



OPEN Facial identity recognition using StyleGAN3 inversion and improved tiny YOLOv7 model

Akhil Kumar¹, Swarnava Bhattacharjee², Amrisha Kumar¹ & Dushantha Nalin K. Jayakody^{3,4,5}✉

Facial identity recognition is one of the challenging problems in the domain of computer vision. Facial identity comprises the facial attributes of a person's face ranging from age progression, gender, hairstyle, etc. Manipulating facial attributes such as changing the gender, hairstyle, expressions, and makeup changes the entire facial identity of a person which is often used by law offenders to commit crimes. Leveraging the deep learning-based approaches, this work proposes a one-step solution for facial attribute manipulation and detection leading to facial identity recognition in few-shot and traditional scenarios. As a first step towards performing facial identity recognition, we created the Facial Attribute Manipulation Detection (FAM) Dataset which consists of twenty unique identities with thirty-eight facial attributes generated by the StyleGAN3 inversion. The Facial Attribute Detection (FAM) Dataset has 11,560 images richly annotated in YOLO format. To perform facial attribute and identity detection, we developed the Spatial Transformer Block (STB) and Squeeze-Excite Spatial Pyramid Pooling (SE-SPP)-based Tiny YOLOv7 model and proposed as FIR-Tiny YOLOv7 (Facial Identity Recognition-Tiny YOLOv7) model. The proposed model is an improvised variant of the Tiny YOLOv7 model. For facial identity recognition, the proposed model achieved 10.0% higher mAP in the one-shot scenario, 30.4% higher mAP in the three-shot scenario, 15.3% higher mAP in the five-shot scenario, and 0.1% higher mAP in the traditional 70% – 30% split scenario as compared to the Tiny YOLOv7 model. The results obtained with the proposed model are promising for general facial identity recognition under varying facial attribute manipulation.

Keywords Facial identity recognition, Facial attribute manipulation, Face detection, Tiny YOLOv7, StyleGAN3

Facial identity is a way of recognizing a person's identity by using his facial features. Facial identity recognition has always been a demanding area in the domain of computer vision. Although, it has several challenges due to the varying complexities of the facial attributes, however, in recent years several groundbreaking works^{1,2} have been proposed to deal with the arising challenges of complexities associated with facial features. With the growing data especially, the data generated from social media platforms, digital imaging, and surveillance systems, the necessity of accurately analyzing facial attribute manipulation and detection has become important in facial identity recognition in entertainment³ and security and law enforcement⁴. Facial attribute manipulation is the process of altering the facial identity of a person by manipulation of specific features or attributes of a person's face. Facial attribute manipulation is generally performed in images or videos to generate an unrealistic or plausible identity for a person. In this process, specific facial features such as age progression, gender swapping, and facial expressions are manipulated to generate a new facial identity for a person⁵. In the past, facial attribute manipulation was confined to science fiction, however, with the advancements in deep learning, it has become increasingly achievable⁶. However, in contrast to facial identity recognition and facial attribute manipulation, facial attribute detection deals with the identification and classification of facial attributes such as age, gender, facial expressions, and other facial landmarks. The application of facial attribute manipulation and detection in facial identity recognition holds immense potential in the entertainment industry, forensic analysis, and law enforcement.

¹School of Computer Science Engineering and Technology, Bennett University, Greater Noida, India. ²Liverpool John Moores University, Liverpool, England. ³COPELABS, Lusófona University, Lisboa, Portugal. ⁴Center of Technology and Systems (UNINOVA-CTS) and Associated Lab of Intelligent Systems (LASI), 2829-516 Caparica, Portugal. ⁵CIET/DEEE, Faculty of Engineering, Sri Lanka Institute of Information Technology, 10115 Malabe, Sri Lanka. ✉email: dushantha.jayakody@ulusofona.pt

The recent progress in deep learning-based approaches such as image editing using Generative Adversarial Networks (GANs)⁷ and object detection using YOLO (You Only Look Once)⁸ has shown immense potential in several downstream tasks^{9–11}. The advent of generative models has propelled image generation, particularly in human face synthesis, to new levels of realism. However, the challenge of preserving identity amidst facial attribute manipulation still exists. Object detection models like YOLO have demonstrated their potential in precisely detecting objects in various scenarios. Furthermore, the integration of generative models with object detection models such as YOLO is yet to be explored in the domain of facial attribute manipulation and detection for facial identity recognition due to several challenges associated with (1) Face attribute manipulation and identity preservation; (2) Scarcity of synthetic face datasets; (3) Synthetic face detection and; (4) Leveraging few-shot detection in scarce facial images scenarios. To address these challenges, several works have been proposed in recent years. A brief outline of the related work in regard to the specified challenges is presented below.

1) Face attribute manipulation and identity preservation The domain of face attribute manipulation and identity preservation presents daunting challenges. To address this challenge, the researchers^{12,13} have employed identity losses to preserve identity during face style manipulation. The works^{14–20} specify that maintaining identity becomes increasingly difficult when multiple facial attributes are altered at one time. Moreover, the researchers^{21–29} highlighted the challenge of maintaining the balance between the degree of facial attribute manipulation and facial identity preservation while performing facial identity recognition in manipulated faces scenarios.

2) Synthetic face dataset The usage of Generative Adversarial Networks (GANs) and plausible images in face detection and recognition has made significant improvements, highlighting their multifaceted contributions. The research work³⁰ focused on the complexities of face recognition in unconstrained environments, employing the Multi-Factor Joint Normalization Network (MFJNN) to synthesize multi-factor normalization while crucially preserving identity information, especially beneficial for addressing large pose changes. Moreover, the work³¹ proposed DigiFace-1 M consisting of a synthetic dataset crafted through a computer graphics pipeline, showcasing a remarkable 52.5% reduction in error rates on the LFW³² dataset. The authors utilized aggressive data augmentation strategies to bridge the gap between synthetic and real images. The authors³³ conducted benchmarking of face recognition systems using the syn-multi-PIE dataset generated via StyleGAN2 inversion, achieving an error rate below 5% on the original protocol, thus emphasizing the importance of preserving identity information within synthetic data. Furthermore, the work³⁴ highlighted the efficacy of synthetic face datasets, particularly SFace, generated through class-conditional GANs for privacy-centric applications, attaining high verification accuracies of 91.87% with multi-class and 99.13% with a combined strategy on the LFW dataset. Collectively, these studies underscore the increasing significance and effectiveness of synthetic datasets and GAN-generated images in tackling various challenges in face detection, recognition, privacy preservation, and bridging domain gaps between synthetic and real data.

3) Synthetic face detection Recent studies across various domains have made significant progress in the detection and identification of synthetic images. The authors³⁵ developed a sophisticated detector tailored specifically to identify synthetic images generated from StyleGAN3. They employed an ensemble of CNNs and achieved an AUC of 99.95% on a combined test set comprising FFHQ, AFHQ, and Metfaces datasets. Additionally, the work³⁶ introduced an Eyes-Based Siamese Neural Network, utilizing semantic-based methodologies to scrutinize inter-eye symmetries and inconsistencies in synthetic images generated from ProGAN, StyleGAN2, and StyleGAN3. Their approach demonstrated an AUC exceeding 99% with the Xception-based model, highlighting the robustness of their technique in discriminating synthetic content. Furthermore, the researchers in³⁷ conducted a comprehensive comparative analysis of GANs for synthetic image detection, and training on StyleGAN2, and achieved exceptional accuracies of 99.7% for high-resolution FFHQ and 99.9% for low-resolution images, emphasizing the efficacy of GAN-specific models in detecting synthetic content. Moreover, the work³⁸ explored the problem of synthetic face identification and biometrics by employing GAN inversion and classification methodologies to attain a global accuracy above 88%, specifically on FFHQ and other datasets, showcasing the potential for accurate identification of synthetic faces through intricate classification strategies. These collective efforts highlight the progression in detecting and identifying synthetic imagery, crucial for ensuring authenticity and integrity in digital image analysis and recognition systems.

4) Few-shot object detection scenarios Recent progress in Few-Shot Object Detection (FSOD) has witnessed significant advancements, with YOLO detectors playing an important role. The work³⁹ introduced the Meta-YOLO framework, which integrates a Feature Decorrelation Module (FDM) and a three-head module, resulting in an impressive 39.8% Mean Average Precision (mAP) improvement in few-shot scenarios. The researchers⁴⁰ explored content assessment and detection of various objects, including faces and firearms, across multiple social media platforms. They achieved an 80.39% mAP at IoU 0.5 and a 35.22% mAP at IoU 0.50:0.95 using YOLOv5 in few-shot settings. Moreover, the work⁴¹ proposed BC-YOLO for real-time FSOD. The proposed BC-YOLO algorithm incorporates bi-path detection and an Attentive Drop Block. Their approach demonstrated superior performance and inference trade-offs on benchmark datasets such as PASCAL VOC 2007 and MS COCO 2014. Together, these studies showcase diverse approaches to enhancing object detection accuracy, speed, and adaptability within few-shot contexts across various domains. The work⁴² proposed a diffusion-based framework to maintain fidelity and face control in zero shot learning. The proposed model was capable of maintaining facial attributes by preserving the identities.

Focusing on the problem of facial identity recognition and challenges of facial attribute manipulation and detection, this work aims to propose a one-step solution that can address the bottleneck of the scarcity of

synthetic face datasets consisting of high-quality manipulated facial images with varying facial attributes and detection of synthetic faces for determining the facial identity in few-shot scenarios such as one-shot, three-shot, five-shot and traditional 70% – 30% split scenario. To propose this solution, we first created the Facial Attribute Manipulation Detection (FAM) Dataset by performing inversion using the StyleGAN3⁴³. The created dataset consists of 11,560 images with twenty original identities and thirty-eight varying facial manipulations such as gender swapping, age progression, makeup, etc. Moreover, the created dataset is annotated in YOLO format for all the identities which makes it suitable for usage with different Bounding-Box regression-based object detectors. The created FAM Detection Dataset has image for one-shot, three-shot, five-shot, and traditional 70%-30% split scenarios. Further, we developed an improved variant of the Tiny YOLOv7 model⁴⁴ by incorporating Swin Transformer Block (STB)⁴⁵ and Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) into its feature extraction network and proposed as the FIR-Tiny YOLOv7 (Facial Identity-Recognition-Tiny YOLOv7) model. The Swin Transformer Block (STB) in the proposed model enhanced its capabilities by accurately learning the position of the Bounding-Boxes to be used for prediction of the identities and Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) allowed the model to learn on the large feature maps and perform better detection. We have chosen the Tiny YOLOv7 model because it has a lesser number of parameters and high detection speed. To perform facial identity recognition in few-shot scenarios, we trained and tested the Tiny YOLOv7 and the proposed FIR-Tiny YOLOv7 and obtained the detection results. The quantitative and qualitative results obtained with the proposed model were fascinating as compared to the Tiny YOLOv7 and Tiny YOLOv8 models promising it to be capable of accurate recognition of facial identities in synthetic facial images having varying facial characteristics. For better understanding of the readers, we state that face recognition presents a broader view that encompasses face classification and detection as its subsets.

The major contributions of this work are:

1. Proposal of Facial Attribute Manipulation (FAM) Detection Dataset consisting of human face variations using attribute manipulation and StyleGAN3 inversion. The dataset consists of twenty original identities with 11,560 images richly annotated in YOLO format for thirty-eight varying facial attribute manipulations.
2. Proposal of Swin Transformer Block (STB) and Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) inspired improved Tiny YOLOv7 (FIR-Tiny YOLOv7) model for facial identity recognition in few-shot and traditional scenarios. The integration of Swin Transformer Block (STB) in the Tiny YOLOv7 model advances its bounding-box localization abilities which is one of the significant contributions of this work. Furthermore, the Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) is a novel contribution of this work which aided in overall increment in the detection accuracy of the proposed model.
3. Exploration of Tiny YOLOv7 and proposed FIR-Tiny YOLOv7 model for facial identity recognition in one-shot, three-shot, five-shot, and traditional scenarios. In comparison to the Tiny YOLOv7 model, the proposed model achieved a 10.0% higher mAP in the one-shot scenario, a 30.4% higher mAP in the three-shot scenario, a 15.3% higher mAP in the five-shot scenario, and a 0.1% higher mAP in traditional 70% – 30% split scenario.
4. Exhaustive experiments with state-of-the-art Tiny YOLO variants for facial identity recognition in one-shot, three-shot, five-shot, and traditional scenarios have been conducted. In comparison to the state-of-the-art Tiny YOLOv8 models, the proposed model achieved a 0.2–3.2% higher mAP in different few-shot and traditional 70% – 30% split scenarios. Specifically for five-shot and traditional 70% – 30% split scenarios, the proposed FIR-Tiny YOLOv7 model utilized 5.1–12.9 M (Million) and 4.9–12.7 M (Million) lesser training parameters as compared to the YOLOv8 Small and YOLOv8 NAS Small models. Further in terms of detection speed, the proposed model achieved 1.1–11.4(ms) and 1.2–11.4(ms) lesser inference time as compared to the YOLOv8 Small and YOLOv8 NAS Small models.

The rest of this work is organized into following sections: Sect. 2 presents the materials and method highlighting the proposed dataset and FIR-Tiny YOLOv7 model; Sect. 3 presents the experiments and results along with the comparison with the related work, and; Sect. 4 presents the conclusions along with the future scope of this work.

Materials and methods

This section describes about the dataset and the proposed model utilized for the recognition of facial identities. To carry out this work, we collected a selected number of facial images from the publicly available FFHQ (Flickr Faces High Quality) dataset and performed facial attribute manipulation using the StyleGAN3. The StyleGAN3 provided us with facial images with different attributes. Further, to perform facial identity detection with high accuracy we improvised the Tiny YOLOv7 model by applying Swin Transformer Block (STB) and Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) into its feature extraction network. In order to make the facial images generated by the StyleGAN3 suitable to the YOLO-based detectors, we annotated all the images in YOLO format and performed training and testing to obtain the detection results for facial identity recognition using the proposed model. The detailed process flow for performing facial identity recognition using the methodology adopted in this work is presented in Fig. 1. Further details about the created dataset and the proposed model are discussed in subsequent subsections.

Dataset

FFHQ (Flickr Faces High Quality) dataset by NVIDIA⁴⁶ contains 70,000 RGB images of resolution 1024 × 1024 (PNG images). In order to generate synthetic images with varying facial attributes, we employed StyleGAN3 inversion on selected images of the FFHQ dataset. The StyleGAN3 is an autoencoder-based deep neural network that generates latent space of a given input image and further generates the image back again using a series of operations such as convolutions, non-linearities, upsampling, and per-pixel noise. Though,

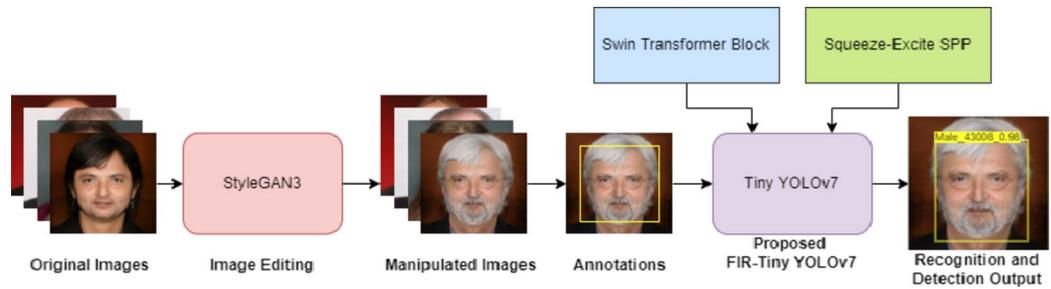


Fig. 1. Facial identity recognition methodology.

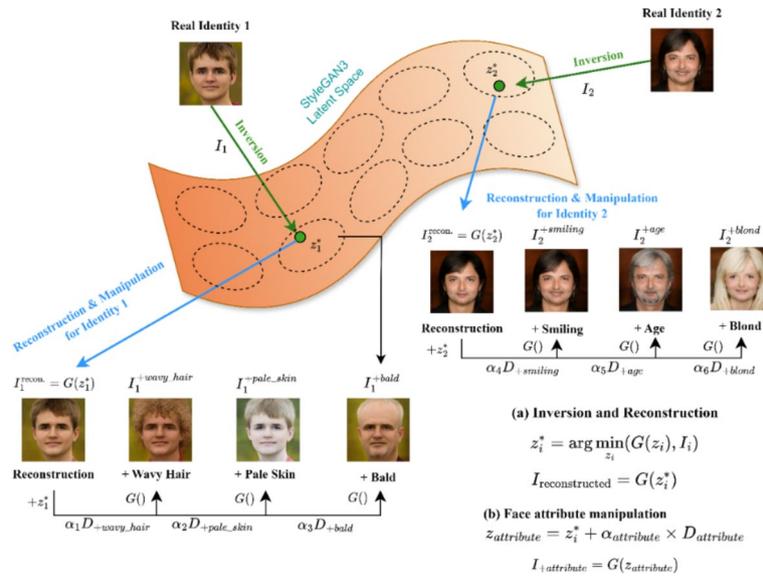


Fig. 2. StyleGAN3 inversion and manipulation process.

StyleGAN3 is composed of a generator and discriminator, in this work we utilized its generator network only to generate the synthetic facial images (reconstructed images). The process of obtaining latent space using the StyleGAN3 inversion is presented in Fig. 2. For inversion using the StyleGAN3, we evaluated the MS-SISM (Multi-Scale Structural Similarity Index) and PSNR (Peak-Signal-to-Noise) metrics to compare the similarity between the original and generated images (reconstructed images). While utilizing StyleGAN3, we utilized the below-specified losses to achieve the benchmark results for MS-SISM and PSNR metrics. Moreover, we set the thresholds of MS-SSIM to 80% and PSNR to 23dB for image quality assessment and selected 20 identities equally divided into male and female as demonstrated in Table 1; Fig. 3.

1) ID loss The ID Loss function, denoted as \mathcal{L}_{ID} , serves as a pivotal component for aligning and comparing feature representations extracted from a batch of images. Let \hat{y} , y , and x symbolize the predicted, reference, and input images, respectively, each associated with their respective feature representations $\widehat{y_{feats}}$, y_{feats} , and x_{feats} . The formulation of \mathcal{L}_{ID} for a single sample within the batch is represented as (1).

$$\mathcal{L}_{ID}(\hat{y}, y, x) = \frac{1}{N} \sum_{i=1}^N (1 - \widehat{y_{feats, i}} \cdot y_{feats, i}) \tag{1}$$

Where, N signifies the number of samples in the batch, and $\widehat{y_{feats, i}}$ and $y_{feats, i}$ represent the respective feature vectors extracted from \hat{y} and y for the i^{th} sample. The dot product operation (\cdot) quantifies the similarity between these feature representations, and the subtraction from one delineates the dissimilarity or loss metric. The primary objective of \mathcal{L}_{ID} is to minimize the disparity between the feature representations of the predicted image \hat{y} and the reference image y , contextualized by the information encapsulated in the input image x . This strategic alignment of representations aims to enrich the learning process pertaining to identity-based features.

2) W-norm loss The W-norm loss serves as a measure of the magnitude of deviation of latent representations from a reference point. By penalizing deviations, it encourages the model to produce latent vectors that align more closely with a desired distribution or target representation. The W-norm loss is expressed as (2).

Image IDs	MS-SSIM (%)	PSNR (dB)
04580	79.74	23.37
04108	75.97	23.23
66,195	74.95	23.81
42,708	79.62	23.26
03118	80.93	23.51
04516	76.02	22.90
42,001	80.52	23.67
02069	79.15	23.13
00218	84.16	23.28
05155	89.94	25.47
04865	81.98	23.16
66,311	85.27	24.82
25,527	73.51	23.21
69,207	80.72	23.30
09544	76.15	22.79
25,033	84.25	25.15
04751	79.63	23.25
69,069	78.52	24.02
42,496	83.09	23.63
03094	73.69	23.92

Table 1. Reconstructed image quality metrics.



Fig. 3. Original and reconstructed images selected IDs.

$$\mathcal{L}_{w_norm} = \frac{1}{N} \sum_{i=1}^N |(\text{latent}_i - \text{latent}_{\text{avg}})|_2 \quad (2)$$

Where, N denotes the batch size, latent_i signifies each individual latent representation within the batch, and $\text{latent}_{\text{avg}}$ represents the average latent space vector. This loss function's utility extends to tasks involving style manipulation, generative model training, and latent space optimization. It aids in controlling the structure and properties of generated outputs, ensuring they align with specific characteristics defined by the latent space distribution or average.

3) MoCo loss: The MoCo (Momentum Contrast) loss mechanism is instrumental in self-supervised learning paradigms which is specifically designed for unsupervised visual representation learning in computer vision tasks. The core principle of MoCo revolves around leveraging contrastive learning techniques to enhance the quality of learned representations from unlabelled data. The MoCo loss function comprises two primary components: the loss term and the similarity improvement term expressed as (3–5).

$$\text{Loss Term} = \frac{1}{N} \sum_{i=1}^N (1 - \text{diff}_{\text{target}_i}) \quad (3)$$

Where, N signifies the number of samples, and $\text{diff}_{\text{target}_i}$ represents the dot product similarity between the predicted features and the features from the target view of the same image for the (i) th sample.

$$\text{Similarity Improvement Term} = \frac{1}{N} \sum_{i=1}^N (\text{diff}_{\text{target}_i} - \text{diff}_{\text{views}_i}) \quad (4)$$

Here, $\text{diff}_{\text{views}_i}$ denotes the dot product similarity between the features from the target view and the features from a different view of the same image for the (i) th sample.

The overall MoCo loss is a combination of these terms:

$$\mathcal{L}_{\text{MoCo}} = \frac{\text{Loss Term}}{N}, \quad \text{Similarity Improvement Term} \quad (5)$$

Along with the above specified losses, in this work we utilized the Learned Perceptual Image Patch Similarity (LPIPS) to evaluate the similarity between the two images. The LPIPS is a deep learning-based network trained to determine human perceptual judgements of similarity between images by measuring the distance between the feature representations of images under consideration. As proposed in⁴⁷, the lower LPIPS value indicates higher perceptual similarity between the two images. The LPIPS is expressed as (6).

$$\text{LPIPS}(x, y) = \sum_{i=1}^L \frac{1}{H_i W_i} \sum_{h=1}^{H_i} \sum_{w=1}^{W_i} \|\varphi_i(x)_{h,w} - \varphi_i(y)_{h,w}\|_2^2 \quad (6)$$

Here, x and y are the images under consideration for comparison, L signifies the number of layers in the deep learning-based network, H_i and W_i signifies the height and width of the feature representation at i th layer of the network, and feature representation of image x at layer i is specified using $\varphi_i(x)$.

In this work, we randomly selected over three hundred images consisting of male and female faces. Through StyleGAN3 inversion we reconstructed the latent representations of those original faces. Further, to filter high quality and identity preserved reconstructions, we have used MS-SSIM (Multi Scale Structural Similarity Index Measure) and PSNR (Peak Signal-to-Noise Ratio) with above specified thresholds. Image structures at different scales poses challenge to the human visual systems. Therefore, we leveraged the Multi-Scale Structural Similarity Index (MS-SSIM)⁴⁸ that checks for structural similarity at multiple scales by using multiple down-sampling and up-sampling operations. The MS-SSIM is expressed as (7).

$$\text{MS-SSIM}(x, y) = \frac{1}{L} \sum_{i=1}^L w_i \prod_{j=1}^i \text{SSIM}(x_j, y_j)^\alpha \quad (7)$$

Where, x and y are the two images under consideration for comparison, L represents the number of scales, w_i signifies a set of weights that reduces with increasing scale, and α is a variable that controls the corresponding value of each scale.

In order to measure the quality of reconstructed and compressed images, we have used Peak-Signal-to-Noise Ratio (PSNR) metric. It evaluates the ratio between the signal over noise in an image that signifies the quality of the image and its representation. The PSNR value with a high score represents a higher quality image. The PSNR is expressed as (8).

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right) \quad (8)$$

Where, MAX_I represents the maximum pixel value of the image and MSE represents the mean squared error between the original and reconstructed images.

Using the latent space manipulation of the StyleGAN3, we created 38 facial attributes ('receding hairline', 'age', 'smiling', 'eyeglasses', 'straight hair', 'no beard', 'pale skin', 'pose', 'pointy nose', 'narrow eyes', 'oval face', 'male', 'mouth slightly open', 'moustache', 'heavy makeup', 'high cheekbones', 'bushy eyebrows', 'chubby', 'double chin', 'blurry', 'brown hair', 'blond hair', 'big lips', 'black hair', 'bags under eyes', 'bald', 'arched eyebrows', 'young', 'bangs', 'wavy hair', 'rosy cheeks', 'sideburns', 'gray hair', 'goatee', 'attractive', '5 o'clock shadow', 'big nose', 'wearing lipstick') and further categorized into three broad facial representations: 'aging', 'lifestyle changes', and 'medical procedures' (can be due to surgical/non-surgical/cosmetic) with various intensities ($-\alpha_{\text{attribute}}$ to $+\alpha_{\text{attribute}}$) (as demonstrated in Fig. 2) for predefined vector directions ($D_{\text{attribute}}$) for each attributes, and latent vector (z_i^*) in the StyleGAN3 latent space, we manipulated each faces (I_i) to mimic real world face diversity of a person ($I_i^{\pm \text{attribute}}$) and created "FAM Dataset". The image samples from the created "FAM Dataset" are presented in Fig. 4. The StyleGAN3 inversion and manipulation are expressed as (9) and face attribute manipulation is expressed as (10–11).

$$z_i^* = \operatorname{argmin}_{z_i} (G(z_i), I_i) \quad (9)$$

$$z_{\text{attribute}} = z_i^* \pm \alpha_{\text{attribute}} \times D_{\text{attribute}} \quad (10)$$

$$I_i^{\pm \text{attribute}} = G(z_{\text{attribute}}) \quad (11)$$

Where, G is StyleGAN3 generator function.

The created FAM Dataset contains twenty original identities and 11,560 facial attributes manipulated PNG (RGB) images of twenty identities with resolution 416×416 . This is a synthetic dataset we proposed and derived from original FFHQ by NVIDIA (for samples of FAM Dataset refer Fig. 3). Further, we propose few-shot approaches with three distinct criteria for face identification (detection as well as recognition) to represent real-world data paucity.

1) One-shot approach: only 'original' images present per class in training split (total one image).

2) Three-shot approach: with 'original' extremes of 'age' manipulation are present in training set (total three images per class).

3) Five-shot approach: including criteria for three-shot, extremes of 'gender' manipulation are present in training set (total five images per class).

As synthetic dataset for face recognition benchmarking proven effective in³³, we considered traditional 70% – 30% splits as benchmark to assess the performance of the proposed FIR-Tiny YOLOv7 and other trained and tested Tiny YOLO models in few-shot approaches. Out of thirty-eight attributes, we have chosen twenty-eight attributes for training split and ten attributes for test split. We manually annotated FAM Dataset in YOLO annotation format to create "FAM Detection Dataset." The FAM Detection Dataset contains twenty classes in nomenclature of 'gender {FFHQ_{image id}}'. For example, if the image id in original FFHQ is '42001' and the person is 'male' then in our dataset class it is named as 'male_42001'. FAM Detection dataset has twenty original identities and 11,560 manipulated face images in JPEG format (RGB) with resolution 416×416 and their corresponding annotations in TEXT file.

Proposed improved tinyYOLOv7 model

You Only Look Once (YOLO)⁸ is a state-of-the-art object detection model that has shown its competencies in various downstream tasks such as general object detection⁴⁹, face detection⁵⁰, faces with mask⁵¹, etc. The YOLO

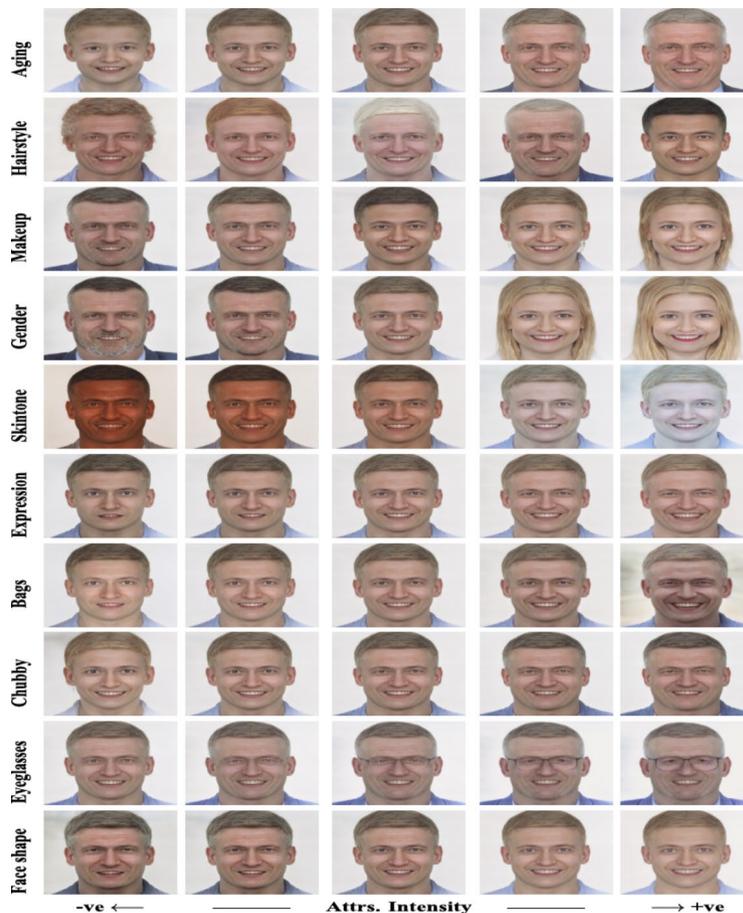


Fig. 4. Manipulated sample images from FAM Dataset.

detection model is popular because of its high detection accuracy and speed. To improve the detection speed and to train under a limited computation resources environment, the developers of YOLO have proposed small variants of the YOLO object detection model as the Tiny YOLO model. Amongst, variously proposed Tiny YOLO models, the Tiny YOLOv7 model⁴⁴ recently proposed has only 6.2 M parameters. However, the bottleneck with the Tiny YOLOv7 model is its low detection accuracy. Considering the advantage of Tiny YOLOv7 having fewer parameters and addressing the challenge of improving its detection accuracy, in this work we have chosen it for the task of facial attribute manipulation detection leading to facial identity recognition.

The Tiny YOLOv7 model is a single-stage object detection model comprising of feature extraction and detection network. Its feature extraction network is composed of fifty-three convolutional layers, maxpooling, and upsampling layers. Whereas, its detection network is composed of five convolutional layers and the C-IoU loss function. The C-IoU loss function in the Tiny YOLOv7 model penalize the network for any incorrect prediction and helps in faster regression. In this work, to improve the detection accuracy of the Tiny YOLOv7 model for few-shot and traditional scenarios, we have improvised its feature extraction network while utilizing the detection network without any changes. In order to improve its feature extraction network, we added a Swin Transformer Block (STB) and Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) which allowed the network to learn residual features and perform feature aggregation. We added the Swin Transformer Block (STB) after the twenty-sixth convolutional layer and the Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) on top of the last convolutional layer of the Tiny YOLOv7 feature extraction network. The Swin Transformer Block (STB) has been added after the twenty-sixth layer of the feature extraction network because of the reason that after this layer the network splits into three diversified branches of convolutional layers used by the YOLO detection heads to make predictions. The Swin Transformer Block (STB) at this location adds attention mechanism to the feature extraction network thereby, providing it with rich features for enhanced interpretability. Further, the Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) has been added on top of the last convolutional layer of the feature extraction network because the last convolutional layer holds the most crucial and contextual representations of the feature maps used by the YOLO detection heads to make accurate predictions by drawing precise bounding boxes. The Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) is a novel contribution of this work which enhanced the detection accuracy of the Tiny YOLOv7 model by multi-folds. With the proposed enhancements in the Tiny YOLOv7 model, we named the developed architecture as FIR-Tiny YOLOv7 model. Since, the Tiny YOLOv7 is a bounding-box regression-based object detection model therefore, for its three detection layers we calculated the anchor boxes for the created “FAM Detection Dataset” and passed the values to the detection network for accurate localization of the bounding-boxes in the detected images. We calculated the anchor boxes using the k-means++ clustering and obtained the following values for the anchor boxes: [257 × 403, 263 × 407, 271 × 401, 269 × 407, 271 × 409, 273 × 407, 277 × 409, 277 × 413, 279 × 413]. Moreover, the filter size for the convolutional layer of the YOLO detection network has been calculated using the formula: $((Classes + 5) * 3)$. As the number of classes in the created “FAM Detection Dataset” is twenty, therefore, the filter size has been set to 75. In the proposed model, the output feature maps of the three YOLO detection heads correspond to three scales: $13 \times 13 \times 75$, $26 \times 26 \times 75$, and $52 \times 52 \times 75$. The entire model is trained with the ReLU (Linear Rectified Unit) activation function while the activation function for the filter layer is set to Logistic. The detailed outline of the proposed model is presented in Fig. 5. The details about the added Swin Transformer Block (STB) and Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) are discussed below.

a) Swin Transformer Block.

This work aims to recognize identities in image samples having similar faces with manipulated facial attributes. The Tiny YOLOv7 model can perform the detection of faces however, it struggles to perform correct predictions when the face image samples under consideration are similar thereby, leading to missed predictions for a few image samples. One approach to address this challenge is to add an attention mechanism in the feature extraction network of the Tiny YOLOv7 model. The attention mechanism enables the model to focus on certain parts of the images used for training and pay more weight to those parts while making the prediction. Inspired by the recent advancements in visual recognition systems, in this work, we have utilized Swin Transformer Block (STB)⁴⁵ with the feature extraction network of the Tiny YOLOv7 model. We specifically used the Swin Transformer Block (STB) due to its ability to maintain speed-accuracy trade-off over other attention mechanisms. The Swin Transformer Block (STB) is composed of Window-based Multi-head Self-Attention (W-MSA) and Sliding Window-based Multi-head Self-Attention (SW-MSA) which acts as a central module for computation of attention. It functions by replacing the large feature maps with windows and layers i.e. small patches of feature maps which further improves the speed-accuracy trade-off. In the Swin Transformer Block (STB), the Window-based Multi-head Self-Attention (W-MSA) decreases the complexity of self-attention by splitting the image into windows i.e. small patches consisting of feature maps for pixels and further computing attention for each window. Whereas, the Sliding Window-based Multi-head Self-Attention (SW-MSA) enables the connections between the windows to perform the calculation of attention among different windows thereby, resulting in higher efficiency. The detailed illustration of the Swin Transformer Block (STB) is presented in Fig. 6.

To improve the detection accuracy of the Tiny YOLOv7 for recognition of facial images having manipulated attributes, we applied the Swin Transformer Block (STB) after the twenty-sixth convolutional layer of its feature extraction network. We specifically applied the Swin Transformer Block (STB) at this position because, till this layer, the Tiny YOLOv7 feature extraction network utilizes the local feature transformations by retaining the contextual pixel representations of the input images. The Swin Transformer Block (STB) splits the feature maps (pixels) for the facial images having manipulated attributes into smaller windows and provides rich feature maps by applying the attention mechanism which can be used by further layers of the network to learn upon the features for same faces with varying facial attributes and leading to accurate classification, detection and recognition.

b) Squeeze-Excite Spatial Pyramid Pooling.

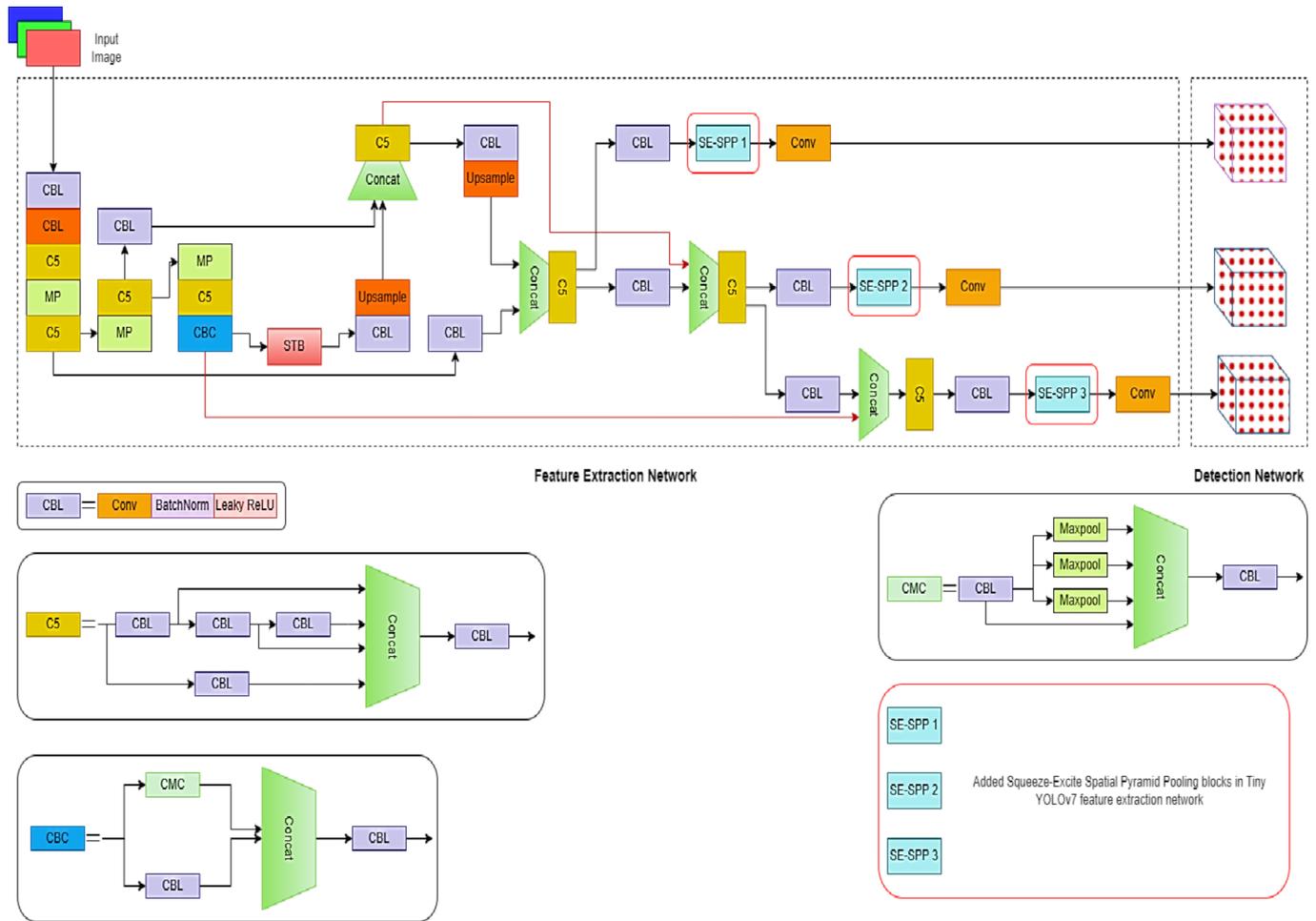


Fig. 5. Proposed FIR-Tiny YOLOv7 model.

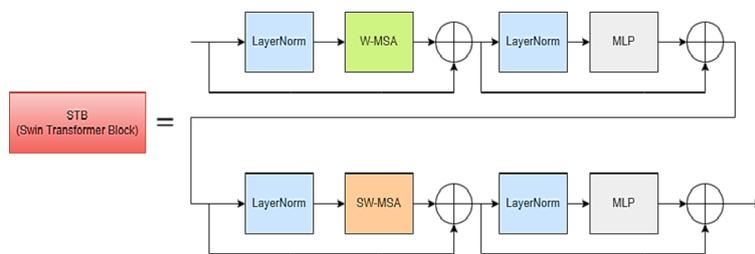


Fig. 6. Swin Transformer Block.

To increase the detection accuracy of the proposed FIR-Tiny YOLO v7 model we developed a novel spatial pyramid pooling⁵² mechanism. Since the feature extraction network of the proposed model is based on CNNs (Convolutional Neural Networks), the spatial pyramid pooling helps in removing the fixed-size constraints of the CNN network. It aids the network by providing variable-size input feature maps. In this work, we further improved the spatial pyramid pooling by utilizing it in a squeeze-excite manner i.e. a few of the input feature maps generated by the spatial pyramid pooling are squeezed and others are excited. The detailed outline of the developed Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) is presented in Fig. 7. The developed Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) is added on top of the last convolutional layer of each detection head of the YOLO detection network. The developed Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) is constituted of three maxpool layers each of filter size 5×5 , 7×7 , and 9×9 with a stride of 1. The output of the first maxpool layer is passed to a convolutional layer of size 128 that squeezes the feature map and the output of the second and third maxpool layers are passed to two convolutional layers of size 512 that excites the feature maps. Further, the feature map of the CNN layer after which the proposed Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) is applied is concatenated with the feature maps of the CNN layers applied after the three maxpool

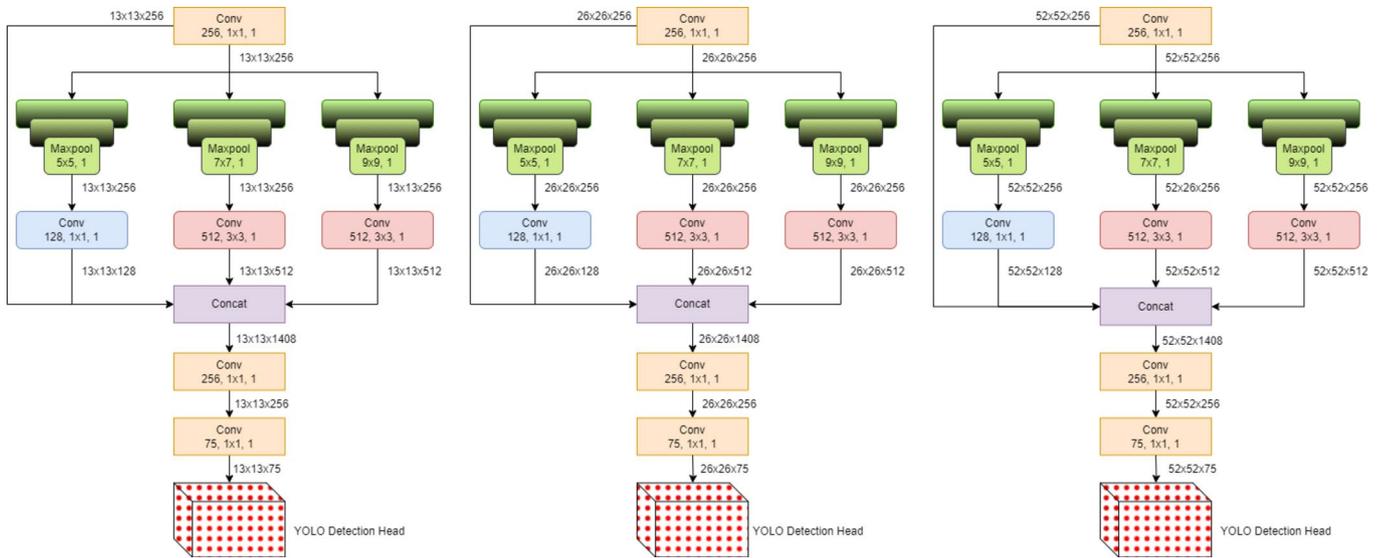


Fig. 7. Squeeze-Excite Spatial Pyramid Pooling.

layers. The proposed model has three detection layers therefore, this entire operation enhanced the feature map from $13 \times 13 \times 256$, $26 \times 26 \times 256$, and $52 \times 52 \times 256$ to $13 \times 13 \times 1408$, $26 \times 26 \times 1408$, and $52 \times 52 \times 1408$ thereby, providing the YOLO detection network of the proposed model to perform detection on larger feature maps and perform feature aggregation, and produce accurate detection results. Mathematically, the proposed Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) can be expressed as (12–16)

$$z_i = (f * x_i) + b_i \tag{12}$$

Where, z_i represents the output of the convolutional layer after which the proposed Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) is applied, f represents its filter size, x_i represents its feature map, and b_i represents the bias term

$$y_{ij} = \max_{m, n} (x_{(i.s+m)(j.s+n)}) \tag{13}$$

Where, y_{ij} signifies the output of the maxpooling operation at position (i, j) in the output feature map, x signifies the input feature map, s signifies the stride of the maxpooling operation, and m and n are the two index values used to traverse over the local region within the feature map.

$$z_{squeezed(i)} = (f * x_i) + b_i \tag{14}$$

Where, $z_{squeezed(i)}$ represents the output of the convolutional layer which squeezed the feature map of the preceding maxpooling layer.

$$z_{excited(i)} = (f * x_i) + b_i \tag{15}$$

Where, $z_{excited(i)}$ represents the output of the convolutional layer which excited the feature map of the preceding maxpooling layers.

$$Y = Concat(z_i, z_{squeezed(i)}, z_{excited(i)}) \tag{16}$$

Where, Y represents the concatenated feature map of the convolutional layer after which the proposed Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) have been applied and the feature maps of the squeezed and excited maxpooling layers generated by the subsequent convolutional layers.

Experiments and evaluations

To implement the proposed FIR-Tiny YOLOv7 and other tested detection models, we have used the DarkNet repository available at³³. Further, the training and testing have been done on the Google Colaboratory which provided access to 12 GB RAM and 16 GB Tesla K-80 GPU. For all the models, the batch size was set to 32 with a sub-division of 8; the learning rate was set to 0.0001; decay of 0.005, and; momentum of 0.9. Moreover, to obtain realistic results we applied the “random function” with all the trained and tested models that varied the training images to the scale of 512×512 and 608×608 from the original size of 416×416 . Further, all the models have been evaluated using the standard object detection performance metrics to gauge their validity and effectiveness. The details about the evaluation metrics and the obtained results are presented in the subsequent subsections.

Evaluation metrics

Performance evaluation metrics in object detection play a pivotal role in assessing the accuracy and efficiency of models. Precision signifies the total number of true positive predictions among all positive predictions, while recall signifies the proportion of true positives among all ground truths, both essential for achieving a balance between accuracy and comprehensiveness. Mean Average Precision (mAP) assesses object classification and localization across different thresholds, placing emphasis on both accurate detection and precise localization. Inference time is crucial for real-time applications such as video surveillance or autonomous driving. Each metric—precision, recall, mAP, and inference time—provides distinct insights into a model's performance, ensuring accurate object detection while addressing practical usability concerns. A high mAP indicates precise identification and accurate localization of objects, which are essential for ensuring robust model performance, particularly in scenarios like face detection in image spaces. The mathematical expressions for the stated performance evaluation metrics are expressed as (17–19).

$$\text{Precision } (P) = \frac{\text{True Positives } (TP)}{\text{True Positives } (TP) + \text{False Positives } (FP)} \quad (17)$$

$$\text{Recall } (R) = \frac{\text{True Positives } (TP)}{\text{True Positives } (TP) + \text{False Negatives } (FN)} \quad (18)$$

$$\text{Mean Average Precision } (mAP) = \frac{1}{n} \sum_{k=1}^{k=n} AP_k \quad (19)$$

Evaluation results

In order to assess the improvements with the proposed FIR-Tiny YOLOv7 model in comparison to the Tiny YOLOv7 model, we evaluated the precision, recall, and mAP on the test set of the created FAM Detection Dataset for few-shot and traditional 70% – 30% split scenarios. Across all the comparisons for different performance metrics, the proposed model achieved better results as compared to the Tiny YOLOv7 model. For the one-shot scenario, the proposed model achieved a 21.5% higher value for precision, a 21.9% higher value for recall, and a 10% higher value for mAP. For the three-shot scenario, the proposed model achieved a 30.4% higher value for precision, a 33.1% higher value for recall, and a 30.4% higher value for mAP. For the five-shot scenario, the proposed model achieved a 9.5% higher value for precision, a 14.8% higher value for recall, and a 15.3% higher value for mAP. Moreover, for the traditional 70% – 30% split scenario, the proposed model achieved a 0.3% higher value for precision, a 0.8% higher value for recall, and a 0.1% higher value for mAP. Furthermore, the proposed model utilized only 0.1 M extra parameters with the added enhancements while maintaining the inference time and achieving better values for detection accuracy (mAP). The results of the proposed model are fascinating for one-shot, three-shot, and five-shot scenarios thus, it can be stated that the proposed model can achieve better results as compared to the Tiny YOLOv7 model for facial identity recognition under the limited training data availability. The detailed comparative results of the proposed model with the Tiny YOLOv7 model are presented in Table 2.

In an effort to benchmark the performance of the proposed FIR-Tiny YOLOv7 model for facial identity recognition, we trained and tested different Tiny YOLO variants on the created FAM Detection Dataset and evaluated the performance metrics on its test set. Further, we carried out a comparison between the proposed model and other trained and tested Tiny YOLO models to gauge the efficacy of the proposed model. We have specifically discussed for performance metric mAP as it is the only metric that gauges the validity of the overall model for recognition and detection accuracy in object detection tasks. For the one-shot approach, the proposed model performed better as compared to Tiny YOLOv7, YOLOv7-P5, and YOLOv8 NAS Small only by achieving 10–23.13% higher mAP. However, in comparison to other Tiny YOLO variants, the proposed model performance was comparatively poor. For the three-shot approach, the proposed model performed better as compared to Tiny YOLOv7 and YOLOv7-P5 models only by achieving 30.4–32.8% higher mAP. However, in comparison to other Tiny YOLO variants, the proposed model performance was unsatisfactory. However, for the five-shot and traditional 70% – 30% split approach, the proposed model surpasses the performance of all the tested Tiny YOLO variants by achieving a 0.2–81.9% higher mAP in the five-shot scenario and a 0.2–25.5% higher mAP in traditional 70% – 30% split scenario by beating the state-of-the-art YOLOv8 Small model. Interestingly, the

Model	Approach	Precision (%)	Recall (%)	mAP @ 0.50 (%)	Parameters (M)	Inference (ms)
Tiny YOLOv7	One-shot	21.3	19.6	13.5	6.2	5.2
Proposed		42.8	41.5	23.5	6.3	5.2
Tiny YOLOv7	Three-shot	25.4	17.1	16.4	6.2	5.6
Proposed		55.8	50.2	46.8	6.3	5.6
Tiny YOLOv7	Five-shot	89.1	83.3	83.1	6.2	5.2
Proposed		98.6	98.1	98.4	6.3	5.3
Tiny YOLOv7	Traditional 70%-30%	98.6	97.3	98.6	6.2	5.5
Proposed		98.9	98.1	98.7	6.3	5.5

Table 2. Proposed FIR-Tiny YOLOv7 comparison with tiny YOLOv7 model.

Model	Approach	Precision (%)	Recall (%)	mAP @ 0.50 (%)	Parameters (M)	Inference (ms)
Tiny YOLOv4	One-shot	88.7	80.2	86.2	6.0	3.0
Tiny YOLOv5		82.3	75.6	79.8	7.5	2.8
Tiny YOLOv7		21.3	19.6	13.5	6.2	5.2
YOLOv7-P5		17.2	18.5	12.8	36.9	18.5
YOLOv8 Small		90.9	82.8	90.7	11.2	6.4
YOLOv8 Nano		89.7	73.7	85.2	3.2	5.2
YOLOv8 NAS Small		0.20	0.42	0.37	19	16.7
Proposed		42.8	41.5	23.5	6.3	5.2

Table 3. Proposed FIR-Tiny YOLOv7 comparison with YOLO models for one-shot approach.

Model	Approach	Precision (%)	Recall (%)	mAP @ 0.50 (%)	Parameters (M)	Inference (ms)
Tiny YOLOv4	Three-shot	98.2	96.4	92.7	6.0	3.1
Tiny YOLOv5		84.7	78.1	80.4	7.5	2.8
Tiny YOLOv7		25.4	17.1	16.4	6.2	5.6
YOLOv7-P5		20.2	22.7	14.8	36.9	18.6
YOLOv8 Small		97.5	95.5	97.4	11.2	6.4
YOLOv8 Nano		98.6	92.1	95.9	3.2	5.2
YOLOv8 NAS Small		87.9	87.5	85.8	19.0	17.0
Proposed		55.8	50.2	46.8	6.3	5.6

Table 4. Proposed FIR-Tiny YOLOv7 comparison with YOLO models for three-shot approach.

Model	Approach	Precision (%)	Recall (%)	mAP @ 0.50 (%)	Parameters (M)	Inference (ms)
Tiny YOLOv4	Five-shot	98.2	96.7	94.8	6.0	3.3
Tiny YOLOv5		87.3	81.4	82.3	7.6	3.1
Tiny YOLOv7		89.1	83.3	83.1	6.2	5.2
YOLOv7-P5		24.7	26.1	16.5	36.9	18.5
YOLOv8 Small		98.4	97.8	98.2	11.2	6.4
YOLOv8 Nano		98.2	97.8	98.2	3.2	5.2
YOLOv8 NAS Small		95.1	95.9	95.2	19.0	16.7
Proposed		98.6	98.1	98.4	6.3	5.3

Table 5. Proposed FIR-Tiny YOLOv7 comparison with YOLO models for five-shot approach.

proposed model achieved a better speed-accuracy tradeoff in comparison to the state-of-the-art YOLOv8 and YOLOv8 NAS Small model by achieving 1.1–11.4(ms) and 1.2–11.4(ms) lesser inference time. The performance results for the different comparisons are presented in Tables 3, 4 and 5, and 6.

For a more justifiable comparison of the performance of the proposed model in few-shot and traditional 70% – 30% split scenarios, we performed recognition and detection on a few image samples of the test set of the created FAM Detection Dataset and obtained realistic qualitative results. The qualitative results with the proposed model are illustrated in Fig. 8. For the varying facial attributes of the same face in few-shot and traditional 70% – 30% split scenarios, the proposed model accurately recognized the facial identity of the same face with manipulated facial attributes with high accuracy. However, there were a few incorrect and missed predictions but those were only in one-shot and three-shot scenarios where the proposed model achieved a relatively low detection accuracy (mAP). However, for the five-shot and traditional 70% – 30% split scenarios, the proposed model recognized all the identities correctly despite having extensive facial attribute manipulation leading to a complete change of the identity of the person's face.

Ablation tests

In order to test the individual contribution of the added enhancements to the Tiny YOLOv7 model, we performed two ablation tests: (1) Tiny YOLOv7 with Swin Transformer Block (STB). (2) Tiny YOLOv7 with Squeeze-Excite Spatial Pyramid Pooling (SE-SPP). Further, we compared the results of the two ablation tests with the performance of the proposed FIR-Tiny YOLOv7 model. The detailed results of the ablation tests are presented in Table 7.

As shown in Table 7, in the first ablation test with the combination of Tiny YOLOv7 and Swin Transformer Block (STB), for the one-shot approach, the model achieved a precision of 24.8%, a recall of 22.5%, and a mAP

Model	Approach	Precision (%)	Recall (%)	mAP @ 0.50 (%)	Parameters (M)	Inference (ms)
Tiny YOLOv4	Traditional 70%-30%	98.8	97.2	98.6	6.0	3.4
Tiny YOLOv5		98.2	98.1	98.0	7.6	3.2
Tiny YOLOv7		98.6	97.3	98.6	6.2	5.5
YOLOv7-P5		95.9	70.0	73.4	36.9	18.6
YOLOv8 Small		98.8	97.9	98.5	11.2	6.7
YOLOv8 Nano		98.4	98.5	98.2	3.2	5.4
YOLOv8 NAS Small		98.6	98.4	98.4	19.0	16.9
Proposed		98.9	98.1	98.7	6.3	5.5

Table 6. Proposed FIR-Tiny YOLOv7 comparison with YOLO models for traditional approach.

value of 14.8%, for three-shot approach, the model achieved a precision of 28.8%, a recall of 20.2%, and a mAP value of 18.1%, for five-shot approach, the model achieved a precision of 93.5%, a recall of 87.2%, and a mAP value of 87.9%, and for the traditional 70%-30% split approach, it achieved a precision of 98.1%, a recall of 97.9%, and a mAP value of 98.7%. The addition of the Swin Transformer Block (STB) to the Tiny YOLOv7 model increased the baseline model performance by 0.1–4.8% in terms of detection accuracy (mAP). However, the number of parameters and inference time remained the same across all the approaches as compared to the baseline Tiny YOLOv7 model. For the second ablation test with the combination of Tiny YOLOv7 and Squeeze-Excite Spatial Pyramid Pooling (SE-SPP), for one-shot approach, the model achieved a precision of 30.8%, a recall of 29.9%, and a mAP value of 21.2%, for three-shot approach, the model achieved a precision of 37.4%, a recall of 32.6%, and a mAP value of 29.8%, for five-shot approach, the model achieved a precision of 95.2%, a recall of 90.8%, and a mAP value of 91.9%, and for the traditional 70%-30% split approach, it achieved a precision of 98.8%, a recall of 98.0%, and a mAP value of 98.7%. The addition of Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) to the Tiny YOLOv7 model increased the baseline model performance by 7.7–13.4% in terms of detection accuracy (mAP) for the one-shot, three-shot, and five-shot approach. For, the traditional 70%-30% split the mAP value remained the same. However, the number of parameters and inference time remained the same across all the approaches as compared to the baseline Tiny YOLOv7 model. Moreover, with the two ablation tests, there was a significant improvement in precision and recall across all the split approaches. The results of the ablation tests indicate the individual contribution of Swin Transformer Block (STB) and Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) in improving the performance of baseline Tiny YOLOv7 for different evaluation metrics without sacrificing on number of trainable parameters and inference time.

Comparison with related work

To verify the results of the proposed FIR-Tiny YOLOv7 model, we performed a direct comparison with the related works trained and tested on the FFHQ dataset. We have selected the FFHQ dataset for comparison because the FAM Detection Dataset as created and proposed in the subset of FFHQ dataset. In the work⁵⁴, the researchers have proposed an improved variant of the Xception model for the recognition of faces generated using the GAN-based approach. The researchers improvised the Xception model by applying three variations (1) M1 – removing four residual blocks from the Xception model; (2) M2 – replacing the common convolutional layer with the Inception module with the dilated convolution and; (3) M3 - addition of feature pyramid network to obtain multi-level features. With the proposed model they achieved an accuracy of 89.2% on the FFHQ dataset in the GAN inpainting scenario. In the work⁵⁵, the researchers utilized deep learning-based models for face forensics detection on the images of the FFHQ dataset for real v/s fake v/s edited images. They exploited Xception, Encoder-Decoder, and MesoNet and obtained detection accuracy of 99.7% with the Xception model, 98.6% with the Encoder-Decoder model, and 86.0% with the MesoNet model for real v/s fake v/s edited images extracted and generated from the FFHQ dataset. In⁵⁶, the authors explored Bayesian linear regressor to classify between natural and synthetic faces and achieved a classification accuracy of 78.1% for synthetic faces. The authors⁵⁷ proposed GAN discriminator-based model for recognizing facial identities for different face emotions. On the employed dataset, the authors achieved a combined accuracy of 89.5%. The work⁵⁸ proposed a combination of VGG and CBAM attention mechanism to classify facial attributes. On the CK+ and FER-2013, the authors achieved an accuracy of 69.0%. In⁵⁹, the authors proposed multi-channel attention mechanism based ResNet18 architecture for classification of facial identities. On the CK+ dataset, the authors achieved an accuracy of 98.8%. The work⁶⁰ proposed a local patch attention-based CNN architecture for recognition of facial identities. In real world scenarios dataset, it achieved an accuracy of 86.58%. The comparison of the proposed model with the related work is highlighted in Table 8.

As shown in Table 8, the proposed model performed better or equal to the models proposed in related work for the detection of synthetic faces generated using the GAN-based models. Therefore, this can be stated that the proposed model is a useful tool for the detection of facial images generated using the GAN-based approach.

Generalization results

The proposed FIR-Tiny YOLO v7 model was trained and tested on the FAM Detection dataset which is a subset of the FFHQ dataset. To gauge the efficacy of the proposed model, we tested the proposed model with the weights trained on the FAM Detection dataset on 10,000 facial images each of CelebA⁶¹ and LFW⁶² datasets to evaluate the proposed model's generalization ability on the other datasets. To get the generalization results,

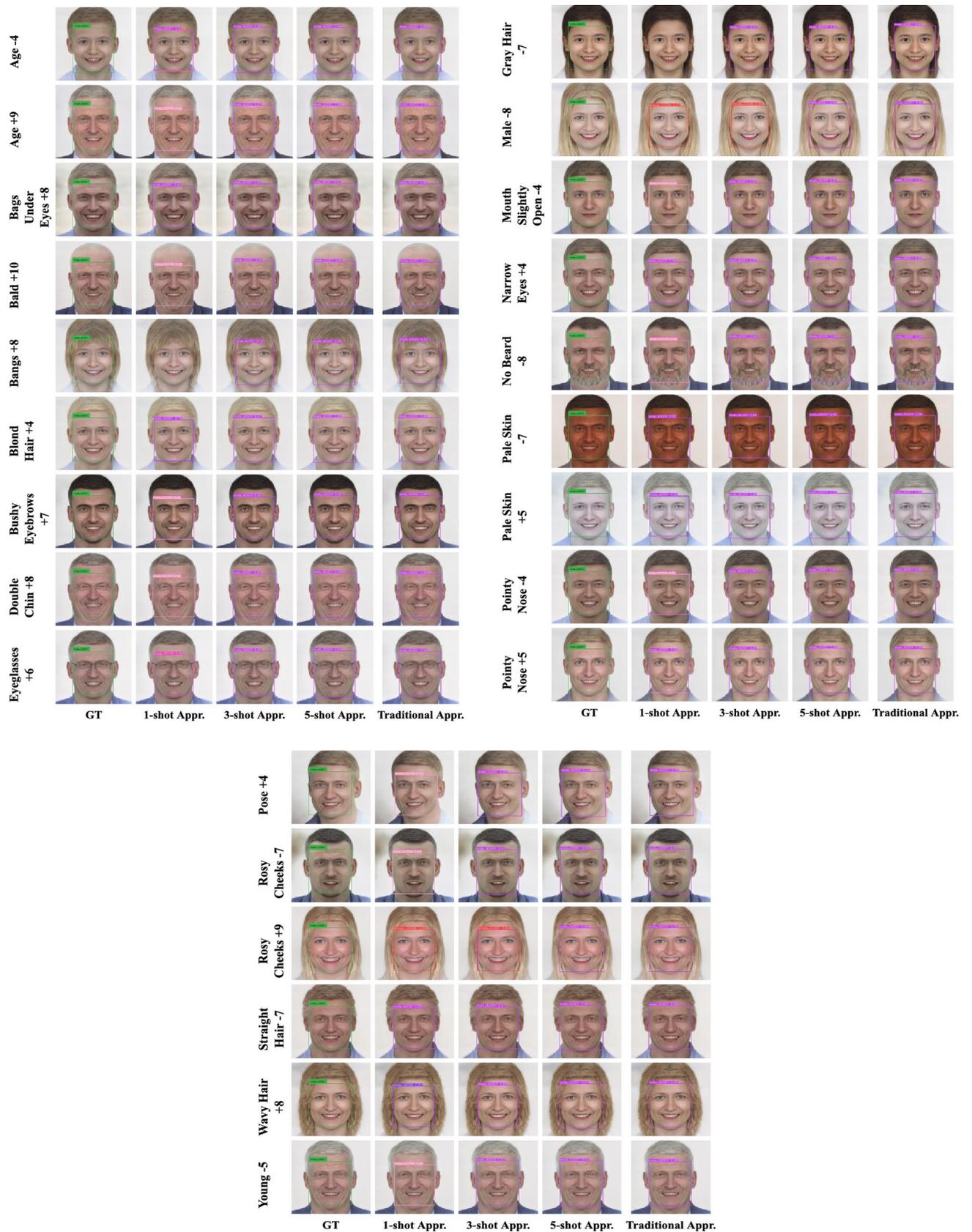


Fig. 8. Facial identity recognition with the proposed FIR-Tiny YOLOv7.

Model	Approach	Precision (%)	Recall (%)	mAP @ 0.50 (%)	Parameters (M)	Inference (ms)
Tiny YOLOv7	One-shot	21.3	19.6	13.5	6.2	5.2
Tiny YOLOv7 + STB		24.8	22.5	14.8	6.2	5.2
Tiny YOLOv7 + SE-SPP		30.8	29.9	21.2	6.3	5.2
Proposed		42.8	41.5	23.5	6.3	5.2
Tiny YOLOv7	Three-shot	25.4	17.1	16.4	6.2	5.6
Tiny YOLOv7 + STB		28.8	20.2	18.1	6.2	5.2
Tiny YOLOv7 + SE-SPP		37.4	32.6	29.8	6.3	5.4
Proposed		55.8	50.2	46.8	6.3	5.6
Tiny YOLOv7	Five-shot	89.1	83.3	83.1	6.2	5.2
Tiny YOLOv7 + STB		93.5	87.2	87.9	6.2	5.2
Tiny YOLOv7 + SE-SPP		95.2	90.8	91.9	6.3	5.3
Proposed		98.6	98.1	98.4	6.3	5.3
Tiny YOLOv7	Traditional 70%-30%	98.6	97.3	98.6	6.2	5.5
Tiny YOLOv7 + STB		98.1	97.9	98.7	6.3	5.5
Tiny YOLOv7 + SE-SPP		98.8	98.0	98.7	6.3	5.5
Proposed		98.9	98.1	98.7	6.3	5.5

Table 7. Ablation tests results.

Work	Model	Accuracy/ mAP (%)
Chen et al. ⁵⁰	Xception	71.4
	Xception + M1 + M2 + M3	89.2
Songsri-in and Zafeiriou ⁵¹	Xception	99.7
	Encoder-Decoder	98.6
	MesoNet	86.0
Uittenhove et al. ⁵⁸	Bayesian linear regressor	78.1
Villegas-Ch et al. ⁵⁹	GAN discriminator	89.5
Cao et al. ⁶⁰	VGG + CBAM	69.0
Shen and Xu ⁶¹	ResNet18 + MCAM	98.8
Liu et al. ⁶²	GMS + LPA	86.58
Proposed	FIR-Tiny YOLOv7	98.7

Table 8. Comparison with related work.

the proposed model is evaluated for mAP across all the split approaches. The generalization results with the proposed model on the CelebA and LFW datasets are presented in Table 9. The generalization results show that the model was able to get a good detection accuracy (mAP) when tested with the trained weights of the FAM Detection dataset in the detection of different faces present in the CelebA and LFW datasets due to similarity in facial images across the trained and tested datasets.

Conclusion

We proposed a one-step approach for face attribute manipulation and detection leading to facial identity recognition in few-shot and traditional scenarios. The proposed approach has been developed by leveraging deep learning-based approaches for image editing and object detection. The entire work has been carried out on a self-created Facial Attribute Manipulation (FAM) Detection Dataset generated using the latent space representation of StyleGAN3 inversion. Further, to perform facial identity recognition, we developed the FIR-Tiny YOLOv7 model which is an improvised variant of the Tiny YOLOv7 model. The Tiny YOLOv7 model has been improved by incorporating Swin Transformer Block (STB) and Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) into its feature extraction network. The Swin-Transformer Block (STB) was added to the proposed model to introduce the mechanism of attention and focus on localized features of the same face having varying attribute manipulations. The Squeeze-Excite Spatial Pyramid Pooling (SE-SPP) was developed to perform feature aggregation and produce larger feature maps that can be further utilized by the later layer of the proposed model. With the added enhancements, the proposed model achieved a 10.0% higher mAP in the one-shot scenario, a 30.4% higher mAP in the three-shot scenario, a 15.3% higher mAP in the five-shot scenario, and a 0.1% higher mAP in the traditional 70% – 30% split scenario as compared to the Tiny YOLOv7 model. Further, in comparison to the state-of-the-art Tiny YOLOv8 models, the proposed model achieved 0.2–3.2% higher mAP in different few-shot and traditional 70% – 30% split scenarios. Specifically for five-shot and traditional 70%-30% split scenarios, the proposed model utilized 5.1-12.9 M (Million) and 4.9-12.7 M (Million) lesser

Model	Dataset	Approach	mAP @ 0.50 (%)
FIR-Tiny YOLO v7 (Proposed)	CelebA ⁵²	One-shot	19.8
		Three-shot	38.2
		Five-shot	92.1
		Traditional (70%-30%)	94.5
	LFW ⁵³	One-shot	20.8
		Three-shot	41.2
		Five-shot	95.1
		Traditional (70%-30%)	96.5
	FAM	One-shot	23.5
		Three-shot	46.8
		Five-shot	98.4
		Traditional (70%-30%)	98.7

Table 9. Generalization results on celeba and LFW datasets.

training parameters as compared to the YOLOv8 Small and YOLOv8 NAS Small models. Further, in terms of detection speed, the proposed model achieved 1.1–11.4(ms) and 1.2–11.4(ms) lesser inference time as compared to the YOLOv8 Small and YOLOv8 NAS Small models. Future work in this domain can be further extended by improving the proposed model for facial identity recognition in one-shot and three-shot scenarios. The other scope is to collect images of law offenders and generate synthetic images using the StyleGAN3 inversion by manipulating their facial identities and developing a system that can be used by law enforcement agencies to catch suspects with manipulated identities. Moreover, Conditional-GANs (Co-GANs) can also be explored to reconstruct the real identities of the manipulated identities by inverting the facial features. The other potential lies in testing and improving the tiny variant of YOLOv9 which is yet to be released in the public domain.

Data availability

The created FAM Detection Dataset is made publicly available on the link: https://drive.google.com/file/d/1ThHs1MYbAECz-66aOnYQh1o0IYbJ_neD/view?usp=sharing Or, it can be requested from the First Author [Akhil Kumar].

Received: 26 November 2024; Accepted: 4 March 2025

Published online: 17 March 2025

References

- Davis, E. E., Matthews, C. M. & Mondloch, C. J. Ensemble coding of facial identity is robust, but May not contribute to face learning. *Cognition* **243**, 105668. <https://doi.org/10.1016/j.cognition.2023.105668> (Feb. 2024).
- Beltrán, M. & Calvo, M. A privacy threat model for identity verification based on facial recognition. *Computers Secur.* **132**, 103324. <https://doi.org/10.1016/j.cose.2023.103324> (Sep. 2023).
- de Ruiter, A. The distinct wrong of deepfakes. *Philos. Technol.* **34** (4), 1311–1332. <https://doi.org/10.1007/s13347-021-00459-2> (Jun. 2021).
- Nador, J. D., Vomland, M., Thielgen, M. M. & Ramon, M. Face recognition in Police officers: who fits the bill? *Forensic Sci. International: Rep.* **5**, 100267. <https://doi.org/10.1016/j.fsir.2022.100267> (Jul. 2022).
- Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. & Ortega-Garcia, J. Deepfakes and beyond: A Survey of face manipulation and fake detection, *Information Fusion*, vol. 64, pp. 131–148, Dec. (2020). <https://doi.org/10.1016/j.inffus.2020.06.014>
- Hou, X., Shen, L., Ming, Z. & Qiu, G. Deep generative image priors for semantic face manipulation. *Pattern Recogn.* **139**, 109477. <https://doi.org/10.1016/j.patcog.2023.109477> (Jul. 2023).
- Goodfellow, I. J. et al. Generative Adversarial Networks, arXiv.org, Jun. 10, 2014. <https://arxiv.org/abs/1406.2661>
- Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: unified, Real-Time object detection. *ArXiv Org.* **Jun 08**, (2015). <https://arxiv.org/abs/1506.02640>
- Aggarwal, A., Mittal, M. & Battineni, G. Generative adversarial network: an overview of theory and applications. *Int. J. Inform. Manage. Data Insights.* **1** (1), 100004. <https://doi.org/10.1016/j.jjime.2020.100004> (Apr. 2021).
- Rizvi, S. K. J., Azad, M. A. & Fraz, M. M. Spectrum of Advancements and Developments in Multidisciplinary Domains for Generative Adversarial Networks (GANs), *Archives of Computational Methods in Engineering*, vol. 28, no. 7, pp. 4503–4521, Apr. (2021). <https://doi.org/10.1007/s11831-021-09543-4>
- Jiang, P., Ergu, D., Liu, F., Cai, Y. & Ma, B. A review of Yolo algorithm developments. *Procedia Comput. Sci.* **199**, 1066–1073. <https://doi.org/10.1016/j.procs.2022.01.135> (2022).
- Pernuš, M., Štruc, V. & Dobrišek, S. MaskFaceGAN: High-Resolution face editing with masked GAN latent code optimization. *IEEE Trans. Image Process.* **32**, 5893–5908. <https://doi.org/10.1109/tip.2023.3326675> (2023).
- Han, L. et al. AE-StyleGAN: Improved Training of Style-Based Auto-Encoders, 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, pp. 955–964, (2022). <https://doi.org/10.1109/WACV51458.2022.00103>
- Abdal, R., Qin, Y. & Wonka, P. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space? 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 4431–4440, (2019). <https://doi.org/10.1109/ICCV.2019.00453>
- Abdal, R., Qin, Y. & Wonka, P. Image2StyleGAN++: how to edit the embedded images? *ArXiv Org.* **Nov 26**, (2019). <https://arxiv.org/abs/1911.11544>
- He, Z., Zuo, W., Kan, M., Shan, S. & Chen, X. AttGAN: Facial Attribute Editing by Only Changing What You Want, *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464–5478, Nov. (2019). <https://doi.org/10.1109/tip.2019.2916751>

17. Liu, M. et al. STGAN: A unified selective transfer network for arbitrary image attribute editing. *ArXiv Org. Apr* **22**, (2019). <https://arxiv.org/abs/1904.09709>
18. Li, X. et al. Image-to-image translation via hierarchical style disentanglement. *ArXiv Org. Mar.* **02**, (2021). <https://arxiv.org/abs/2103.01456>
19. Saha, R., Duke, B., Shkurti, F., Taylor, G. W. & Aarabi, P. LOHO: Latent Optimization of Hairstyles via Orthogonalization, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 1984–1993, (2021). <https://doi.org/10.1109/CVPR46437.2021.00202>
20. Alaluf, Y., Patashnik, O. & Cohen-Or, D. ReStyle: A Residual-Based stylegan encoder via iterative refinement. *ArXiv Org. Apr* **06**, (2021). <https://arxiv.org/abs/2104.02699>
21. Sun, R., Huang, C., Zhu, H. & Ma, L. Mask-aware photorealistic facial attribute manipulation. *Comput. Visual Media.* **7** (3), 363–374. <https://doi.org/10.1007/s41095-021-0219-7> (Apr. 2021).
22. Zhu, P., Abdal, R., Femiani, J. & Wonka, P. Barbershop, ACM Transactions on Graphics, vol. 40, no. 6, pp. 1–13, Dec. (2021). <https://doi.org/10.1145/3478513.3480537>
23. Dalva, Y., Altindis, S. F. & Dundar, A. VecGAN: Image-to-Image translation with interpretable latent directions. *ArXiv Org. Jul* **07**, (2022). <https://arxiv.org/abs/2207.03411>
24. Dalva, Y., Pehlivan, H., Moran, C., Hatipoğlu, Ö. I. & Dündar, A. Face attribute editing with disentangled latent vectors. *ArXiv Org. Jan* **11**, (2023). <https://arxiv.org/abs/2301.04628>
25. Yang, N., Luan, X., Jia, H., Han, Z. & Tang, Y. CCR: Facial image editing with continuity, consistency and reversibility. *ArXiv Org. Sep.* **22**, (2022). <http://arxiv.org/abs/2209.10734>
26. Jabberi, M., Wali, A. & Alimi, A. M. Generative Data Augmentation applied to Face Recognition, 2023 International Conference on Information Networking (ICOIN), Bangkok, Thailand, 2023, pp. 242–247. <https://doi.org/10.1109/ICOIN56518.2023.10049052>
27. Guo, Q. & Gu, X. Enhancing accuracy, diversity, and random input compatibility in face attribute manipulation, *Engineering Applications of Artificial Intelligence*, vol. 134, p. 108683, May (2024). <https://doi.org/10.1016/j.engappai.2024.108683>. Available: <https://doi.org/10.1016/j.engappai.2024.108683>.
28. Mohammadbagheri, N., Ayar, F., Nickabadi, A. & Safabakhsh, R. Identity-preserving editing of multiple facial attributes by learning global edit directions and local adjustments. *Comput. Vis. Image Underst.* **246**, 104047. <https://doi.org/10.1016/j.cviu.2024.104047> (Jun. 2024).
29. Xie, Y. et al. Controllable facial protection against malicious translation-based attribute editing. *Knowl. Based Syst.* 112873. <https://doi.org/10.1016/j.knsys.2024.112873> (Dec. 2024).
30. Liu, Y. & Chen, J. Multi-factor joint normalisation for face recognition in the wild, IET Computer Vision, vol. 15, no. 6, pp. 405–417, Apr. (2021). <https://doi.org/10.1049/cvi2.12025>
31. Bae, G. et al. DigiFace-1 M: 1 Million Digital Face Images for Face Recognition. *ArXiv Org. Oct.* **05**, (2022). <https://arxiv.org/abs/2210.02579>
32. Huang, G. B. et al. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, 2008 European Conference on Computer Vision (ECCV), Marseille, France, 2008. Available at: <https://hal.inria.fr/inria-00321923>
33. Colbois, L., Freitas Pereira, T. & Marcel, S. On the use of automatically generated synthetic image datasets for benchmarking face recognition, 2021 IEEE International Joint Conference on Biometrics (IJCB), Shenzhen, China, pp. 1–8, (2021). <https://doi.org/10.1109/IJCB52358.2021.9484363>
34. Boutros, F., Huber, M., Siebke, P., Rieber, T. & Damer, N. SFace: Privacy-friendly and Accurate Face Recognition using Synthetic Data, 2022 IEEE International Joint Conference on Biometrics (IJCB), Abu Dhabi, United Arab Emirates, pp. 1–11, (2022). <https://doi.org/10.1109/IJCB54206.2022.10007961>
35. Mandelli, S., Bonettini, N., Bestagini, P. & Tubaro, S. Detecting Gan-Generated Images by Orthogonal Training of Multiple CNNs, 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022, pp. 3091–3095. <https://doi.org/10.1109/ICIP46576.2022.9897310>
36. Wang, J., Tondi, B. & Barni, M. An Eyes-Based Siamese neural network for the detection of GAN-Generated face images. *Front. Signal. Process.* **2** <https://doi.org/10.3389/frsip.2022.918725> (Jul. 2022).
37. Gragnaniello, D. et al. Are GAN Generated Images Easy to Detect? A Critical Analysis of the State-Of-The-Art, 2021 IEEE International Conference on Multimedia and (ICME), Shenzhen, China, 2021, pp. 1–6. <https://doi.org/10.1109/ICME51207.2021.9428429>
38. Pasquini, C. et al. Jan., Identifying synthetic faces through GAN inversion and biometric traits analysis, *Applied Sciences*, **13**, 2, p. 816, (2023). <https://doi.org/10.3390/app13020816>
39. Ren, X., Zhang, W., Wu, M., Li, C. & Wang, X. May, Meta-YOLO: Meta-Learning for Few-Shot traffic sign detection via decoupling dependencies, *Applied Sciences*, **12**, 11, p. 5543, (2022). <https://doi.org/10.3390/app12115543>
40. Chatterjee, R., Chatterjee, A., Islam, S. H. & Khan, M. K. An object detection-based few-shot learning approach for multimedia quality assessment. *Multimedia Syst.* **29** (5), 2899–2912. <https://doi.org/10.1007/s00530-021-00881-8> (Jan. 2022).
41. Xia, R., Li, G., Huang, Z., Meng, H. & Pang, Y. Bi-path combination YOLO for Real-time Few-shot object detection. *Pattern Recognit. Lett.* **165**, 91–97. <https://doi.org/10.1016/j.patrec.2022.11.025> (Jan. 2023).
42. Xu, Y. et al. Diff-PC: Identity-preserving and 3D-aware controllable diffusion for zero-shot portrait customization. *Inform. Fusion.* **117**, 102869. <https://doi.org/10.1016/j.inffus.2024.102869> (Dec. 2024).
43. Karras, T. et al. Alias-Free generative adversarial networks. *ArXiv Org. Jun* **23**, (2021). <https://arxiv.org/abs/2106.12423>
44. Wang, C. Y., Bochkovskiy, A. & Liao, H. Y. M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *ArXiv Org. Jul* **06**, (2022). <https://arxiv.org/abs/2207.02696>
45. Liu, Z. et al. Swin transformer: hierarchical vision transformer using shifted windows. *ArXiv Org. Mar.* **25**, (2021). <https://arxiv.org/abs/2103.14030>
46. Karras, T., Laine, S. & Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 43, no. 12, pp. 4217–4228, (2021). <https://doi.org/10.1109/TPAMI.2020.2970919>
47. Zhang, R., Isola, P., Efros, A., Shechtman, E. & Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 586–595. (2018) pp.
48. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image Quality Assessment: From Error Visibility to Structural Similarity, IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, Apr. (2004). <https://doi.org/10.1109/tip.2003.819861>
49. Diwan, T., Anirudh, G. & Tembhurne, J. V. Object detection using YOLO: challenges, architectural successors, datasets and applications. *Multimedia Tools Appl.* **82** (6), 9243–9275. <https://doi.org/10.1007/s11042-022-13644-y> (Aug. 2022).
50. Anusudha, K. & A.N and Real time face recognition system based on YOLO and insightface. *Multimedia Tools Appl.* **83** (11), 31893–31910. <https://doi.org/10.1007/s11042-023-16831-7> (Sep. 2023).
51. Kumar, A., Kalia, A., Verma, K., Sharma, A. & Kaushal, M. Scaling up face masks detection with YOLO on a novel dataset. *Optik* **239**, 166744. <https://doi.org/10.1016/j.ijleo.2021.166744> (Aug. 2021).
52. He, K., Zhang, X., Ren, S. & Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904–1916, Sep. (2015). <https://doi.org/10.1109/tpami.2015.2389824>

53. AlexeyAB GitHub - AlexeyAB/darknet: YOLOv4 / Scaled-YOLOv4 / YOLO - Neural networks for object detection (Windows and Linux version of Darknet), GitHub. <https://github.com/AlexeyAB/darknet>
54. Chen, B. et al. Locally GAN-generated face detection based on an improved Xception, *Information Sciences*, vol. 572, pp. 16–28, Sep. (2021). <https://doi.org/10.1016/j.ins.2021.05.006>
55. Songsri-In, K. & Zafeiriou, S. Complement face forensic detection and localization with FacialLandmarks. *ArXiv Org. Oct. 12*, (2019). <https://arxiv.org/abs/1910.05455>
56. Uittenhove, K., Shahreza, H. O., Marcel, S., Ramon, M. & SYNTHETIC AND NATURAL FACE IDENTITY PROCESSING SHARE COMMON MECHANISMS. *Computers Hum. Behav. Rep.*, **100563**, <https://doi.org/10.1016/j.chbr.2024.100563> (Dec. 2024).
57. Villegas-Ch, W., Navarro, A. M. & Mera-Navarrete, A. Using generative adversarial networks for the synthesis of emotional facial expressions in virtual educational environments. *Intell. Syst. Appl.* p. **200479** <https://doi.org/10.1016/j.iswa.2025.200479> (Feb. 2025).
58. Cao, W., Feng, Z., Zhang, D. & Huang, Y. Facial expression recognition via a CBAM embedded network. *Procedia Comput. Sci.* **174**, 463–477. <https://doi.org/10.1016/j.procs.2020.06.115> (Jan. 2020).
59. Shen, T. & Xu, H. Facial expression recognition based on Multi-Channel attention residual network. *Comput. Model. Eng. Sci.* **135** (1), 539–560. <https://doi.org/10.32604/cmescs.2022.022312> (Sep. 2022).
60. Liu, C., Liu, X., Chen, C. & Zhou, K. Deep global Multiple-Scale and local patches attention Dual-Branch network for Pose-Invariant facial expression recognition. *Comput. Model. Eng. Sci.* **139** (1), 405–440. <https://doi.org/10.32604/cmescs.2023.031040> (Dec. 2023).
61. Yang, S., Luo, P., Loy, C. C. & Tang, X. From Facial Parts Responses to Face Detection: A Deep Learning Approach. 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, Dec. (2015). <https://doi.org/10.1109/iccv.2015.419>
62. Liu, Z., Luo, P., Wang, X. & Tang, X. Deep Learning Face Attributes in the Wild, IEEE International Conference on Computer Vision (ICCV), IEEE, Dec. (2015). <https://doi.org/10.1109/iccv.2015.425>

Acknowledgements

Not applicable.

Author contributions

Conceptualization, implementation and data analysis were performed by [Akhil Kumar] and [Swarnava Bhattacharjee]. Material and data preparation were performed by [Akhil Kumar], [Swarnava Bhattacharjee] and [Amrisha Kumar]. The supervision of the work was performed by [Dushantha Nalin K. Jayakody]. The first draft of the manuscript was written by [Akhil Kumar] and [Swarnava Bhattacharjee]. All authors reviewed the final manuscript.

Funding

This work was supported, in part, by the Scheme for Promotion of Academic & Research Collaboration (SPARC), Government of India, via grant no. SPARC/2024-2025/NXTG/P3524, by the Sri Lanka Institute of Information Technology through the grantPVC(R&I)/RG/2024/12, by the COFAC - Cooperativa de Formação e Animação Cultural, C.R.L. (University of LusófonaUniversity), via the project PortuLight (COFAC/ILIND/COPELABS/2/2023), by the national funds through FCT - Fundação para a Ciência e a Tecnologia - as part of the projects ULLC-UAV (2023.08191.CEECIND), UNINOVA-CTS (UIDB/00066/2020) and COPELABS (no. UIDB/04111/2020).

Declarations

Competing interests

The authors declare no competing interests.

The authors declare that they have no known potential conflict of interest(s) to disclose.

Ethical statement

No human participants have directly participated in this research work. The FAM dataset and the corresponding images used in this manuscript for facial identities have been taken from Flickr-Faces-HQ (FFHQ) dataset. The FFHQ dataset is a publicly available dataset for facial imagery released by NVIDIA Labs and no permission is required to use these images. The readers can access the FFHQ dataset from: <https://github.com/NVLabs/ffhq-dataset>.

Additional information

Correspondence and requests for materials should be addressed to D.N.K.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025