This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

On Which Data Distribution (Synthetic or Real) We Should Rely for Soft Biometric Classification

Manju R A*, Atul Kumar*, Akshay Agarwal
Trustworthy BiometraVision Lab, IISER Bhopal
{manjura, atulk23, akagarwal}@iiserb.ac.in

Abstract

Identification of gender is critical not only for human-computer interaction but also for scrutinizing the search space in which an identity needs to be determined. Traditionally, "real" facial images are employed for gender identification by computer vision algorithms. Due to the tremendous rise of privacy and advancement in generative networks, synthetic face images are heavily developed and can be used for several face-related studies including gender classification. However, their effectiveness compared to real images is still unexplored for gender classification. In response, this study explores the effectiveness of gender classification networks trained on real and synthetic face images, offering novel insights into the effectiveness of these two data distributions. For that, we implemented several state-of-the-art gender classification architectures covering convolutional neural networks (CNNs) and vision transformers (ViT). Our research builds on the rigorous evaluation of 8 Deep Neural Networks (DNNs) across 4 diverse datasets and 6 types of image corruptions. To make the research interpretable, we have also used several explainable mechanisms, including Grad-CAM and t-SNE visualizations. In brief, the impact of the proposed research is multifold: (i) understand the effectiveness of real vs. synthetic data distributions in network training and (ii) whether the synthetic models reflect the true physical world distribution to ensure that the models trained on them are resilient against image perturbations.

1. Introduction

The impact of social media and machine learning applications in our day-to-day lives on different genders is ubiquitous. It ranges from highlighting a variety of jobs to recommending products on social media platforms. For gender identification, face is one of the primary data sources for several machine learning algorithms. Not only gender but also face covers several important attributes, including expression, identity, and ethnicity [6]. Therefore, the use of face images raised several ethical and privacy concerns. Further, the collection of face images of every demographic entity present in the physical world and labeling is a challenging task. These challenges can be overcome by utilizing the strength of Generative Artificial Intelligence (GenAI) architectures and generating synthetic face images [8, 31]. Research studies recently have started exploring the use of generative networks to produce high-quality face images that can reflect the variation present in the physical world faces [9, 21, 28, 41].

Inspired by the success of synthetic image generation algorithms and the importance of the collection of physical world datasets, we have performed an extensive gender classification study using these two domains of the dataset. We assert that understanding the effectiveness of these domain datasets in training machine learning classifiers can ensure the balance between privacy and the effectiveness of an image classifier. Henceforth, we have selected several state-of-the-art machine learning classifiers covering convolutional neural networks and transformer architectures. Further, we are aware, that the physical world is unconstrained due to distortions that occur during image collection or transmission [33], and images might undergo unknown corruptions [30]. Consequently, it is essential to evaluate the impact of wrong decisions made by DNNs before their deployment in these uncontrolled and variable environments. To ensure that the proposed models are not only effective but also resilient against unseen corruptions, we tested the trained classifier against test images that are corrupted through 6 image corruptions. We believe that this experiment will also help in understanding whether synthetic images are useful in covering the physical world distributions better than real-world images, which might make them resilient against physical world corruption. In brief, the primary contributions of this paper are:

- To the best of our knowledge, this is the first work that paves the way in understanding the effectiveness of

*Equal author contribution.

real vs. synthetic face images in training deep learning architectures, including both CNN and transformer, for gender classification;

- Further, for the first time, we evaluated the robustness of deep architectures aimed at gender classification through understanding which domain (real vs. synthetic) reflects better physical world image distributions;

- An extensive experimental evaluation has been conducted to demonstrate the strength of pure convolutional architectures and attention architectures.

2. Related Work

The research community has recently begun to explore using convolutional neural networks (CNNs) for analyzing gender from facial data under unrestricted conditions [13]. [2, 14, 26] analyzed age and gender from facial data in challenging environments. Levi and Hassner [24] investigated a five-layer CNN for classifying age and gender in uncontrolled settings. Shaik and Micheal [35] developed a 5-layer CNN model for this purpose using the Adience database. Zhang et al. [42] proposed a new CNN model called "Residual Networks of Residual Networks (RoR)" for the same application. Conversely, Duan et al. [12] introduced "CNN2ELM," a hybrid model that combines three CNN models with two Extreme Learning Machine (ELM) architectures for predicting gender. This model is effective in analyzing facial data in real-world scenarios, including images from constrained environments. Recent advancements in age classification techniques are thoroughly reviewed in [3]. Initially, researchers focused on identifying manually adjusted facial features and leveraged their variations [1]. While these early methods achieved significant accuracy on limited datasets, few addressed the challenges posed by variations in image quality and clarity found in real-world scenarios [1]. The work in [16] provides an extensive review of gender classification techniques, including the use of neural networks dating back to 1990 [4]. Support vector machines (SVMs) have also been used to achieve low error rates in gender prediction from low-resolution images [18]. Kumar and Arthanariee [23] employed a feature-based approach, identifying critical facial points like nostrils, mouth, and eyes, to locate missing children. Apart from the CNN-based architecture, very limited studies showcase the effectiveness of transformers for gender classification. Singh and Singh [36] use a hybrid transformer-sequencer model to classify gender from in-the-wild facial images by capturing both local and global features. Suravarapu and Patil [37] apply vision transformers for gender classification using periocular images, leveraging their ability to capture fine-grained visual features.

Dataset	Images	Gender Attribute	
		Male	Female
HDA_SynChildFaces [17]	8260	4925	3335
UTKFace [43]	8802	4651	4151
AgeDB [27]	6596	3916	2680
FairFace [20]	8122	4186	3936

Table 1. Comprehensive overview of the diverse datasets used.

As we mentioned earlier, limited studies have been conducted to understand the impact of data distribution in the training of DNN for gender classification. Moreover, no existing study has demonstrated the robustness of gender classification models whether trained on real or synthetic images against noise corruption. A few studies that have been done preliminary studies using synthetic images are: Oulad-Kaddour et al. [29] demonstrate that facial synthetic datasets can effectively train gender recognition models. Davarci and Anarim [10] generate the gait synthetic data to train the gender detection model. Escalante and Wiskot [15] proposed slow feature analysis (SFA), a versatile unsupervised learning algorithm to estimate gender and age from synthetic face images. Recently, Atzori et al. [5] studied the impact of the demographically balanced dataset on face recognition; however, due to the use of real data in combination with synthetic data left concerns about privacy leakage.

3. Proposed Gender Classification Setup

This section first describes the datasets used for training and testing, followed by the perturbations applied to the test data to assess the robustness of the gender classification models. Next, the ViT and CNN-based models utilized for gender classifications are described followed by the description of the experimental protocol used for the evaluation of these models.

3.1. Datasets

The characteristics of the real and synthetic datasets used in this research are given in Table 1.

3.1.1 Real Datasets

To comprehensively evaluate gender classification models under various conditions, we utilize three diverse real-world datasets: UTKFace [43], AgeDB [27], and FairFace [20]. Each dataset offers unique characteristics and challenges, enabling a thorough analysis of model performance across demographic attribute prediction using different network modalities. The UTKFace [43] dataset contains face images with a wide age range, spanning from 0 to 116 years. It consists of over 20,000 images, each annotated with age, gender, and ethnicity labels. The dataset falls into the wild category, encompassing various factors such as pose, facial

Figure 1. A visual exploration of the diversity present in real and synthetic datasets (2 images of each). From left: face images from the UTKFace [43], AgeDB [27], FairFace [20], and HDASynChildFace dataset [17].

expression, illumination, occlusion, and resolution. For this study, a subset of 8802 images from UTKFace is selected to balance gender representation (4651 males, 4151 females) and to ensure a mix of age, pose, and illumination variations, making the data representative of real-world diversity. The AgeDB [27] dataset consists of 6596 facial images from 440 subjects in celebrity, politics, and science. Gender annotations include 2680 females and 3916 males. The FairFace [20] dataset is a balanced dataset with 8122 face images. It addresses the imbalance issue in the UTKFace dataset. Gender annotations include 3936 for females and 4186 for males. Each of the real datasets is divided into training and testing splits and only a real test set is used for evaluation.

3.1.2 Synthetic Dataset

The HDASynChildFace dataset [17] is a Synthetic (Syn) dataset created using the generative model StyleGAN3 and advanced techniques to simulate various facial features. The dataset includes $1,652$ subjects categorized into six age groups: 20+, 16-13, 13-10, 10-7, 7-4, and 4-1 years, providing a diverse range of facial data across different age brackets. In this study, we used 8260 images covering all age groups, with 3335 females and 4925 males. While the dataset exhibits some imbalance between male and female representations, it offers significant advantages. Firstly, HDASynChildFace enables the creation of large-scale, privacy-friendly datasets that are particularly valuable when dealing with sensitive demographics, such as children. Additionally, the dataset captures facial characteristics from a broad range of age groups, including both children and adults, ensuring diversity in facial features. This diversity allows models trained on this dataset to generalize across various age ranges, making the synthetic dataset applicable even when test sets predominantly contain adult faces. By incorporating a wide age range, HDASynChildFace provides a comprehensive dataset that enhances the model's robustness in real-world scenarios. The dataset is also used to train the models. Figure 1 showcases a few samples of the real and synthetic data distribution.

3.2. Perturbations

To rigorously assess the robustness of gender classification models, we applied various perturbations to the real test subsets (UTKFace, AgeDB, and FairFace). These perturbations simulate common distortions encountered in real-world scenarios. The perturbations applied are: *Color Saturation (CS)*, *Color Contrast (CC)*, *Blockwise Distortion (BW)*, *Gaussian Noise in Color Component (GNC)*, *Gaussian Blur (GB)*, and *JPEG Compression (JPEG)*. Examples of these perturbed images are shown in Figure 2, and the description of the perturbations applied over the real datasets is as follows:

1. ***Color Saturation (CS)***: Color saturation adjusts the intensity of colors, either enhancing or dulling them. This is useful for testing how models handle images with altered color richness. The saturation parameter controls the degree of adjustment. The color saturation distortion is defined as: $\mathrm{CS}(x, \alpha) = x \cdot \alpha$. Where x is the input image and α is the saturation factor. The parameter values for α used for this perturbation are:$\alpha \in \{0.4, 0.3, 0.2, 0.1, 0.0\}$ which represent decreasing levels of saturation from high to low. Among all the specified values, a perturbation value of 0.2 has been utilized.

2. ***Color Contrast (CC)***: Color contrast modification alters the contrast levels of the image, simulating varied lighting conditions. The contrast parameter controls the adjustment. The color contrast distortion is defined as: $\mathrm{CC}(x, \beta) = x \cdot \beta + (1 - \beta) \cdot \mu$. Where, β is the contrast factor, and μ is the mean intensity of the image. The parameter values for β used are: $\beta \in \{0.85, 0.725, 0.6, 0.475, 0.35\}$ representing varying contrast levels, from high to low contrast. A perturbation value of 0.6 has been utilized.

3. ***Blockwise Distortion (BW)***: Blockwise distortion applies a block-wise corruption to the image, degrading specific sections to simulate partial occlusion or artifacts. The parameter for blockwise distortion is the block size. The blockwise distortion is defined as:

$$\mathrm{BW}(x, b) = \sum_{i=1}^{n/b} \sum_{j=1}^{m/b} \frac{x_{i,j}}{b^2}$$

   where, b is the block size, and $x_{i,j}$ is a block of pixels in the image. The block size values b used for this distortion are:$b \in \{16, 32, 48, 64, 80\}$ with increasing block sizes, representing varying levels of corruption from small to large block occlusions. A perturbation value of 48 has been utilized.

4. ***Gaussian Noise in Color Component (GNC)***: Gaussian noise adds random noise with a Gaussian distribution to the image's color components, mimicking sensor noise or poor image quality. The noise level parameter determines the standard deviation of the noise.

Figure 2. Example perturbed images of UTKFace [43], AgeDB [27] and FairFace [20] dataset. From left to right, images of clean and six different perturbations (leftmost is BW, followed by CC, CS, GB, GNC, and JPEG) are showcased to reflect the variation they bring to the visual appearance of faces.

The Gaussian noise in the color component distortion is defined as:$\mathrm{GNC}(x, \sigma) = x + \mathcal{N}(0, \sigma^2)$. Where σ is the standard deviation of the Gaussian noise. The noise standard deviation σ is 0.005 has been utilized.

5. ***Gaussian Blur (GB)***: Gaussian blur smooths the image by averaging pixels with their neighbors, simulating defocus or motion blur. The blur level parameter determines the kernel size of the Gaussian filter. The Gaussian blur distortion is defined as: $\mathrm{GB}(x, k) = x * G(k)$. Where, k is the kernel size which is set to the value of 13, and $G(k)$ is the Gaussian kernel.

6. ***JPEG Compression (JPEG)***: JPEG compression reduces the image quality by approximating the image data, which is typical in lossy compression formats. The compression level parameter controls the quality factor of JPEG encoding. In this research, a compression quality value of 4 has been utilized.

3.3. Gender Attribute Recognition Architectures

In this work, we explore the efficacy of various deep-learning architectures for gender classification. We investigate two prominent approaches: ViT-based models and CNN-based models. Each model is trained on synthetic and real-world datasets to evaluate the effectiveness of each data distribution. Figure 3 illustrates the architecture diagram of the gender classification setup, highlighting the flow from training on synthetic and real datasets to testing on clean and perturbed real images. This setup ensures a comprehensive evaluation of the model's robustness and effectiveness in gender prediction across different scenarios.

3.3.1 Vision Transformers For Gender Recognition

The vision transformer models leverage self-attention mechanisms to process images, providing a robust framework for capturing complex patterns. BEiT (Bidirectional Encoder representation from Transformers) [7] employs a

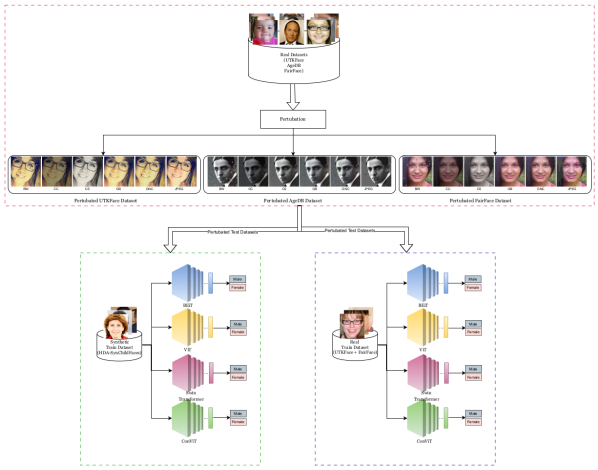

Figure 3. Schematic diagram of the proposed gender classification setup.

pre-training method that leverages masked image modeling, akin to masked language modeling in NLP. It splits images into patches and processes them using transformer blocks, making it adept at capturing intricate visual features. ViT [11] similarly splits images into patches and processes them using standard transformer blocks, effectively capturing global dependencies within images. Swin-T (Shifted Window Transformer) [25] utilizes a hierarchical architecture with shifted windows, enabling it to capture both local and global dependencies. This design enhances its ability to model spatial hierarchies in images. CViT (Compact Vision Transformer) [32] is designed to be more efficient, reducing the computational overhead of standard ViTs while maintaining high performance, making it particularly useful for scenarios requiring computational efficiency. These vision transformer models are pre-trained on the ImageNet subset dataset and fine-tuned on the training set of synthetic and real datasets.

3.3.2 CNNs For Gender Recognition

The CNN models are another form of popular deep networks that rely on convolutional layers to extract features from images and we have several CNNs having wide architectural configurations. For instance, ResNet-50 [22] introduces residual connections, which help mitigate the vanishing gradient problem and allow the training of deeper networks, facilitating the learning of robust features across different scales. DenseNet-121 [19] connects each layer to every other layer in a feed-forward fashion, promoting feature reuse and enhancing gradient flow, resulting in efficient feature propagation and reduced parameter redundancy. InceptionV3 (IV3) [38] employs a combination of convolutions of different sizes within its architecture, capturing multi-

Train	Models	Test Dataset																	
		UTKFace						AgeDB						FairFace					
		CS	CC	BW	GNC	GB	JPEG	CS	CC	BW	GNC	GB	JPEG	CS	CC	BW	GNC	GB	JPEG
Real	BEiT	94.39	94.45	94.41	92.44	93.52	87.14	97.02	96.99	96.10	92.66	94.22	89.20	90.21	90.43	90.56	84.90	88.41	72.11
	ViT	93.64	93.98	93.91	92.27	93.37	88.65	96.11	96.24	96.02	94.23	94.36	90.85	88.18	88.62	89.02	83.66	87.25	70.10
	SwinT	92.76	92.83	93.01	91.25	92.11	86.58	96.64	96.68	96.52	93.53	91.75	88.49	86.63	87.51	87.98	81.74	85.82	66.98
	CViT	87.21	88.44	88.52	83.7	87.26	79.59	91.04	91.38	90.74	84.54	84.14	82.04	78.31	79.41	80.16	73.53	78.60	64.87
Syn	BEiT	88.53	88.41	88.41	78.49	85.28	78.14	93.61	93.60	91.02	79.97	90.68	83.96	81.40	80.94	81.23	61.73	75.64	65.02
	ViT	90.08	90.30	90.35	87.31	88.30	86.40	94.15	94.66	93.75	86.71	88.85	84.67	82.03	83.08	83.49	79.72	81.14	72.21
	SwinT	91.66	91.66	91.86	89.21	91.18	85.28	96.08	96.02	95.85	92.14	92.77	89.00	83.78	84.63	85.12	79.74	83.45	70.42
	CViT	84.34	80.88	80.15	82.50	77.72	71.29	90.01	88.5	87.74	85.43	80.23	79.23	75.33	73.65	73.28	72.22	71.86	64.94

Table 2. Performance of transformer-based models across various perturbations. The models are evaluated under six types of perturbations: Color Saturation (CS), Color Contrast (CC), Blockwise Distortion (BW), Gaussian Noise in Color Component (GNC), Gaussian Blur (GB), and JPEG Compression (JPEG). The green cells indicate the best performance, while the blue cells represent the second-best performance in each category.

scale features effectively. This multi-path design allows for efficient computation and high performance. EfficientNet-B0 (Eff-B0) [39] scales the depth, width, and resolution of the network in a balanced manner, achieving high accuracy with fewer parameters and optimizing performance across various image classification tasks.

3.4. Experimental Protocol

In this research, two data distributions are considered: one is the synthetic distribution and another is the real data distribution. When the synthetic distribution is considered, we used 8260 face images from the HDASynChildFace dataset to train the models. For the real distribution training, we have first divided the UTKFace and FairFace datasets into training and testing subsets. Later, the combination of the train sets of both UTKFace and FairFace datasets, with 4130 images from UTKFace and 4130 images from FairFace train sets, respectively are used for training the models. The testing is conducted on the remaining/unseen test set of all the real datasets with 8802 images from UTKFace, 6596 images from AgeDB, and 8122 images from FairFace for clean data. Further, for the fair robustness evaluation, the same test sets are perturbed with six different types of distortions: blockwise distortion, color saturation, color contrast, Gaussian noise in the color component, Gaussian blur, and JPEG compression. The training process of each network is carried out using a batch size of 32, a learning rate of 0.0001, and an Adam optimizer over 10 epochs.

4. Experimental Results and Analysis

This section provides a comprehensive evaluation of soft biometric gender attribute classification models trained on synthetic and real-world datasets, tested on both clean and perturbed real-world datasets. The performance metrics, particularly accuracy, are used to assess the robustness and effectiveness of the models.

The performance of models trained on real-world and synthetic datasets is analyzed to understand the impact of

Train	Models	Test Dataset			
		UTKFace	AgeDB	FairFace	Average
Real	BEiT	94.45	96.96	90.55	**93.32**
	ViT	93.90	96.36	89.07	93.11
	SwinT	93.04	96.74	87.98	92.59
	CViT	88.50	91.48	80.13	86.04
Syn	BEiT	88.28	93.71	81.25	87.75
	ViT	90.39	94.56	83.55	89.50
	SwinT	91.77	96.09	84.87	**90.91**
	CViT	80.14	88.65	73.33	80.71

Table 3. Comprehensive overview of transformer-based model performance across various real-world datasets. The green color indicates the best-performing model for both training setups.

training data quality on model accuracy. The findings can be broadly divided into multiple categories based on the deep network used for classification, data distribution of the training data, robustness of the models against corruption, and role of corruption augmentation in training.

The performance of Vision Transformer (ViT) models on clean and corrupted data are summarized in Table 2 and Table 3, respectively. Whereas, the findings of CNNs on clean and perturbed real test sets are presented in Table 4 and Table 5, respectively.

4.1. Analysis of Training Data Distribution

Broadly, the findings suggest that models trained on real data consistently outperform those trained on synthetic datasets, regardless of whether CNN or transformer architectures are used. For example, when the BEiT model is used for classification, it yields an average gender classification accuracy of 93.32% when the real distribution is used for training. However, the network observed an accuracy drop and yield and average accuracy of 87.75% when the synthetic distribution is used for training the model. A similar observation can be observed in the case of CNN models where each model yields higher classification when they are trained on the real distribution as compared to the synthetic

Train	Models	Test Dataset																	
		UTKFace						AgeDB						FairFace					
		CS	CC	BW	GNC	GB	JPEG	CS	CC	BW	GNC	GB	JPEG	CS	CC	BW	GNC	GB	JPEG
Real	ResNet-50	70.79	69.96	73.65	61.34	68.55	58.79	69.70	69.20	68.54	57.74	65.11	60.20	61.74	62.50	63.85	37.76	59.96	54.69
	DNet-121	79.37	79.50	83.09	52.79	78.15	54.03	77.24	77.66	75.66	57.44	69.57	58.70	69.92	68.82	71.65	51.53	68.39	52.52
	IV3	73.11	75.21	76.82	54.19	73.68	62.29	75.36	75.60	74.37	52.95	63.12	57.95	64.63	65.02	66.07	53.25	63.83	56.32
	Eff-B0	76.32	78.18	80.75	49.64	70.19	56.08	69.14	70.96	71.52	46.48	63.96	57.11	65.58	66.30	68.49	50.54	63.80	51.95
Syn	ResNet-50	59.31	60.09	61.62	56.56	60.31	57.33	67.98	66.66	65.08	59.80	59.46	57.86	57.05	58.97	58.36	51.28	56.43	52.10
	DNet-121	48.56	49.00	49.73	48.23	49.00	47.18	46.11	47.51	45.87	51.10	43.08	40.64	49.51	49.97	50.54	48.51	49.61	48.43
	IV3	54.13	53.11	54.40	49.18	49.84	48.71	60.67	58.80	58.27	46.72	50.42	48.37	52.89	53.00	54.19	50.62	50.84	50.62
	Eff-B0	51.93	56.48	55.04	54.41	47.64	53.47	59.80	59.91	58.68	51.09	44.13	56.33	52.73	55.90	55.22	51.09	49.76	50.75

Table 4. Performance of CNN models across various perturbations. The models are evaluated under six types of perturbations: Color Saturation (CS), Color Contrast (CC), Blockwise Distortion (BW), Gaussian Noise in Color Component (GNC), Gaussian Blur (GB), and JPEG Compression (JPEG). The green cells indicate the highest accuracy, while the blue cells denote the second-highest accuracy for each perturbation type.

Train	CNN Models	Test Dataset			
		UTKFace	AgeDB	FairFace	Average
Real	ResNet-50	73.64	71.81	63.78	69.08
	DNet-121	83.00	78.30	72.21	**77.84**
	InceptionV3	76.83	76.42	65.47	72.91
	EfficientNetB0	80.59	73.01	68.52	74.71
Syn	ResNet-50	61.69	68.25	58.49	**62.81**
	DNet-121	49.82	48.60	50.68	49.70
	InceptionV3	53.78	61.30	53.80	56.29
	EfficientNetB0	55.05	59.77	55.06	56.63

Table 5. Comprehensive overview of CNN model performance across various real-world datasets. The green color indicates the best-performing model for both training setups.

distribution. For instance, the performance of the best performing CNN model, i.e., DenseNet-121 (DNet-121) drops from 77.84% to 49.70% as soon the training distribution changed from real to synthetic.

4.2. Analysis of Classifier

As observed above the real distribution is effective compared to the synthesis distribution when training the gender classifier. However, the disparity in performance is high when the CNN models are used for classification. For example, the InceptionV3 model shows a performance gap of 16.62% when the training of the model moved from real to synthetic distribution. It is interesting to note that, only the ResNet-50 model significant level of resiliency against training distribution and yields the lowest level of reduction in performance as compared to the other CNNs. However, the performance of CNNs can drop up to 28.14% when the distribution of training data changes, where the real distribution outperforms the synthetic distribution. It is interesting to note that DNet-121 which shows the best performance when trained on real distribution yields performance close to a random level when the synthetic distribution is used for training. The ViT model also shows the trend of accuracy reduction when the models trained on the synthetic distribution as compared to their training on real distribution. However, the reduction is significantly lower than what we observed in the case of CNNs. For example, the best performing ViT model, i.e., BEiT, yields a drop of 5.57% which is a maximum reduction across all the ViTs used. *This indicates that transformer models exhibit less sensitivity to the distribution of training data, maintaining stronger performance even when trained on synthetic datasets.* Further, as compared to the BEiT model, on the synthesis distribution, the SwinT model performs best yielding an average accuracy of 90.91% on the clean real test sets.

4.3. Robustness Analysis

The robustness analysis can be broadly done based on the classifier architecture, i.e., CNN vs Transformer, and distribution of training data, i.e., real vs. synthetic. As mentioned, the BEiT model demonstrated the highest level of effectiveness in performing gender classification, the model is found robust against the majority of the corruptions. Surprisingly, not just the BEiT model but each model is found robust in handling each of the unseen common corruption which is only used for evaluation (testing), except JPEG compression. Another corruption namely Gaussian noise in color component (GNC) corruption also able to reduce the performance of BEiT to 2.01%, 4.3%, and 7.65% on UTKFace, AgeDB, and FairFace datasets, respectively. Similar findings are observed across ViTs where the models are highly sensitive to JPEG compression and are found robust in handling other noise corruptions. The same form of vulnerability against corruption is found in CNNs as well where they show significant reduction when the test images are perturbed with GNC and JPEG compression.

Furthermore, the models are found less effective when trained on synthetic than the models trained on real distribution, their sensitivity against corruption is similar to what they are when trained on the real distribution. For example, the SwinT model which yields the best performance when trained on synthetic distribution suffers a drop of 6.49%, 7.09%, 14.45% at most on the UTKFace. AgeDB, and FairFace datasets, respectively. Across all the corruptions,

Train	Models	Test Dataset																	
		UTKFace						AgeDB						FairFace					
		CS	CC	BW	GNC	GB	JPEG	CS	CC	BW	GNC	GB	JPEG	CS	CC	BW	GNC	GB	JPEG
Syn	BEiT	90.69	90.53	90.82	86.16	89.15	85.06	93.89	93.66	93.84	90.31	90.29	87.62	81.45	81.77	82.07	77.93	81.13	65.47
	ViT	90.27	90.43	90.26	89.03	89.35	86.13	91.79	92.52	90.61	88.90	89.90	85.27	81.08	81.66	81.99	78.93	80.38	69.35
	SwinT	81.72	73.68	72.99	80.60	73.29	74.26	87.81	83.41	82.70	81.89	78.41	79.53	72.78	70.42	69.08	69.53	70.54	62.06
	CViT	88.93	88.53	88.36	84.51	89.45	81.28	92.19	92.07	90.97	87.29	89.37	83.11	79.35	79.25	78.51	71.18	79.24	67.98
Syn	ResNet-50	67.93	67.40	67.55	63.21	67.35	58.99	70.19	72.08	68.32	64.20	62.76	58.24	62.10	62.71	62.47	59.18	61.47	55.19
	DNet-121	64.16	57.88	60.47	53.64	59.08	56.82	66.54	62.20	62.28	51.83	58.02	55.57	61.25	57.65	59.24	49.96	57.44	56.67
	IV3	57.38	57.48	61.21	51.28	56.16	56.10	65.22	63.87	62.11	57.23	55.38	57.26	55.61	55.92	57.60	50.14	55.28	55.55
	Eff-B0	59.91	64.15	64.59	58.41	62.75	54.83	64.29	64.46	61.74	60.43	56.80	55.32	57.84	59.08	59.55	55.39	57.91	55.29

Table 6. Performance of transformer and CNN models across various perturbations trained on the augmented data. The models are evaluated under six types of perturbations: Color Saturation (CS), Color Contrast (CC), Blockwise Distortion (BW), Gaussian Noise in Color Component (GNC), Gaussian Blur (GB), and JPEG Compression (JPEG). The green cells indicate the highest accuracy, while the blue cells denote the second-highest accuracy for each perturbation type.

Train	Models	Test Dataset		
		UTKFace	AgeDB	FairFace
Syn	BEiT	90.75	93.78	82.13
	ViT	90.29	92.14	82.06
	SwinT	73.14	83.58	69.20
	CViT	88.54	92.17	78.56
Syn	ResNet-50	67.62	72.31	62.59
	DNet-121	60.61	66.97	59.06
	InceptionV3	61.11	64.46	57.35
	EfficientNetB0	64.51	65.29	59.20

Table 7. Comprehensive overview of transformer and CNN models performance across various real-world datasets trained on augmented data. The green color indicates the best-performing model for both training setups.

JPEG compression is found to be more stealthy than others. It is also observed that the models are sometimes not only able to retain the accuracy on the corruption set but showcase slightly better accuracy than the clean test set. Out of several CNNs used, ResNet-50 is found best when the models are trained on the synthetic distribution, and similar to SwinT, the model is found less sensitive to corruption with a drop of at most 5.13%, 10.36%, and 7.21% on UTKFace, AgeDB, and FairFace datasets, respectively.

4.4. Data Augmentation for Robustness

To enhance the robustness of the models trained on the HDASynChildFace dataset, we applied data augmentation techniques during training. Real-world distortions and training data variability are simulated using the six perturbations listed in Section 3.2 (CS, CC, BW, GNC, GB, and JPEG), hence augmenting the training data variability. These augmentations significantly increase the training set and expose the models to a wider spectrum of facial variances, therefore ensuring that the models are robust against usual real-world problems. Table 6 and Table 7 show the performance improvements observed in models trained using augmented data compared to those trained without augmentation. From both tables, it is evident that models trained with data augmentation show a substantial increase in accuracy compared to those trained without augmentation. For CNN-based models, DenseNet-121 demonstrates a significant improvement when trained with augmented data, achieving 64.16% accuracy on the UTKFace test set, compared to 48.56% without augmentation. Similarly, ResNet-50 increases from 59.31% to 67.93% on the same test set. Despite these gains, CNN models trained on real-world data still outperform their augmented synthetic counterparts, with DenseNet-121 achieving 83.00% on the UTKFace dataset. For transformer-based models, the impact of augmentation is pronounced. BEiT, the best-performing model, improves from 88.53% to 90.69% on UTKFace with augmentation, demonstrating the effectiveness of these perturbations in boosting model robustness. However, while BEiT trained on real data achieves 94.45%, the performance of the model trained on augmented synthetic data, reaching 90.69%, is nearly comparable to the results obtained from real data training. This suggests that the application of augmentations significantly narrows the performance gap between synthetic and real-world datasets.

4.5. Model Interpretability

To gain deeper insights into the decision-making processes and the internal feature representations of our gender classification models, we use two complementary visualization techniques: Gradient-weighted Class Activation Mapping (Grad-CAM) [34] and t-distributed Stochastic Neighbor Embedding (t-SNE) [40]. These methods help to illustrate how the models focus on key facial features and how they separate different classes under various types of perturbations.

4.5.1 Grad-CAM Visualizations

Grad-CAM allows us to visually examine which regions of the face the models attend to when making predictions. In Figure 4, we present Grad-CAM visualizations for the best-performing model, BEiT, which is trained on real-world

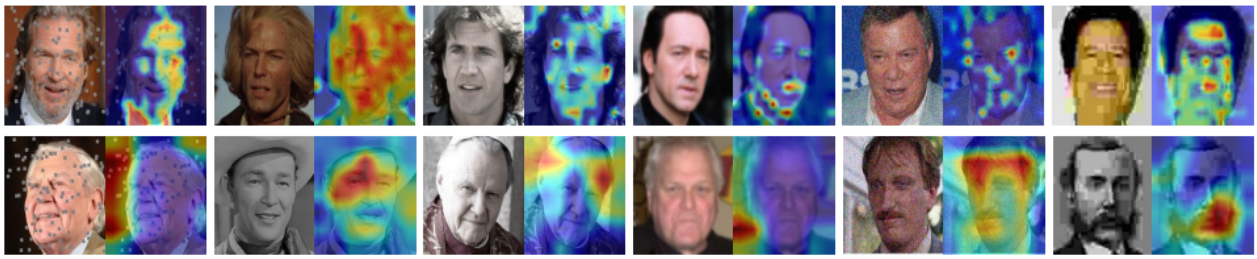

Figure 4. Grad-CAM visualizations for BEiT and ResNet-50 across different perturbations. The first row shows BEiT, and the second row shows ResNet-50, from left: Blockwise Distortion (BW), Color Contrast (CC), Color Saturation (CS), Gaussian Blur (GB), Gaussian Noise in Color component (GNC), and JPEG Compression.

data alongside the ResNet-50 model, trained on synthetic data under different perturbations. The results reveal that BEiT consistently focuses on critical facial areas, such as the eyes and mouth, even when images are distorted by blockwise occlusions, color saturation, or Gaussian noise. BEiT's ability to maintain its attention on relevant facial attributes, even when the image quality is compromised, demonstrates the model's resilience to such distortions and the effectiveness of its attention mechanisms. In contrast, the Grad-CAM visualizations for ResNet-50, trained on synthetic data, show that while the model successfully highlights important facial features, its attention shifts depending on the type of distortion. This suggests that the synthetic data allows for effective generalization, but the model still has room to improve when it comes to handling severe distortion.

Figure 5. t-SNE plots feature representations for models (first row: BEiT and second row: ResNet-50) trained on real and synthetic datasets across different perturbations, from left: Blockwise Distortion (BW), Color Contrast (CC), and Color Saturation (CS).

4.5.2 t-SNE Feature Representation Analysis

In Figure 5, we present t-SNE plots for the BEiT and ResNet-50 models. The results demonstrate that BEiT maintains well-separated clusters for different classes, even when subjected to significant perturbations such as blockwise occlusion, color contrast changes, and Gaussian noise. This highlights BEiT's ability to preserve the integrity of its feature representations, reflecting its robustness and the advantages of its advanced architecture. In contrast, the ResNet-50 model, trained on synthetic data, shows less distinct clustering. While it can differentiate between classes, the clusters are less defined, indicating that synthetic data, while useful, may not capture the full range of real-world variability. This comparison underscores the superiority of Vision Transformers like BEiT in achieving better-defined feature representations and greater resilience when confronted with diverse perturbations.

5. Conclusion

The robustness and efficiency of gender classification models trained using synthetic and real-world data are systematically evaluated in this paper. Based on accuracy and robustness to several perturbations, the results show a clear advantage for models trained using real-world data that frequently outperform those trained using synthetic data. However, the point to note here is that using real datasets brings a privacy concern. ViT models performed better than CNN models in managing clean and distorted data scenarios even trained on synthetic face images, solving privacy concerns to some extent. We further observed that augmenting different perturbations in the training set results in improved performance of the models, especially the ViT-based model, and is almost close to the real data training performance. Also, the explainability studies, including Grad-CAM and t-SNE visualizations, demonstrate the superiority of ViTs in discriminating the feature space of males and females and utilizing critical face regions for classification.

References

[1] Olatunbosun Agbo-Ajala and Serestina Viriri. Age estimation of real-time faces using convolutional neural network. In *International Conference on Computational Collective Intelligence*, pages 316–327. Springer, 2019. 2

[2] Yaman Akbulut, Abdulkadir Şengür, and Sami Ekici. Gender recognition from face images with deep learning. In *2017 International artificial intelligence and data processing symposium (IDAP)*, pages 1–4. IEEE, 2017. 2

[3] Abhinav Anand, Ruggero Donida Labati, Angelo Genovese, Enrique Munoz, Vincenzo Piuri, and Fabio Scotti. Age estimation based on face images and pre-trained convolutional neural networks. In *2017 IEEE symposium series on computational intelligence (SSCI)*, pages 1–7. IEEE, 2017. 2

[4] Grigory Antipov, Moez Baccouche, Sid-Ahmed Berrani, and Jean-Luc Dugelay. Effective training of convolutional neural networks for face-based gender and age prediction. *Pattern Recognition*, 72:15–26, 2017. 2

[5] Andrea Atzori, Pietro Cosseddu, Gianni Fenu, and Mirko Marras. The impact of balancing real and synthetic data on accuracy and fairness in face recognition. *arXiv preprint arXiv:2409.02867*, 2024. 2

[6] Safaa Azzakhnini, Lahoucine Ballihi, and Driss Aboutajdine. Combining facial parts for learning gender, ethnicity, and emotional state based on rgb-d information. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1s):1–14, 2018. 1

[7] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*. 4

[8] Fadi Boutros, Vitomir Struc, Julian Fierrez, and Naser Damer. Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing*, 135:104688, 2023. 1

[9] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[10] Erhan Davarci and Emin Anarim. Gender detection based on gait data: A deep learning approach with synthetic data generation and continuous wavelet transform. *IEEE Access*, 2023. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[12] Mingxing Duan, Kenli Li, and Keqin Li. An ensemble cnn2elm for age estimation. *IEEE Transactions on Information Forensics and Security*, 13(3):758–772, 2017. 2

[13] Mingxing Duan, Kenli Li, Canqun Yang, and Keqin Li. A hybrid deep learning cnn–elm for age and gender classification. *Neurocomputing*, 275:448–461, 2018. 2

[14] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security*, 9(12):2170–2179, 2014. 2

[15] Alberto N Escalante B and Laurenz Wiskott. Gender and age estimation from synthetic face images. In *Computational Intelligence for Knowledge-Based Systems Design: 13th International Conference on Information Processing and Management of Uncertainty, IPMU 2010, Dortmund, Germany, June 28-July 2, 2010. Proceedings 13*, pages 240–249. Springer, 2010. 2

[16] Sergio Escalera, Jordi Gonzalez, Xavier Baró, Pablo Pardo, Junior Fabian, Marc Oliu, Hugo Jair Escalante, Ivan Huerta, and Isabelle Guyon. Chalearn looking at people 2015 new competitions: Age estimation and cultural event recognition. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2015. 2

[17] Magnus Falkenberg, Anders Bensen Ottsen, Mathias Ibsen, and Christian Rathgeb. Child face recognition at scale: Synthetic data generation and performance benchmark. *Frontiers in Signal Processing*, 4:1308505, 2024. 2, 3

[18] Guodong Guo, Yun Fu, Charles R Dyer, and Thomas S Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, 17(7):1178–1188, 2008. 2

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 4

[20] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019. 2, 3, 4

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[22] Brett Koonce and Brett Koonce. Resnet 50. *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 63–72, 2021. 4

[23] RAJIV KUMAR and AM ARTHANARIEE. Recognition of missing children using feature-based method. *Journal of Theoretical & Applied Information Technology*, 96(7), 2018. 2

[24] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–42, 2015. 2

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[26] Shubham Mittal and Shiva Mittal. Gender recognition from facial images using convolutional neural network. In *2019 Fifth International Conference on Image Information Processing (ICIIP)*, pages 347–352. IEEE, 2019. 2

[27] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017. 2, 3, 4

[28] Sergey I Nikolenko. *Synthetic data for deep learning*, volume 174. Springer, 2021. 1

[29] Mohamed Oulad-Kaddour, Hamid Haddadou, Cristina Conde Vilda, Daniel Palacios-Alonso, Karima Benatchba, and Enrique Cabello. Deep learning-based gender classification by training with fake data. *IEEE Access*, 11:120766–120779, 2023. 2

[30] Anibal Pedraza, Oscar Deniz, and Gloria Bueno. Really natural adversarial examples. *International Journal of Machine Learning and Cybernetics*, 13(4):1065–1077, 2022. 1

[31] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. Synface: Face recognition with synthetic data. In *IEEE/CVF International Conference on Computer Vision*, pages 10880–10890, 2021. 1

[32] Sanjay Roka and Manoj Diwakar. Cvit: a convolution vision transformer for video abnormal behavior detection and localization. *SN Computer Science*, 4(6):829, 2023. 4

[33] Denys Rozumnyi, Martin R Oswald, Vittorio Ferrari, and Marc Pollefeys. Motion-from-blur: 3d shape and motion estimation of motion-blurred objects in videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15990–15999, 2022. 1

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359, 2020. 7

[35] Subhani Shaik and Anto A Micheal. Automatic age and gender recognition in human face image dataset using convolutional neural network system. *International journal of advance research in computer science and management studies*, 4(2), 2016. 2

[36] Aakash Singh and Vivek Kumar Singh. A hybrid transformer–sequencer approach for age and gender classification from in-wild facial images. *Neural Computing and Applications*, 36(3):1149–1165, 2024. 2

[37] Vasu Krishna Suravarapu and Hemprasad Yashwant Patil. Person identification and gender classification based on vision transformers for periocular images. *Applied Sciences*, 13(5):3116, 2023. 2

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4

[39] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5

[40] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 1(10):e2, 2016. 7

[41] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 1

[42] Ke Zhang, Ce Gao, Liru Guo, Miao Sun, Xingfang Yuan, Tony X Han, Zhenbing Zhao, and Baogang Li. Age group and gender estimation in the wild with deep ror architecture. *IEEE Access*, 5:22492–22503, 2017. 2

[43] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017. 2, 3, 4