UDC 00 (004.85)

ИСПОЛЬЗОВАНИЕ СИНТЕТИЧЕСКИХ ДАННЫХ ДЛЯ РАСПОЗНАВАНИЯ АТРИБУТОВ ЛИЦА

Р. М. Черепенников

Белорусский государственный университет, Беларусь, Минск,
charapennikaurm@gmail.com

В статье исследуется использование диффузионных моделей для создания синтетических наборов данных для распознавания факторных признаков с акцентом на прогнозирование возраста, пола и этнической принадлежности. Мы сравниваем модели, обученные на реальных данных, синтетических данных и их комбинации. Мы демонстрируем, что предварительное обучение на синтетических данных с последующей точной настройкой на реальных образцах превосходит модели, обученные исключительно на реальных данных. Наши результаты подчеркивают потенциал синтетических данных для повышения производительности нейронных сетей в задачах регрессии и классификации.

Ключевые слова: нейронные сети; компьютерное зрение; синтетические данные; диффузионные модели.

USING SYNTHETIC DATA FOR FACE ATTRIBUTES RECOGNITION

R. M. Charapennikau

Belarusian State University, Belarus, Minsk, charapennikaurm@gmail.com

The paper explores the usage of diffusion models to create synthetic datasets for facial attribute recognition, focusing on age, gender and ethnicity prediction. We compare models trained on real-world data, synthetic data, and a combination of both. We demonstrate that pretraining on synthetic data followed by fine-tuning on real samples outperforms models trained solely on real-world data. Our results highlight the potential of synthetic data to enhance neural network performance in regression and classification tasks.

Keywords: neural networks; computer vision; synthetic data; diffusion models.

Introduction

In recent years, a wide variety of datasets have been developed for computer vision tasks, supporting advances in object detection, image classification, and many other tasks. While many of these datasets are accessible for educational and scientific research, their use in commercial applications is often restricted. Moreover, publicly available datasets frequently suffer from

limitations, including insufficient size, class imbalance, and inconsistent labeling quality, which can hamper model performance and generalization.

At the same time, diffusion models and generative adversarial networks (GANs) have achieved remarkable success in generating high-quality synthetic images [1-4] and videos [5;6] that closely resemble real-world data. These generative models open the possibility of creating balanced, customizable synthetic datasets that can potentially complement or substitute real-world data in training deep learning models for various tasks.

In this work, we explore the feasibility of using synthetic data generated by text-to-image diffusion models to train neural networks for face attribute recognition, specifically targeting age, gender, and ethnicity classification. We compare the performance of models trained solely on synthetic data, on real-world data, and models trained using combinations of both. Our goal is to determine whether synthetic data can mitigate common dataset issues such as imbalance and limited diversity while improving model performance.

The implementation is publicly available at https://github.com/charapenni-kaurm/synth-far.

Dataset

Fig. 1. Selected samples from generated dataset

To construct the dataset, we employ the Humans fine-tuned version of Stable Diffusion 1.5 [7; 1]. To ensure that the generated facial images are properly framed, we employ the OpenPose ControlNet model [8], which aligns poses using 24 predefined pose templates.

The dataset generation process follows a structured prompt-based approach. The base prompt template is formulated as follows: *close-up photorealistic, raw photo, amateur photo, face, beautiful (age y.o)*

($gender$) (from $country$) named $name$, $hair_length$ $hair_style$ $hair_color$ hair, $suffix$.

All placeholders marked with $ are substituted with randomly selected attributes corresponding to target labels for classification and regression tasks.

To mitigate artifacts and ensure high image quality, we apply a negative prompt explicitly discouraging undesirable characteristics, such as anatomical distortions, artificial rendering effects, and other visual artifacts.

We use following negative prompt, provided by Humans model author: *(hands), (3d, render, cgi, doll, painting, fake, 3d modeling:1.4),(worst quality, low quality:1.4),monochrome, child, deformed, malformed, deformed face, bad teeth, bad hands, bad fingers, bad eyes, long body, blurry, duplicated, cloned, duplicate body parts, disfigured, extra limbs, fused fingers, extra fingers, twisted, distorted, malformed hands, mutated hands and fingers, conjoined, missing limbs, bad anatomy, bad proportions, logo, watermark, text, copyright, signature, lowres, mutated, mutilated, artifacts, gross, ugly, malformed genital.*

For image sampling, we utilize the Diffusers library [9] with the following hyperparameters, all of which were selected empirically based to ensure image quality and diversity: Guidance scale – 7.5; Number of diffusion steps – 15; ControlNet conditioning scale – 0.8; Scheduler – DPMSolverMultistepScheduler.

On Figure 1 we present selected samples from the generated dataset.

We compare demographic attribute distributions (age, gender, and ethnicity) in our dataset with two existing datasets: UTKFace [10] and FairFace [11]. Figure 2 illustrates these distributions, showing that our dataset provides a more balanced representation of demographic attributes compared to traditional datasets. This balance is crucial for mitigating biases in downstream facial attribute recognition tasks.

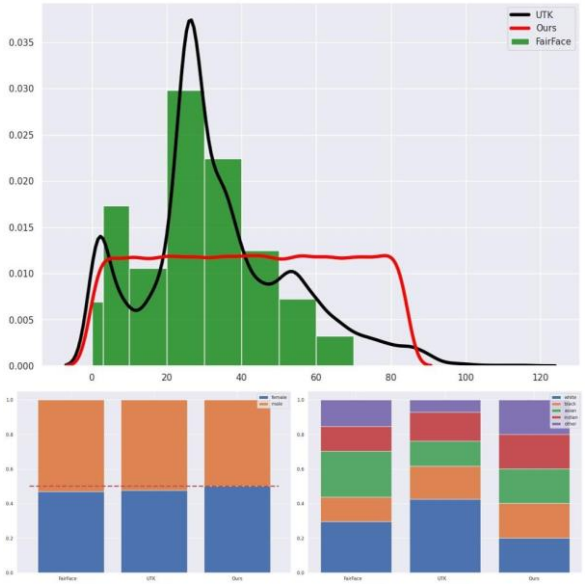

Fig.2. Distributions of Age, Gender and Ethnicity attributes across different datasets

Neural Network Architecture

Our network architecture follows a multi-task learning paradigm, leveraging a shared image encoder to extract visual features, which are subsequently processed by three independent classification or regression (depending on task) heads to predict age, gender, and ethnicity.

The model consists of the following key components:

- Image Encoder: A convolutional or visual transformer backbone (in particular, we use ResNet-50 [12]), is employed as a feature extractor. This encoder processes input facial images and generates a high-dimensional feature representation.
- Task-Specific Prediction Heads: The extracted features are fed into three separate branches, each responsible for predicting a distinct facial attribute. The Age Head is designed to estimate the subject's age, which can be formulated as either a regression (for UTKFace and Ours datasets) or classification (for
FairFace dataset) problem. The Gender Head performs binary classification to distinguish between male and female subjects. The Ethnicity Head predicts the subject's ethnicity as a multi-class classification task. Each prediction head is a fully connected layer that transform the shared feature representations into task-specific outputs.

The model is trained using a weighted combination of loss functions corresponding to each prediction task, facilitating joint optimization across multiple facial attributes.

Experimental Results

To ensure a fair comparison between different training strategies, we employ a consistent training protocol across all models. Specifically:

- Optimizer: AdamW
- Computational Budget: Maximum of 100,000 optimization steps for all models
- Data Augmentations: Identical augmentations applied across all models
- For models fine-tuned on rea-world data, we use 50,000 pretraining steps on synthetic data and up to 50,000 fine-tuning steps on real-world data

Further details on training parameters, including hyperparameter choices and augmentation strategies, can be found in the provided code repository.

Our experimental results are presented in the Table, name of model represents the dataset it was trained on, For SynthFAR-<X><dataset>: <X>

denotes percentage of <dataset> used. The best result is highlighted in **bold**, while the second-best result is highlighted in *italic*.

Experimental Results

	UTK-Face			FairFace		
Model	Age MAE	Ethnicity Accuracy	Gender Accuracy	Age Accuracy	Ethnicity Accuracy	Gender Accuracy
(a) SynthFAR	10.27	0.58	0.76	0.33	0.47	0.77
(b) UTK	**5.33**	**0.8**	*0.9*	0.28	0.29	0.71
(c) FairFace	-			0.71		
(d) SynthFAR-25FairFace	-	0.63	*0.9*	*0.57*	**0.79**	**0.95**
(e) SynthFAR-50FairFace	-	0.7	0.86	0.51	0.71	0.91
(f) SynthFAR-50UTK	*5.9*	*0.77*	0.89	0.56	*0.78*	*0.94*
(g) SynthFAR-75FairFace	-	0.71	0.89	0.37	0.5	0.8

1) The model trained exclusively on synthetic data (a) demonstrates significantly inferior performance across all metrics. In particular, it exhibits a high age MAE (10.27) and lower accuracy in both ethnicity and gender classification compared to models trained solely on real data. These results indicate that, while synthetic data provides a useful pre-training signal, they are insufficient for achieving competitive performance without additional adaptation to real-world samples.

2) Models pre-trained on synthetic data and subsequently fine-tuned on varying proportions of real-world data (d, e, f, g) exhibit substantial performance improvements over the synthetic-only model. Even with as little as 25% of real data for (d) and 50% for (e, f), the model surpasses the synthetic-only approach across all metrics and begins to approach the performance of real word data only baselines (b, c).

3) Increasing the proportion of real-world data for fine-tuning to 75% of FairFace for (g) results in further performance gains, ultimately leading to model that match or exceed the accuracy of real-world-data-only models.

Notably, (g) outperforms (c), despite being trained on only 75% of the available real-world dataset.

References

1. *Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B.* High-Resolution Image Synthesis with Latent Diffusion Models // arXiv preprint, 2021. URL: https://arxiv.org/abs/2112.10752 (date of access: 10.04.2025).

2. *Podell D., English Z., Lacey K., Blattmann A., Dockhorn T., Müller J., Penna J., Rombach R.* SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis // arXiv preprint, 2023. URL: https://arxiv.org/abs/2307.01952 (date of access: 10.04.2025).

3. *Karras T., Laine S., Aittala M., Hellsten J., Lehtinen J., Aila T.* Analyzing and Improving the Image Quality of StyleGAN // arXiv preprint, 2019. URL: https://arxiv.org/abs/1912.04958 (date of access: 10.04.2025).

4. *Karras T., Aittala M., Laine S., Härkönen E., Hellsten J., Lehtinen J., Aila T.* Alias-Free Generative Adversarial Networks // arXiv preprint, 2021. URL: https://arxiv.org/abs/2106.12423 (date of access: 10.04.2025).

5. OpenAI. Video generation models as world simulators // OpenAI Research, 2024. URL: https://openai.com/index/video-generation-models-as-world-simulators/ (date of access: 10.04.2025).

6. The Movie Gen team @ Meta. Movie Gen: A Cast of Media Foundation Models // Meta AI Research, 2024. URL: https://ai.meta.com/static-resource/movie-gen-research-paper (date of access: 10.04.2025).

7. Humans (Reloaded) // CivitAI, 2024. URL: https://civitai.com/models/542430/humans-reloaded (date of access: 10.04.2025).

8. *Zhang L., Rao A., Agrawala M.* Adding Conditional Control to Text-to-Image Diffusion Models // arXiv preprint, 2023. URL: https://arxiv.org/abs/2302.05543 (date of access: 10.04.2025).

9. *Von Platen P., Patil S., Lozhkov A., Cuenca P., Lambert N., Rasul K., Davaadorj M., Nair D., Paul S., Berman W., Xu Y., Liu S., Wolf T.* Diffusers: State-of-the-art diffusion models // GitHub, 2022. URL: https://github.com/huggingface/diffusers (date of access: 10.04.2025).

10. *Zhang Z., Song Y., Qi H.* Age progression/regression by conditional adversarial autoencoder // arXiv preprint, 2017. URL: https://arxiv.org/abs/1702.08423 (date of access: 10.04.2025).

11. *Karkkainen K., Joo J.* FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age // arXiv preprint, 2019. URL: https://arxiv.org/abs/1908.04913 (date of access: 10.04.2025).

12. *He K., Zhang X., Ren S., Sun J.* Deep Residual Learning for Image Recognition // arXiv preprint, 2015. URL: https://arxiv.org/abs/1512.03385 (date of access: 10.04.2025).