samples were used for training. To facilitate data generation, GAN models were trained and used to generate new samples.

Three main approaches were considered and implemented:

- StyleGAN2 model with labels,
- multiple StyleGAN2 models without labels,
- multiple custom GAN models.

StyleGAN2 model was trained to improve the quality of the images [19]. This model was trained on data from the ORL database. However, this model is capable of consuming additional input that can guide it to generate a specific type of images. Therefore, we have introduced a differentiation, as mentioned earlier. The single StyleGAN2-based model was trained for 24 different labels (different human faces coming from the ORL database). In turn, there were 24 StyleGAN2-based models without labels. Each label-less model was capable of generating artificial photo of a single person. Finally, 24 custom GAN model were created and trained to generate artificial images of a given person's face. These models were deployed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.



(a) Original (b) After augmentation (64×64)

Fig. 1. Original ORL image and its augmented 64×64 counterpart, illustrating the level of detail retained after the two-stage scaling.

in order to augment original database containing faces. We used 24 identities to study synthetic augmentation of a face-similarity classifier trained from a deliberately limited set of subjects. The same 24 identities supplied the training data for all three GAN approaches and for the classifier training pairs. The remaining identities were allocated to validation, testing, or left unused as specified in Section 6. Consequently, the comparisons describe this selected identity split and do not constitute an evaluation over all 40 ORL identities.

4. CUSTOM GAN NETWORK ARCHITECTURE

Generative Adversarial Networks (GANs) represent a distinct paradigm in machine learning, specifically designed for synthetic data generation tasks [20]. Unlike traditional neural networks, GANs comprise two components: a generator and a discriminator, which compete adversarially to produce high-fidelity synthetic samples. In comparison to other types of neural networks GAN architecture composes of two architectures. One architecture, generator, is in charge of image generation task. Its purpose is to generate images that would be indistinguishable by a naked eye, thus appearing real and not generated. To accomplish this task, a second network is required. It is responsible for discriminating the generated images by the first network and decide if it is real or fake. The learning process involves in-loop generation and discrimination process, that leads to achieving better results. The both networks are trained in two stages loops. In the first stage, only the discriminator is being trained on a dataset that composes of real true samples and synthetic images produced by the generator. The second phases involves the generator training. In this step samples are generated by the generator. However, the labels are set to be real. During the generator training phase only its weight are affected while discriminator's weights are frozen.

GANs operate as adversarial systems, where the generator aims to produce realistic images while the discriminator strives to distinguish them from real data. It was shown that the further the zero-sum game advances, it might converge to a so-called Nash equilibrium state. In reference to GAN architecture it leads to state in which the generator tends to generate perfect synthetic samples. However, the downside of it is the reduction of diversity in generated images.

The generator architecture, depicted in Figure 2, consists of

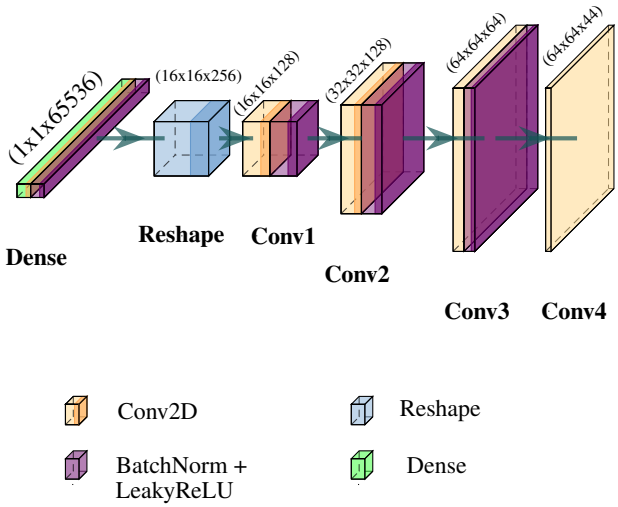


Fig. 2. GAN generator architecture

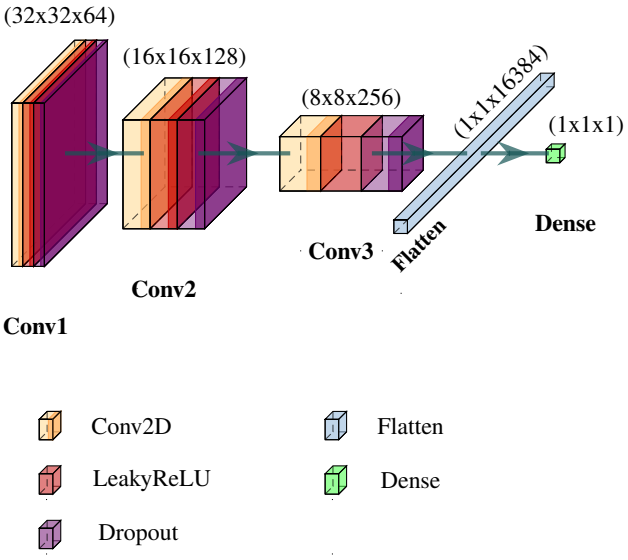


Fig. 3. GAN discriminator architecture

a minimalist design with just four convolutional layers. Despite its simplicity, this structure proved effective for generating high-quality synthetic face images. Additionally, to further improve learning process Leaky ReLU was used. This rectifier type of activation layer was proven to be well suited for image processing tasks [21], [22], [23], [24]. Moreover, batch normalization layer was used to normalize output of each convolutional layer and prevent problems related to vanishing/exploding gradient, thus rendering optimization more challenging [25]. As the input the generator initially consumes random noise, and in the end the generator produces an artificial image of 64x64 shape.

The second network that makes up the GAN architecture is the discriminator. The task of the discriminator is to distinguish if the sample generated by the generator is real or fake.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

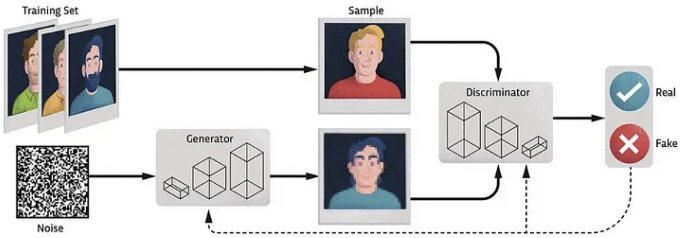


Fig. 4. GAN architecture [26]

The network architecture for the discriminator was presented in Fig. 3. The discriminator can be percieved as a classificator which task is to predict if the provided image is fake or real. This is achieved thanks to decoupled training process that involves two adversarial models – the generator and the discriminator.

Fig. 4 presents a graphical representation of the GAN architecture. The learning process involves generating increasingly accurate samples, which are then evaluated by the discriminator. At the beginning of training, the generator receives a random vector, which is then transformed into an image using a Gaussian distribution. During training, the generator attempts to deceive the discriminator by producing samples that increasingly resemble real data. The improvement in the quality of generated images is achieved by minimizing the loss function, which measures how effectively the generator fools the discriminator. The primary goal of the generator is to produce artificial samples that are so realistic that the discriminator is unable to determine whether a given sample is fake or real.

5. PERFORMANCE ANALYSIS

In order to compare different models we have decided to use two metrics. Accuracy quantifies the overall correctness of a model and is computed from the true-positive (TP, faces belonging to the same person), true-negative (TN, real samples of two different people), false-positive (FP) and false-negative (FN) counts. On the other hand, FP defines the number of samples that were classified as belonging to the same person while in fact belonging to two different people. Analogically, FN represents samples that were misclassified as two different people while in fact representing the same person. The higher the accuracy the better model performs.

To further analyse how trained models perform, we calculated recall and precision for each model. Precision measures the proportion of positive (same-person) decisions that are correct: $TP/(TP+FP)$, where FP denotes different-person pairs incorrectly classified as same-person pairs. Recall measures the proportion of actual same-person pairs that are correctly identified: $TP/(TP+FN)$, where FN denotes same-person pairs incorrectly classified as different-person pairs. In every confusion matrix below, rows represent the actual different-person and same-person classes, in that order, while columns represent the predicted different-person and same-person classes. The upper-right entry is therefore FP, and the lower-left entry is FN; the displayed layout is $\left[\begin{smallmatrix} TN & FP \\ FN & TP \end{smallmatrix}\right]$. Observing how precision and recall behave in a function of assumed threshold allows us to decide which behaviour of the classifier is more desirable. Depending on the selected threshold we can decide if the model should tend to prefer strong cases while some positive samples might not be detected over more robust positive classification with some false positives.

6. EXPERIMENTAL RESULTS AND DISCUSSION

The data allocation was fixed at the identity level before GAN or classifier training. The 24 ORL identities s1–s10 and s27–s40 were assigned to GAN training and were also used to construct the real and synthetic training pairs for the face-similarity classifier. Four different identities, s11–s14, were reserved for classifier validation, and another four, s15–s18, were reserved exclusively for final testing; identities s19–s26 were not used. The standard real pair datasets contained 15120 training, 2520 validation, and 2520 test samples, corresponding to 75%/12.5%/12.5%, or 87.5% development data when training and validation are combined versus 12.5% final test data. Consequently, there was no data pollution of the test set because neither its identities nor their images were used for GAN training or classifier development. This identity allocation applies to the real-only baseline and all nine GAN-augmented classifier scenarios reported below. All ten reported confusion matrices use the same 2520 real-image test pairs from s15–s18, with 1260 same-person and 1260 different-person pairs. Generator training and synthetic classifier-training pairs use only the 24 training identities; validation pairs use only s11–s14. The FID/IS measurements below are separate image-quality comparisons for a training identity and do not use the held-out classifier-test identities.

Figures 5 and 6 show representative accepted images generated by StyleGAN2 and the custom GANs. Image filtering was a manual, human-supervised process: images containing visibly deformed faces or strong noise were excluded before classifier training. Approximately 50% of custom-GAN samples and 35% of StyleGAN2 samples were rejected.

After successful implementation of the GAN model it was used to generate synthetic face images that were used during the training. In order to compare the efficiency of the proposed solution we had conducted several scenarios. Furthermore, to establish a base line to which trained models could be compared to we decided to use face similarity classifier trained only on real images. The accuracy of the model achieved 83.93% on a testing dataset also composed of real images. The confusion matrix for this model was presented in (1).

$$\begin{bmatrix} T_n & F_p \\ F_n & T_p \end{bmatrix} = \begin{bmatrix} 1238 & 22 \\ 383 & 877 \end{bmatrix} \tag{1}$$

For deeper analysis, Figure 7 illustrates the trade-off between precision and recall as a function of the decision threshold, revealing how model behavior shifts with varying sensitivity settings. When setting the threshold to 0.23 the model has 89.1% precision and recall. For the threshold set to 0.5 the precision and recall are 97.55% and 69.60%, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.



Fig. 5. Sample, pre-selected image generated by StyleGAN2 with labels model



Fig. 6. Samples, pre-selected images generated by custom GAN models

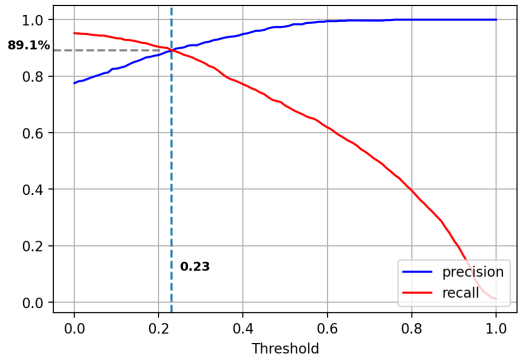


Fig. 7. Precision vs. recall for model trained only on real data

As mentioned earlier, to compare performance of each model a couple of scenarios were created.

Our experimental evaluation encompassed nine distinct scenarios, systematically comparing three GAN-based data augmentation strategies (StyleGAN2 with/without labels, custom GAN) across three training paradigms: mixed real-synthetic datasets, synthetic-only datasets, and fine-tuning on synthetic data post-real-data pretraining.

- mixed dataset based on **StyleGan2 with labels**,
- mixed dataset based on **StyleGan2 without labels**,
- mixed dataset based on **custom models**,
- synthetic-only dataset based on **StyleGan2 with labels**,
- synthetic-only dataset based on **StyleGan2 without labels**,
- synthetic-only dataset based on **custom models**,
- tuning with synthetic images based on **StyleGan2 with labels**,
- tuning with synthetic images based on **StyleGan2 without labels**,
- tuning with synthetic images based on **custom models**,

Three GAN models were used to generate synthetic images of human faces: StyleGan2 with labels, StyleGan2 without labels and custom models. Custom models similar in behaviour to StyleGan2 without labels since for each person images a separately trained model was created. There were three different approaches that allowed us to investigate how synthetic images influence resulting accuracy of the face similarity model. First dataset was consisting of mixed images, real images coming from ORL database after the augmentation process and images generated by GAN model. Second approach involved a dataset solely based on synthetic images generated with the help of GAN architecture. Finally, the last approach was similar to the first one (mixed dataset) but decoupled into two stages. First stage involved training the model on real images. This model was shared for the second stage where it was tuned using generated images by the GAN network. Here a classifier sample is a pair of face images, so the following counts are numbers of pairs, not numbers of distinct faces. Each of the three generator-specific synthetic training sets contained 15120 pairs (7560 same-person and 7560 different-person). The real-only reference used the standard real training set of 15120 pairs; a noise-augmented real training set contained 30240 pairs. In the mixed setting, the labeled StyleGAN2 and custom-GAN configurations used 15120 real plus 15120 synthetic pairs (30240 total); the label-free StyleGAN2 mixed configuration used 30240 noise-augmented real plus 15120 synthetic pairs (45360 total). Synthetic-only training used 15120 synthetic pairs and no real training pairs for each generator. In the fine-tuning setting, a classifier first trained on 15120 standard real pairs (30240 noise-augmented real pairs in the label-free StyleGAN2 path) was subsequently trained on 15120 synthetic pairs from the selected generator; these are consecutive training stages, not a pooled mixed set. The standard validation and final-test sets each contained 2520 real pairs.

For each scenario we provide a separate confusion matrix for the same testing dataset that was used to establish the reference

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

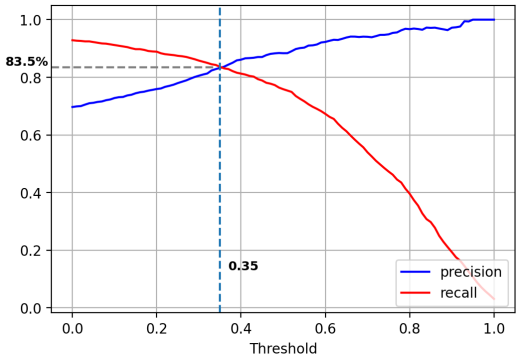


Fig. 8. Precision vs. recall for StyleGan2 with labels trained on mixed dataset

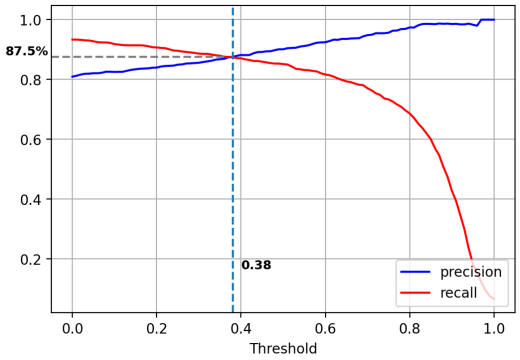


Fig. 9. Precision vs. recall for StyleGan2 without labels trained on mixed dataset

(1).

Confusion matrix for the first scenario was presented in (2). It shows results obtained for mixed training dataset that was extended with synthetic images using StyleGan2 with labels network.

$$\begin{bmatrix} T_n & F_p \\ F_n & T_p \end{bmatrix} = \begin{bmatrix} 1134 & 126 \\ 304 & 956 \end{bmatrix} \tag{2}$$

Moreover, we present recall vs precision in function of threshold in Fig. 8 for better analysis of examined case. When setting the threshold to 0.35 the model has 83.5% precision and recall. For the threshold set to 0.5 the precision and recall are 88.35% and 75.87%, respectively.

The confusion matrixes with results using StyleGan2 without labels networks are presented in (3). In turn, the extended dataset but using our custom GAN model is in (4). In this case the best performance is offered when extending image dataset using StyleGan2 without labels. Accuracy levels when considering StyleGan2 with labels and custom GAN models is comparable.

$$\begin{bmatrix} T_n & F_p \\ F_n & T_p \end{bmatrix} = \begin{bmatrix} 1142 & 118 \\ 187 & 1073 \end{bmatrix} \tag{3}$$

Moreover, we present recall vs precision in function of threshold in Fig. 9 for better analysis of examined case. When setting the threshold to 0.38 the model has 87.5% precision and recall. For the threshold set to 0.5 the precision and recall are 90.09% and 85.16%, respectively.

$$\begin{bmatrix} T_n & F_p \\ F_n & T_p \end{bmatrix} = \begin{bmatrix} 1037 & 223 \\ 228 & 1032 \end{bmatrix} \tag{4}$$

Moreover, we present recall vs precision in function of threshold in Fig. 10 for better analysis of examined case. When setting the threshold to 0.47 the model has 82.1% precision and recall. For the threshold set to 0.5 the precision and recall are 82.23% and 81.90%, respectively.

Next three scenarios involve only training on synthetic images. Confusion matrix face similarity classifier using synthetic images created with StyleGan2 with labels were presented in (5). Similar confusion matrix but for StyleGan2 without labels was presented in (6).

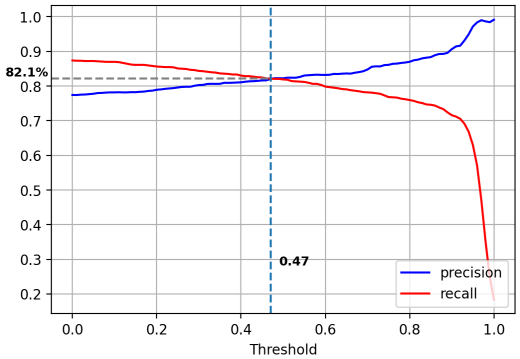


Fig. 10. Precision vs. recall for custom GAN trained on mixed dataset

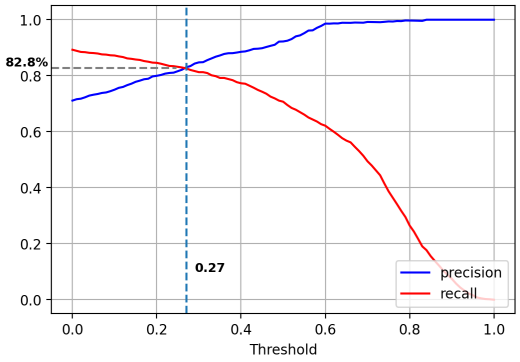


Fig. 11. Precision vs. recall for StyleGan2 with labels trained on synthetic dataset

$$\begin{bmatrix} T_n & F_p \\ F_n & T_p \end{bmatrix} = \begin{bmatrix} 1185 & 75 \\ 369 & 891 \end{bmatrix} \tag{5}$$

Moreover, we present recall vs precision in function of threshold in Fig. 11 for better analysis of examined case. When setting the threshold to 0.27 the model has 82.8% precision and recall. For the threshold set to 0.5 the precision and recall are 92.24% and 70.71%, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

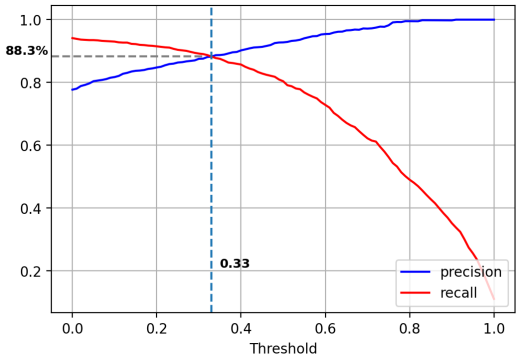


Fig. 12. Precision vs. recall for StyleGan2 without labels trained on synthetic dataset

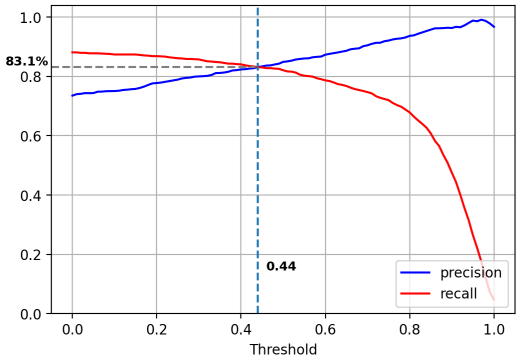


Fig. 13. Precision vs. recall for custom GAN trained on synthetic dataset

$$\begin{bmatrix} T_n & F_p \\ F_n & T_p \end{bmatrix} = \begin{bmatrix} 1181 & 79 \\ 247 & 1013 \end{bmatrix} \tag{6}$$

Moreover, we present recall vs precision in function of threshold in Fig. 12 for better analysis of examined case. When setting the threshold to 0.33 the model has 88.3% precision and recall. For the threshold set to 0.5 the precision and recall are 92.77% and 80.40%, respectively.

Additionally, the same scenario (based only on synthetic images) when the images were generated using custom GANs were presented in (7).

$$\begin{bmatrix} T_n & F_p \\ F_n & T_p \end{bmatrix} = \begin{bmatrix} 1076 & 184 \\ 226 & 1034 \end{bmatrix} \tag{7}$$

Moreover, we present recall vs precision in function of threshold in Fig. 13 for better analysis of examined case. When setting the threshold to 0.44 the model has 83.1% precision and recall. For the threshold set to 0.5 the precision and recall are 84.89% and 82.06%, respectively.

Similarly to the first approach also here the best results are achieved for dataset generated using StyleGan2 without labels (6). The worst accuracy is offered by StyleGan2 with labels (5) while model trained on synthetic images generated using multiple custom GANs models performs slightly better (7).

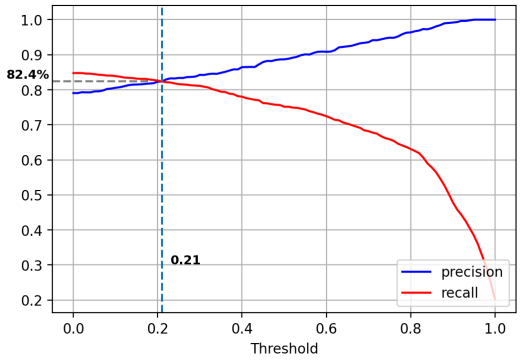


Fig. 14. Precision vs. recall for StyleGan2 with labels tuned with synthetic dataset

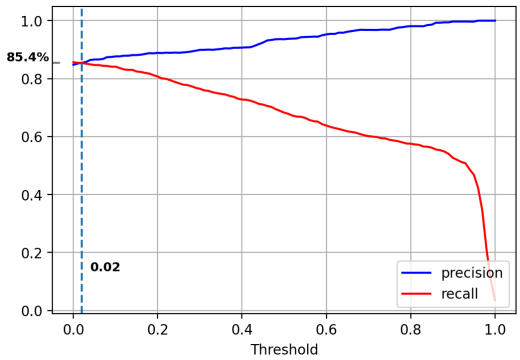


Fig. 15. Precision vs. recall for StyleGan2 without labels tuned with synthetic dataset

The third scenario where initial model was trained on real images and then tuned with synthetic images generated using different GANs was also conducted. The results for variant when synthetic images were generated using StyleGan2 with labels is in (8). The confusion matrix for StyleGan2 without labels was presented in (9), while for synthetic images generated using our custom GAN model is in (10).

$$\begin{bmatrix} T_n & F_p \\ F_n & T_p \end{bmatrix} = \begin{bmatrix} 1139 & 121 \\ 313 & 947 \end{bmatrix} \tag{8}$$

Moreover, we present recall vs precision in function of threshold in Fig. 14 for better analysis of examined case. When setting the threshold to 0.21 the model has 82.4% precision and recall. For the threshold set to 0.5 the precision and recall are 88.67% and 75.16%, respectively.

$$\begin{bmatrix} T_n & F_p \\ F_n & T_p \end{bmatrix} = \begin{bmatrix} 1201 & 59 \\ 400 & 860 \end{bmatrix} \tag{9}$$

Moreover, we present recall vs precision in function of threshold in Fig. 15 for better analysis of examined case. When setting the threshold to 0.02 the model has 85.4% precision and recall. For the threshold set to 0.5 the precision and recall are 93.58% and 68.25%, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

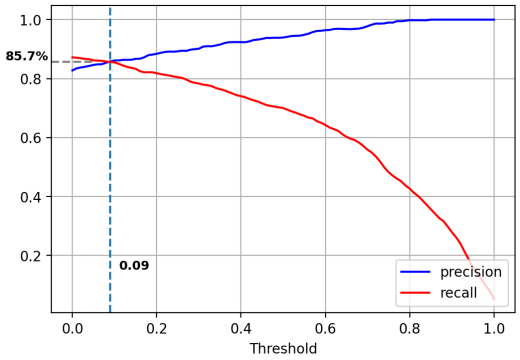


Fig. 16. Precision vs. recall for custom GAN tuned with synthetic dataset

$$\begin{bmatrix} T_n & F_p \\ F_n & T_p \end{bmatrix} = \begin{bmatrix} 1203 & 57 \\ 378 & 882 \end{bmatrix} \tag{10}$$

Moreover, we present recall vs precision in function of threshold in Fig. 16 for better analysis of examined case. When setting the threshold to 0.09 the model has 85.7% precision and recall. For the threshold set to 0.5 the precision and recall are 93.93% and 70.00%, respectively.

For approach when the classifier was trained in two stages all models offer similar accuracy. However, model trained with images generated by StyleGan2 without labels offers the worst performance, while the other two models are comparable.

To further analyse the performance of trained face similarity model and influence of different approaches related to generating synthetic images, the results were collected in Tab. 1. The obtained accuracy needs to be compared with the native face similarity network that achieved 83.93%. Using it as a reference base it is easy to notice that only 2 out of 9 models in all examined scenarios were performing better than a network trained solely on real images.

The best model for generating synthetic images was StyleGan2 without labels for mixed dataset. Using data generated with this model the classifier was able to achieve 5% better results when compared to other synthetic image generators. However, when last approach, namely tuned training is considered, the classifier scored the lowest performance among other synthetic image generators.

Based on the accuracy scores collected in Tab. 1 additional conclusions can be made. The slightly worse model that achieves 87.06%, less then 1% less from the best performing model, was model trained on data generated by StyleGan2 without labels. However, during the training only synthetic images were used in order to train the model.

The most stable models were achieved using data generated with StyleGan2 with labels GAN. The average accuracy is on level of 82.7% while differences between using different scenarios are less than ±0.5%. Other models trained using data generated with StyleGan without labels and ours have higher dispersion.

In Tab. 2 results for each model in terms of equilibrium between precision and recall were collected. Compared to

	Model		
	StyleGan2 with labels	StyleGan2 without labels	custom model
mixed dataset	82.94%	87.90%	82.10%
synthetic only	82.38%	87.06%	83.73%
tuned training	82.78%	81.79%	82.74%

Table 1. Accuracy across GAN models and different approaches

	Model		
	StyleGan2 with labels	StyleGan2 without labels	custom model
mixed dataset	83.5%	87.5%	82.1%
synthetic only	82.8%	88.3%	83.1%
tuned training	82.4%	85.4%	85.7%

Table 2. Precision/recall equilibrium across GAN models and different approaches

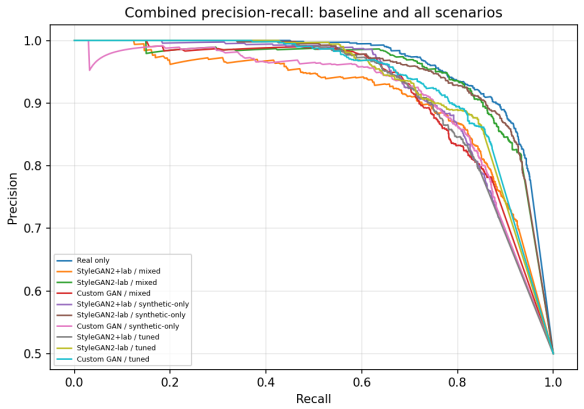


Fig. 17. Consolidated comparison of all configurations, precision–recall curves.

accuracy-oriented results also in this case the model, where StyleGan2 without labels was used to generate synthetic images was the best performing one. However, the model trained solely on synthetic generated data has the highest precision/recall score, 88.3% compared to model trained on mixed dataset that achieved 87.5% precision.

Based on further analysis of figures showing precision/recall curves it can be noticed that for the models tuned on synthetic data, generated using our custom GAN model and StyleGan2 without labels, the balance threshold is unusually low. It is between 2% and 9%. For other models the threshold is always greater than 21% but less than 47%. This suggests that such models exhibit an exaggerated sensitivity to positive predictions, potentially due to over-reliance on synthetic data patterns during fine-tuning.

For an at-a-glance comparison across all configurations,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

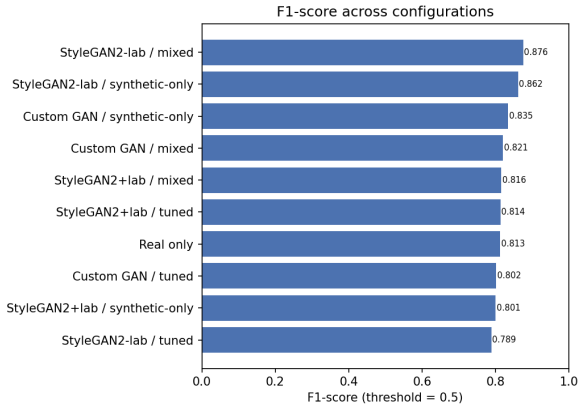


Fig. 18. Consolidated comparison of all configurations, F1-score ranking.

Generator	FID (↓)	IS (↑)
StyleGAN2 without labels	212.62	2.52 ± 0.10
Custom GAN (labeled s1 run)	213.35	2.54 ± 0.11

Table 3. Identity s1 FID and IS measurements for unfiltered synthetic images (1024 label-free StyleGAN2 and 1000 labeled custom-GAN samples, each compared with the same 10 real images; no aggregation across identities).

Method	Reported accuracy	Setting
ArcFace [27]	LFW 99.83%	large-scale, RGB
AdaFace [28]	LFW 99.82%	large-scale, RGB
Proposed (best, mixed)	ORL 87.90%	small-scale, grayscale

Table 4. Positioning of the proposed approach with respect to recent state-of-the-art face-recognition backbones.

Fig. 17 shows a view on the precision–recall curves of the real-data baseline and the nine augmentation scenarios on a single plot. While Fig. 18 presents an F1-score bar chart at a 0.5 decision threshold that ranks all configurations. Together they complement the per-model curves presented above.

6.1. Image quality of the synthetic data

To assess the generated faces quantitatively, Frechet Inception Distance (FID) and Inception Score (IS) were computed using a pretrained torchvision InceptionV3 model. Grayscale images were replicated across three channels, resized to 299×299 pixels, and normalized with ImageNet statistics. FID used the 2048-dimensional pooled features, while IS used the 1000-class softmax output with ten splits. The logged comparisons used one identity at a time: ten real ORL images as the reference and 1024 unfiltered images from the label-free StyleGAN2 model or 1000 unfiltered images from the custom GAN. Lower FID and higher IS are preferable. The results are summarised in Table 3. Both rows of Table 3 concern identity s1 and are individual identity-level comparisons; they are not averages over the 24 training identities. The conditional StyleGAN2 model is included in the downstream classifier experiments but is outside this two-run image-quality comparison, which reports the documented single-identity runs for label-free StyleGAN2 and a labeled custom-GAN sample set. The custom-GAN row is not an aggregate of the 24 identity-specific custom models. A matched conditional-model FID/IS estimate is therefore not reported, and the table does not establish whether conditioning improves or worsens image quality. The uncertainty shown for IS is the standard deviation across ten splits of each fixed synthetic set, not across identities or independent generations.

The FID values are high for both measured cases (above 200), and the IS values are similar (≈ 2.5). Because each FID reference contains only ten real images and the ImageNet feature space is not specialized for grayscale face identity, these values should be interpreted as descriptive identity-level measurements rather than precise estimates for the complete generators. Repeating the calculation on the same fixed image sets and with the same deterministic preprocessing does not provide an independent estimate; meaningful uncertainty would require new generated sets or additional identities. The small difference between the two FID values therefore does not support ranking the generators.

6.2. Comparison with state-of-the-art

To clarify the contribution of this work, Table 4 positions the proposed approach with respect to two recent state-of-the-art face-recognition backbones, ArcFace [27] and AdaFace [28]. These methods are trained on orders-of-magnitude larger, colour datasets and target large-scale recognition, whereas our pipeline addresses the small-data, grayscale regime; they are therefore reported for context rather than for head-to-head comparison.

7. CONCLUSIONS

This study systematically evaluates whether augmenting CNN-based face similarity classifiers with GAN-generated synthetic data can enhance performance, particularly under constrained real-data scenarios. Building upon our prior work, we investigate this question through rigorous experimentation involving three GAN variants and multiple training protocols. Our experiments involved three types of GANs: StyleGAN2 with labels, StyleGAN2 without labels, and custom GAN models- and three training scenarios: mixed datasets (real + generated), generated-only datasets, and fine-tuned models (trained on real data followed by fine-tuning with generated images).

Our key findings reveal that while the integration of GAN-generated data can improve model performance in certain configurations – particularly when combined with real data – other training schedules may lead to degradation. Specifically, StyleGAN2 without labels achieved 87.90% accuracy in the mixed-data setting and 87.06% with synthetic-only training, whereas fine-tuning the real-data model with these generated images produced the lowest accuracy (81.79%). This suggests that a balanced approach, where real and synthetic data are combined judiciously, is essential for optimal performance.

Additionally, our analysis of precision-recall trade-offs indicated that models fine-tuned with generated data often ex-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

hibited lower balance thresholds (between 2% and 9%), suggesting a bias towards positive predictions. In contrast, other models maintained higher thresholds (above 21%), indicating more balanced behavior.

These results contribute to the growing body of research on leveraging synthetic data for training deep learning models, particularly in scenarios where real data is scarce or expensive to collect. However, they also highlight the challenges and limitations associated with this approach, such as the potential for overfitting to synthetic artifacts or loss of generalization to real-world variations.

A critical reading of these results tempers an unconditionally positive interpretation. Only two of the nine configurations surpassed the real-data baseline (83.93%), and several synthetic-only configurations underperformed it; the fine-tuned, label-free StyleGAN2 model in fact produced the lowest accuracy (81.79%). Moreover, the unusually low balance thresholds (2–9%) observed for the fine-tuned models indicate instability in their decision boundaries. These observations are consistent with the broader literature on synthetic data for face recognition, which emphasises that the benefit of synthetic samples is highly configuration-dependent and can introduce negative transfer if not carefully controlled [7, 8].

Guided by this discussion, we identify exactly two directions for further research. First, hybrid, dynamically balanced real/synthetic training schedules that adapt the mixing ratio during training could mitigate the observed trade-offs and stabilise the decision threshold [13, 12]. Second, replacing the GAN generators with more recent, higher-diversity generative models – for example diffusion-based face synthesis – may yield synthetic samples whose distribution better matches real data, reducing over-reliance on synthetic artefacts [16, 7].

Training models exclusively on synthetic data introduces vulnerabilities, including potential overfitting to GAN artifacts or loss of generalization to real-world variations. This can result in inconsistent performance and reduced robustness compared to models trained on balanced (real + synthetic) datasets. While GAN-generated synthetic data holds promise for enhancing model performance in face similarity detection tasks, its integration must be carefully managed to avoid potential pitfalls and ensure robust generalization.

Credit authorship contribution statement

Wojciech Domski: Conceptualization, Methodology, Software, Implementation, Formal analysis, Investigation, Validation, Visualization, Writing - Original draft, Writing - Reviewing and Editing.

Adam Jankowiak: Software, Implementation, Visualization, Data curation, Validation, Writing - Original draft, Writing - Reviewing and Editing.

REFERENCES

[1] W. Domski and A. Jankowiak, "Convolution Neural Network for Face Similarity Estimation," *Journal of Automation, Mobile Robotics and Intelligent Systems*, vol. 19, no. 1, p. 65–72, Mar. 2025. [Online]. Available: https://doi.org/10.14313/jamris-2025-007

[2] H. Giedra and G. Vdoviak, "Leveraging generative adversarial networks to improve training image dataset," *New Trends in Computer Sciences*, vol. 2, no. 1, pp. 31–45, 2024. [Online]. Available: https://doi.org/10.3846/ntcs.2024.20515

[3] AT&T Laboratories Cambridge. (2023) The database of faces. https://cam-orl.co.uk/facedatabase.html.

[4] I. H. Rather and S. Kumar, "Generative adversarial network based synthetic data training model for lightweight convolutional neural networks," *Multimed Tools Appl*, vol. 83, pp. 6249–6271, 2023. [Online]. Available: https://doi.org/10.1007/s11042-023-15747-6

[5] S. Porcu, A. Floris, and L. Atzori, "Evaluation of Data Augmentation Techniques for Facial Expression Recognition Systems," *Electronics*, vol. 9, no. 11, p. 1892, 2020. [Online]. Available: https://doi.org/10.3390/electronics9111892

[6] D. Lin, H. Wang, X. Lei, W. Min, C. Yao, Y. Zhong, and Y. L. Guan, "DSU-GAN: A robust frontal face recognition approach based on generative adversarial network," *Computer Vision and Image Understanding*, vol. 249, p. 104128, 2024. [Online]. Available: https://doi.org/10.1016/j.cviu.2024.104128

[7] F. Boutros, V. Štruc, J. Fierrez, and N. Damer, "Synthetic data for face recognition: Current state and future prospects," *Image and Vision Computing*, vol. 135, p. 104688, 2023. [Online]. Available: https://doi.org/10.1016/j.imavis.2023.104688

[8] P. Melzi, R. Tolosana, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, I. De Andres-Tame, A. Morales, J. Fierrez, J. Ortega-Garcia *et al.*, "FRCSyn-onGoing: Benchmarking and comprehensive evaluation of real and synthetic data to improve face recognition systems," *Information Fusion*, vol. 107, p. 102322, 2024. [Online]. Available: https://doi.org/10.1016/j.inffus.2024.102322

[9] A. Zhalgas, B. Amirgaliyev, and A. Sovet, "Robust Face Recognition Under Challenging Conditions: A Comprehensive Review of Deep Learning Methods and Challenges," *Applied Sciences*, vol. 15, no. 17, p. 9390, 2025. [Online]. Available: https://doi.org/10.3390/app15179390

[10] Y. Heng, M. Yinghua, F. Khan, A. Khan, F. Ali, A. A. AlZubi, and H. Zeng, "Survey: application and analysis of generative adversarial networks in medical images," *Artificial Intelligence Review*, vol. 58, no. 39, pp. 1–81, 2024. [Online]. Available: https://doi.org/10.1007/s10462-024-10992-z

[11] Z. Wu, J. Park, P. R. Steiner, B. Zhu, and J. X. J. Zhang, "Generative adversarial network model to classify human induced pluripotent stem cell-cardiomyocytes based on maturation level," *Scientific Reports*, vol. 14, no. 27016, pp. 2045–2322, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-77943-0

[12] S. Sabnam and S. Rajagopal, "Application of generative adversarial networks in image, face reconstruction

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

and medical imaging: challenges and the current progress," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 12, no. 1, 2024. [Online]. Available: https://doi.org/10.1080/21681163.2024.2330524

[13] W. Zeng, "Image data augmentation techniques based on deep learning: A survey," *Mathematical Biosciences and Engineering*, vol. 21, no. 6, pp. 6190–6224, 2024. [Online]. Available: https://doi.org/10.3934/mbe.2024272

[14] M. K. Nallakaruppan, C. L. Chowdhary, S. K. Somayaji, H. Chaturvedi, R. Sujatha, H. T. Rauf, and M. Sharaf, "Comparative analysis of GAN-based fusion deep neural models for fake face detection," *Mathematical Biosciences and Engineering*, vol. 21, no. 1, pp. 1625–1649, 2024. [Online]. Available: https://doi.org/10.3934/mbe.2024071

[15] H. Zein, S. Chantaf, R. Fournier, and A. Nait-Ali, "Generative adversarial networks for anonymous acneic face dataset generation," *PLOS ONE*, vol. 19, no. 4, p. e0297958, 2024. [Online]. Available: https://doi.org/10.1371/journal.pone.0297958

[16] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, D. Lawatsch, F. Domin, and M. Schaubert, "GANDiff-Face: Controllable Generation of Synthetic Datasets for Face Recognition with Realistic Variations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2023, pp. 3086–3095. [Online]. Available: https://arxiv.org/abs/2305.19962

[17] S. Yang, K.-D. Kim, E. Ariji, N. Takata, and Y. Kise, "Evaluating the performance of generative adversarial network-synthesized periapical images in classifying C-shaped root canals," *Journal of Automation, Mobile Robotics and Intelligent Systems*, vol. 13, no. 18038, pp. 1–10, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-45290-1

[18] A. Bhatta, D. Mery, H. Wu, J. Annan, M. C. King, and K. W. Bowyer, "What's color got to do with it? Face recognition in grayscale," *arXiv preprint arXiv:2309.05180*, 2023. [Online]. Available: https://arxiv.org/abs/2309.05180

[19] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. CVPR*, 2020. [Online]. Available: https://arxiv.org/abs/1912.04958

[20] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[21] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Gordon, D. Dunson, and M. Dudík, Eds., vol. 15. Fort Lauderdale, FL, USA: PMLR, 11–13 Apr 2011, pp. 315–323. [Online]. Available: https://proceedings.mlr.press/v15/glorot11a.html

[22] Z. Hu, Y. Li, and Z. Yang, "Improving convolutional neural network using pseudo derivative relu," in *2018 5th International Conference on Systems and Informatics (ICSAI)*, 2018, pp. 283–287. [Online]. Available: https://doi.org/10.1109/ICSAI.2018.8599372

[23] S. A. Nabi and D. Revathi, "Improved cnn based feature extraction for data mining," in *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, 2022, pp. 1–8. [Online]. Available: https://doi.org/10.1109/ACCAI53970.2022.9752600

[24] R. R. Akram, M. Dwi Sulistiyo, A. Firman Ihsan, P. Eko Yunanto, D. Richasdy, and M. Arzaki, "Exploring relu activation functions in cnn for handwritten sundanese script recognition," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*, 2024, pp. 92–98. [Online]. Available: https://doi.org/10.1109/ICoDSA62899.2024.10652217

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015. [Online]. Available: https://arxiv.org/abs/1502.03167

[26] D. P. Bentley, "Visualization of the GAN network construction concept," https://www.sciencefocus.com/future-technology/how-do-machine-learning-gans-work, 2023.

[27] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694. [Online]. Available: https://doi.org/10.1109/CVPR.2019.00482

[28] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality Adaptive Margin for Face Recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. [Online]. Available: https://arxiv.org/abs/2204.00964

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.