

# Enhancing AI Face Realism: Cost-Efficient Quality Improvement in Distilled Diffusion Models with a Fully Synthetic Dataset

Jakub Wąsala<sup>[0009-0003-4808-6608]</sup>, Bartłomiej Wrzalski<sup>[0009-0007-0130-3379]</sup>,  
Kornelia Noculak<sup>[0009-0005-8616-4528]</sup>, Yuliia Tarasenko<sup>[0009-0004-3027-6742]</sup>,  
Oliwer Krupa<sup>[0009-0009-7546-6140]</sup>, Jan Kocon<sup>[0000-0002-7665-6896]</sup>, and  
Grzegorz Chodak<sup>[0000-0002-9604-482X]</sup>

Department of Artificial Intelligence,  
Wroclaw University of Science and Technology,  
Wroclaw, Poland  
{254554, 254566, 254580, 283338, 254027}@student.pwr.edu.pl,  
{jan.kocon, grzegorz.chodak}@pwr.edu.pl

**Abstract.** In recent years, diffusion models brought unprecedented high-quality image generation, alongside high inference costs. We explore variants of FLUX, a state-of-the-art family of models: baseline FLUX.1-*dev* and distilled FLUX.1-*schnell*. We hypothesize that differences between outputs of baseline and distilled models are consistent within a specialized domain, such as portrait generation. Then, we suggest training a domain-specific image-to-image (I2I) fast translation model from *schnell* to *dev* domain. The paper discusses two potential backbones for the model: UNet, requiring a pairwise dataset of low- and high-quality images of the same scene and subject, and non-pairwise CycleGAN. We demonstrate that results produced by a distilled *schnell* model with our I2I head are perceptually close to what a baseline *dev* model would produce, while cutting 82% computational cost. We also show that results generated by CycleGAN are superior to UNet, which suggests that training images need not be paired in order to achieve satisfactory results.

**Keywords:** deep learning · generative models · fine-tuning



Fig. 1: **Left sides:** example input images from FLUX.1-*schnell*. **Right sides:** outputs from our model based on ESA-CycleGAN. Our model adds skin details, enhances hair, and improves the reflection of eye pupils, contributing to the overall realistic feel of a portrait.

## 1 Introduction

Deep generative models are designed to learn data distributions so they can create realistic images. But since they rely on a fixed set of parameters, they often fall short when it comes to capturing the full complexity and nuances of real-world subjects. This becomes especially clear in portrait generation, where even the most advanced models can miss fine details or produce inaccuracies [4].

In recent years, models such as StyleGAN [18], BigGAN [5], DALL-E [23], and FLUX.1 [1] have made remarkable progress in generating high-quality synthetic images. These advancements have opened up new possibilities in areas like entertainment, advertising, and professional avatar creation [22]. However, their deployment remains constrained by high computational costs, requiring extensive resources for both training and inference [7]. This poses challenges for practical implementation, especially in resource-limited environments.

To address this, many models have their lightweight distilled versions such as FLUX.1-*schnell* [1] that introduce a trade-off between computational demands and image quality. However, these models often struggle more with visible imperfections — including unnatural lighting, texture inconsistencies, and unrealistic facial details — limiting their effectiveness in professional applications.

Our proposal reduces the gap between resource-efficient distilled models and the quality of their higher-quality versions by introducing a dedicated image-to-image (I2I) model trained to recover lost details. As a motivation example, we showcase photorealistic portrait generation using the recent diffusion model FLUX.1. Images generated by its advanced revision, FLUX.1-*dev*, exhibit finer

details in hair, skin, and eyes compared to those produced by FLUX.1-*schnell*, a timestep-distilled variant. On the other hand, FLUX.1-*dev* is 7 times slower than its distilled version, requiring more denoising steps.

To validate our approach, we propose and test two key hypotheses: **(1)** the imperfections of a distilled model are consistent within a specialized domain (e.g., portrait photography) and thus can be learned by an I2I model, **(2)** the combined pipeline — comprising the distilled model and our I2I head — is still significantly faster than the full-scale model.

A crucial factor in training an effective I2I model is the dataset’s quality. To address this, we introduce a novel approach to fully synthetic dataset generation. Our method operates pairwise, producing image pairs featuring the same subject and composition but using different FLUX.1 revisions. This ensures controlled variations in quality and photorealistic detail, forming a strong training set for learning quality restoration. We also demonstrate that a simple prompt engineering technique can significantly enhance dataset diversity across race, ethnicity, gender, and age.



Fig. 2: Example pair of FLUX.1-*schnell* and FLUX.1-*dev* images from our dataset. We observe that the *dev* variant generates portraits featuring more details of skin, hair, and eyes, leading to a more photorealistic look.

We also compare our work to existing solutions aiming at improving the quality of images generated by diffusion models. Typical ways to influence the outputs of a model include prompt engineering and parameter-efficient fine-tuning (PEFT) methods, such as LoRA [13]. We note that in comparison with those methods, our approach is technically model-agnostic, as one can forward any image through our model, regardless of its source.

Our contributions are as follows. **(1)** We demonstrate that, given two versions of a model — baseline and distilled — and a specialized domain, such as portrait generation, it is possible to train an image-to-image translation model that effectively mitigates some of the quality degradation introduced by the distilled model (Figure 1). **(2)** We show that using a distilled version of a generative model with our I2I head leads to 82% savings in computation time. **(3)** We propose a novel approach to dataset generation involving a tailored prompt engineering technique that results in a fully synthetic, pairwise, and diverse dataset. **(4)** We use our approach to generate a dataset of synthetic portraits of 280,000 images, of which samples are shown in the Figure 2. **(5)** We compare different approaches to I2I training, including supervised paired methods based on UNet and unpaired ones based on CycleGAN [33, 30].

## 2 Related Work

The rapid development of generative models has significantly improved the quality and realism of synthesized images. While diffusion models have set new standards in photorealistic image generation due to their iterative refinement process [11, 7], their high computational cost limits their practical deployment [28, 8]. This has led to an increasing focus on efficient alternatives, including model distillation [10], quantization [16], and hybrid fine-tuning approaches [12]. Despite these optimizations, distilled and quantized models often introduce artifacts, such as unnatural lighting and texture inconsistencies [24, 31]. Addressing these issues is critical for applications requiring high-fidelity image synthesis.

### 2.1 Diffusion-Based Image-to-Image Translation

Diffusion models, such as DDPM [11] and DDIM [28], have revolutionized image synthesis by leveraging iterative denoising processes to generate high-quality images. Compared to GANs, they offer superior realism and diversity but at the cost of high computational demands.

One approach to improving efficiency in I2I tasks is the DiffI2I model, which incorporates a compact prior extraction network and a dynamic transformer to produce accurate translations with reduced computational overhead [31]. Another challenge is content preservation, which researchers have tackled by disentangling style and content representations, ensuring that generated images maintain the original content while adopting the desired style [20]. This is particularly relevant for tasks requiring high fidelity to the source image, such as portrait generation.

Recent research has also explored alternative I2I models, including Latent Consistency Models (LCMs) [21], which accelerate diffusion processes while maintaining quality. Additionally, GAN-based approaches, such as pix2pix [15] and CycleGAN [17], have demonstrated effectiveness in translating images across domains, albeit with challenges in preserving fine details. ControlNet [32] has further enhanced diffusion-based I2I tasks by providing more precise control over image modifications.

### 2.2 FLUX: A State-of-the-Art Text-to-Image Model

FLUX is a state-of-the-art generative model optimized for computational efficiency, making it suitable for real-time applications such as avatar generation or creative content workflows [1].

In our work, we utilize two versions of this model: a baseline version with standard quality settings (FLUX.1-*dev*) and a faster distilled version (FLUX.1-*schnell*) optimized for speed. The latter significantly reduces the number of diffusion steps, accelerating generation time while preserving most of the visual fidelity. This efficiency is achieved through model distillation — a compression technique where a smaller “student” model is trained to mimic the behavior of

a larger “teacher” model [10]. In generative models, distillation enables faster inference with minimal quality loss, which is crucial for practical deployment.

Both of these models are open-weight and available through the Hugging Face platform [2, 3].

Additionally, FLUX supports modern fine-tuning, allowing further optimization for specific use cases without retraining from scratch [13].

### 2.3 Fine-Tuning Techniques

Fine-tuning pre-trained models is a prevalent strategy to adapt generative models to specific tasks or styles. Two primary methods have emerged: full fine-tuning and parameter-efficient fine-tuning techniques like Low-Rank Adaptation (LoRA).

Comparisons between full fine-tuning and LoRA training suggest that full fine-tuning yields superior results, with reduced overfitting and improved generalization [27]. However, this approach demands significant computational resources, which may not be feasible in all scenarios. In contrast, LoRA fine-tuning offers a more resource-efficient alternative by introducing low-rank matrices to capture essential adaptations, updating only a subset of parameters, thereby reducing memory usage and training time [13]. Additional studies have explored hybrid fine-tuning strategies, such as adapter-based tuning [12] and layer-wise modifications [19], to balance efficiency and model expressiveness.

Furthermore, specialized fine-tuning techniques such as Realism LoRA [14] and DreamBooth [25] have been developed to enhance realism in generated images. Realism LoRA refines model weights to produce highly detailed and lifelike outputs, making it valuable for portrait generation and professional photography applications. DreamBooth, on the other hand, enables user-specific fine-tuning by incorporating personalized training data, allowing models to generate images with high identity preservation.

### 2.4 Gaps and Contributions

While existing methods have advanced the fields of diffusion-based I2I translation, fine-tuning, and quantization, challenges remain in achieving high-quality image generation with reduced computational costs. Our study addresses these gaps by introducing a novel approach that leverages a fully synthetic paired dataset derived from baseline and distilled versions of a base model. By training an image-to-image translation head in a supervised manner, our method enhances the output quality of a baseline generator to match that of more computationally intensive models. This strategy eliminates the need for manual annotations or real reference photos, offering a cost-effective solution for photorealistic portrait generation.

In summary, our approach contributes to the existing body of work by providing an efficient framework that combines supervised learning and image-to-image translation to improve the quality of diffusion models without incurring significant computational expenses.

### 3 Dataset

The work presents a novel approach to generating a diverse dataset of portraits. Its main idea was to acquire almost identical pairs of images, semantically coherent and distinguished only by details such as a more natural skin texture and a better-depicted beard or hair. Thanks to this technique, there was no need for an outsourced dataset, as we used it to facilitate the assembly of 280,000 portraits — corresponding to 140,000 image pairs.

#### 3.1 Preparation of the dataset

For the purpose of preparing the dataset, we used FLUX.1-*schnell* and FLUX.1-*dev*. To begin with, we generated 10,000 samples using the prompt “A professional business portrait” with a resolution of  $512 \times 512$  pixels. However, we observed a lack of diversity in the generated images, as the dataset consisted of mostly young, slim, and generally flawless people, with additional bias towards stereotypical white men.



Fig. 3: Samples of images created during our research. (a) presents samples created using base prompt and (b) samples generated with prompt containing full names

In the following phase, to increase the diversity of the collection, we applied prompt engineering by adding to the prompts supplemental names and surnames. For preparing the names, we utilized the IMDb Actors and Movies database. By using a subset of this collection, we generated close to 140,000 images with FLUX.1-*schnell* with the prompt “A professional portrait of [FULL NAME]”, where each name corresponded to the individual actor in the IMDb database. We hypothesize that the model, even if not trained to properly depict

specific individuals based on name, may associate names with diverse cultures. Samples of both approaches are shown in Figure 3

Subsequently, using the FLUX.1-*dev* model, we produced target images in I2I mode with the identical predefined prompts. Beyond higher quality, the resulting samples presented improved realism. Using this approach, a dataset of paired images was acquired, consisting of 280,000 images. The pipeline is presented in Figure 5.

### 3.2 Analysis of Dataset Diversity

As mentioned previously, we have increased the diversity of portraits by using prompt engineering. The analysis considered facial attributes such as gender, age, and ethnicity. To perform it, we used a framework provided by the DeepFace library [26]. Results shown in Table 1 demonstrate that the prompt engineering technique led to increased equality in perceived gender occurrences and increased participation of previously underrepresented racial groups, especially Black and Indian. Random samples are presented using t-SNE in Figure 4.

Table 1: Distribution of attributes — perceived gender, age, and dominant ethnicity — in sets based following the baseline prompt and enriched prompt.

Prompt	Perceived Gender	P. Age	P. Dominant Ethnicity
	M : F	18-30 : 31-50 : 50+	Asian : Black : Indian Latino : Middle Eastern : White
Baseline	0.73 : 0.27	0.61 : 0.39 : 0.0	0.12 : 0.01 : 0.01 : 0.26 : 0.02 : 0.59
Enhanced	0.60 : 0.40	0.40 : 0.58 : 0.02	0.10 : 0.05 : 0.04 : 0.08 : 0.02 : 0.7

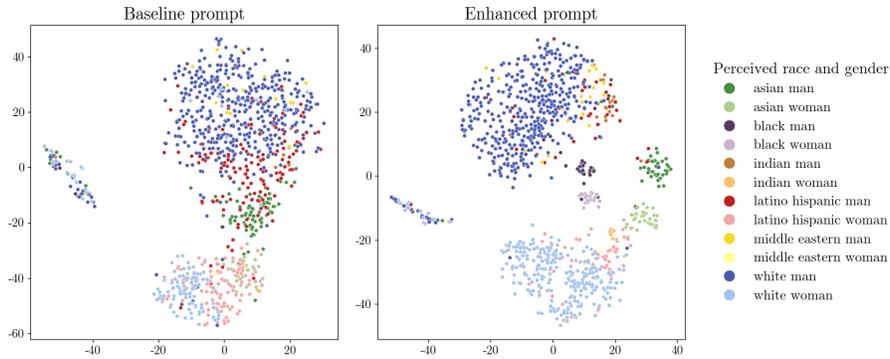


Fig. 4: t-SNE visualization of face embeddings. Representations were acquired using the ArcFace model [6].

## 4 Architecture

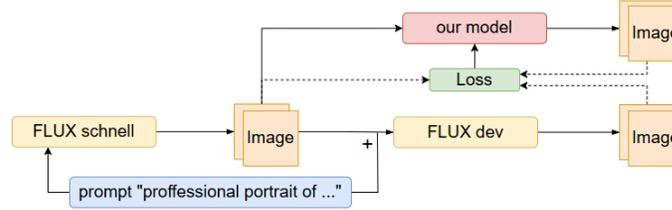


Fig. 5: Pipeline of our solution, including the integration of FLUX.1-*schnell* and our image-to-image model.

### 4.1 Supervised Pairwise Approach

Our initial approach to model design was based on a U-Net architecture in an I2I fashion, taking an image generated by FLUX.1-*schnell* as input and transforming it into an image generated by FLUX.1-*dev*.

The base U-Net architecture was extended with residual connections, which improve gradient flow and training stability, and CBAM (*Convolutional Block Attention Module*) blocks, which allow the model to focus more effectively on important image features. The network had a depth of six layers, enabling it to capture local and global dependencies within the image.

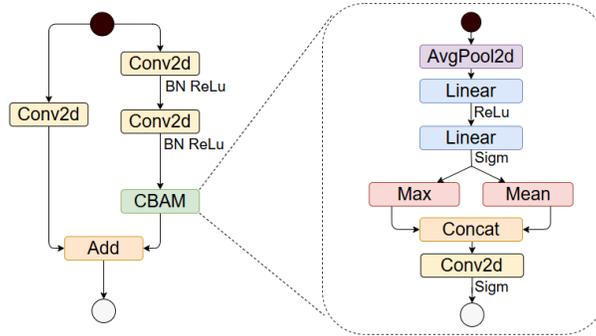


Fig. 6: Single block of our image-to-image U-Net model.

The loss function was formulated as the sum of two components: a gradient loss computed between the input and output images and a perceptual loss – LPIPS (*Learned Perceptual Image Patch Similarity*) measured between the target and generated images. Formally, the function was defined as:

$$\mathcal{L} = GRAD(x, z) + LPIPS(y, z) \quad (1)$$

where  $x$  represents the input image,  $y$  the target image, and  $z$  the image generated by our model. The learning rate was set at  $5 \times 10^{-5}$ , ensuring stable and effective training.

Figure 5 illustrates the general pipeline of our method, including the integration of FLUX.1-*schnell* and our image-to-image model. The architecture of our image-to-image model is shown in Figure 6. It highlights the detailed design of a U-Net block, residual connections, and the CBAM attention mechanism.

## 4.2 Unsupervised Non-Pairwise Approach

In addition to the supervised U-Net-based techniques outlined previously, we also explore *unsupervised non-pairwise* methods for image-to-image translation. Unlike supervised approaches, which require paired training samples, these techniques rely on two separate collections of images — one set per domain — without needing explicit one-to-one correspondences. In our case, we define domain  $A$  as portraits generated by FLUX.1-*schnell*, while domain  $B$  is comprised of images produced by FLUX.1-*dev*.

A leading example of such non-pairwise translation is **CycleGAN** [33], whose core design features two generators: **(1)**  $G$  that converts images from domain  $A$  to domain  $B$ , and **(2)**  $F$  that handles the reverse mapping from  $B$  back to  $A$ . Each generator is trained in tandem with a discriminator: **(1)**  $D_A$  that learns to differentiate real  $A$ -domain images from generated ones, **(2)**  $D_B$  that discerns real  $B$ -domain images from synthesized outputs.

A distinguishing element of CycleGAN is the *cycle consistency loss*, ensuring that an image transformed to the opposite domain and then back again remains close to the original. Formally:

$$\mathcal{L}_{\text{cycle}}(G, F) = \mathbb{E}_{x \sim A} [\|F(G(x)) - x\|] + \mathbb{E}_{y \sim B} [\|G(F(y)) - y\|]. \quad (2)$$

This term prevents the networks from altering content arbitrarily to achieve a plausible look in the target domain.

CycleGAN combines the above term with adversarial objectives for both generators:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GAN}}(G, D_B) + \mathcal{L}_{\text{GAN}}(F, D_A) + \lambda_{\text{cycle}} \cdot \mathcal{L}_{\text{cycle}}, \quad (3)$$

where  $\lambda_{\text{cycle}}$  is a key hyperparameter balancing fidelity to the original image against realism in the transformed domain.

To further refine the translation quality, we consider **ESA-CycleGAN** [30], which integrates an Enhanced Spatial Attention (ESA) module into each generator. This module amplifies important regions of the image through: **(1)** channel reduction and spatial pooling, **(2)** convolutional analysis of the downsampled features, **(3)** bilinear upsampling to restore original dimensions, **(4)** element-wise multiplication to spotlight key structures.

These attention-based enhancements help preserve fine details, particularly in complex or texture-rich images. Empirical evaluations suggest that ESA-CycleGAN improves structural similarity (SSIM) scores and mitigates common artifacts associated with plain CycleGAN. The following sections provide a detailed comparison between the two models, highlighting where the attention-driven approach yields the greatest gains.

## 5 Experiments

### 5.1 Supervised Pairwise Training

We trained a U-Net-like architecture described in 4.1. The training was performed on a fraction (1%) of the dataset with  $512 \times 512$  input size. This approach allowed us to quickly run multiple tests with different losses or architectural modifications to compare their achieved improvements.

All training was performed on a single NVIDIA A100 GPU with early stopping to 10 epochs on validation loss. Single training duration was between 20 and 60 minutes, depending on the parameters.

We sampled probes and performed human evaluations to evaluate the model’s effectiveness and measure improvement. We found out that using standard metrics for image quality, like Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR), does not return reliable values and returns the same values for blurry and sharp images.

To avoid blurry images, we used a combined loss function that measures the structural difference with the input image and the perceptual difference with the target image. Comparing structural difference was crucial, since without it, the model was generating blurry images, “smoothing” it when trying to recreate a better image. To address this, we employed Gradient Loss, although we observed that L1 loss produced similar results in practice.

Our experiments with U-net architecture did not produce acceptable results. Even though quality and photo-realism increased, there were visible artifacts produced by the model. We observed net-like patterns in hairlines and beards, particularly noticeable on light or grey hair. These artifacts negatively impacted the overall realism of the generated images. Example outputs from this pairwise-trained model are shown in the right sides, bottom row of Figure 7.

### 5.2 Unsupervised Non-pairwise Training

In this experiment, we trained and tested CycleGAN-based approaches introduced in 4.2. This included a classic CycleGAN and its modification, ESA-CycleGAN.

Initial trainings were run on a small fraction (0.5% to 2%) of the dataset with  $256 \times 256$  input image size. This allowed us to quickly examine different values of hyperparameters, such as lambda cycle, learning rate, or batch size. The second training phase involved a smaller set of hyperparameters combinations, a 20% subset of the entire dataset, and  $512 \times 512$  input size. We did not rerun training on 100% of the data since we were already satisfied with the outputs from this phase.

All training experiments were conducted on an HPC cluster with 16 NVIDIA A100 GPUs distributed evenly across four nodes. A final run in this setup took 12 hours to complete 190 epochs.

We compare each training’s best CycleGAN loss to select the best setup. We also note SSIM between the input image (SSIM’s reference) and CycleGAN’s full cycle output (SSIM’s target).

Table 2: CycleGAN results.

$\lambda_{cyc}$ \ LR	$10^{-4}$		$2 \times 10^{-4}$		$3 \times 10^{-4}$	
	$\mathcal{L}$	SSIM	$\mathcal{L}$	SSIM	$\mathcal{L}$	SSIM
10	0.96	<b>0.96</b>	0.98	0.95	1.59	0.68
5	0.85	0.95	1.02	0.95	1.44	0.75
2	<b>0.72</b>	0.95	0.90	<b>0.96</b>	1.36	0.80

Table 3: ESA-CycleGAN results.

$\lambda_{cyc}$ \ LR	$10^{-4}$	
	$\mathcal{L}$	SSIM
10	1.22	0.92
5	0.77	<b>0.96</b>
2	<b>0.61</b>	0.95

Tables 2 and 3 present the results from the final training runs. The lowest GAN loss was observed for the ESA-CycleGAN variant with  $\lambda_{cyc} = 2$ . SSIM proved useful for preliminary sanity checks, as values significantly below 0.9 indicated poor-quality results. However, it rapidly converged to approximately 0.95 in higher-quality runs, limiting its effectiveness for more fine-grained performance comparisons.



Fig. 7: **Left sides:** example input images from FLUX.1-*schnell*. **Right sides, top row:** outputs from our non-pairwise model based on ESA-CycleGAN. **Right sides, bottom row:** outputs from our pairwise model based on U-Net.

Representative outputs of the non-pairwise trained model using ESA-CycleGAN are shown in the right sides, top row of Figure 7, demonstrating its effectiveness in enhancing photo-realism. We also include an example of a suboptimal result (top right), where visible artifacts can still occur — such as a ghost edge through the right eye and strong artifacts on the gold chain on the chest.

### 5.3 Computational Complexity Assessment

In this experiment, we evaluate the hypothesis that integrating our model preserves the efficiency benefits of a computationally cheaper distilled backbone. To

this end, we run base FLUX.1-*dev* and FLUX.1-*schnell*. Then, FLUX.1-*schnell* is run with an additional I2I head — in this case, its ESA-CycleGAN variant — with the same configuration. We repeat the procedure for multiple image sizes.

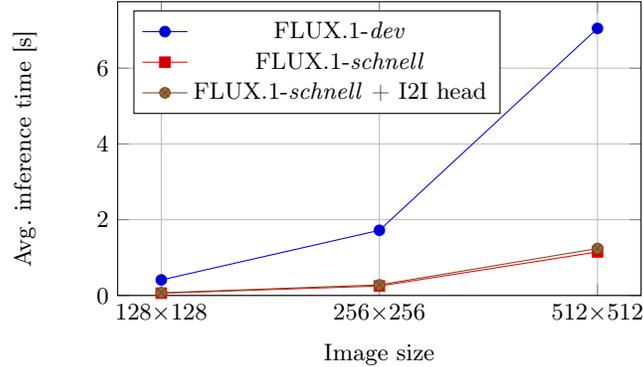


Fig. 8: Comparison of average inference time [s] for different pipelines.

Figure 8 highlights the efficiency of our I2I head, demonstrating a significantly lower computational overhead compared to diffusion-based backbones, FLUX.1-*dev* and FLUX.1-*schnell*. At a resolution of 512×512, our I2I head achieves a processing speed of 11.2 images per second, whereas FLUX.1-*schnell* generates only 0.25 images per second. This results in our proposed pipeline achieving an average speedup of 82% over the standalone FLUX.1-*dev*.

#### 5.4 Quality Assessment

To assess how our enhancement model brings the images closer to what the reference model would output, we calculate Fréchet inception distance (FID) [9] twice — once with FLUX.1-*schnell* as reference, and once with FLUX.1-*dev* as reference. We aim at maximizing the distance from the source dataset to *schnell* images while minimizing the distance to reference *dev* images. To this end, we introduce a  $FID_{diff}$  metric, which is simply a subtraction of  $FID_{dev}$  from  $FID_{schnell}$ . We compare different variations of enhanced FLUX.1-*schnell*: **(1)** pairwise trained I2I head (ours), **(2)** non-pairwise trained I2I head (ours), **(3)** *flux-lora-realism* [14] adaptation available at the Huggingface platform. Results are presented in the Table 4.

We also employed an additional non-reference metric, namely CLIP-IQA [29]. It works by evaluating the overall quality of the images without comparing them directly to a reference dataset.

Our non-pairwise method produces images perceptually closer to those generated by base FLUX.1-*dev* than to those from base FLUX.1-*schnell*. In contrast, the pairwise-trained model yields outputs that exhibit similar perceptual distances to both reference sets. Notably, the LoRA-based approach results in a

Table 4: Quality comparison of different approaches to I2I enhancement.

Model	$\mathbf{FID}_{schnell}$	$\mathbf{FID}_{dev}$	$\mathbf{FID}_{diff}$	$\mathbf{CLIP-IQA}$
FLUX.1- <i>schnell</i>	-	0.37	-	0.35
FLUX.1- <i>dev</i>	0.37	-	-	0.36
LoRA Realism	0.32	0.59	-0.27	0.34
Ours (pairwise)	0.54	0.53	0.01	0.34
Ours (non-pairwise)	0.75	<b>0.34</b>	<b>0.41</b>	0.35

negative  $\mathbf{FID}_{diff}$ , indicating that its outputs are perceptually closer to FLUX.1-*schnell* without adaptation than to FLUX.1-*dev*. Our secondary metric, CLIP-IQA, proved ineffective for quality assessment, as it produced similar scores regardless of the target. More critically, it fails to reflect the superior quality of FLUX.1-*dev*, with score differences falling within a reasonable error margin.

Importantly, the trends captured by  $\mathbf{FID}_{diff}$  align well with our perceptual evaluation, reinforcing its reliability as a quality measure.

## 6 Discussion

The study establishes a cost-efficient method to improve portrait photos, testing paired and unpaired approaches. We show that a lightweight I2I enhancement head can be used to bring the outputs of a smaller, distilled diffusion model closer to the general look and feel of a larger, full-scale base model — while not diminishing the computational save. Unlike existing methods, like Realism LoRA, our model works completely agnostic to the generation method and can be applied at any time later.

The results achieved with CycleGAN and ESA-CycleGAN backbone demonstrate that pairing photos was not necessary to improve realism — indeed, non-pairwise training delivered even better results. The samples obtained with this approach contained fewer artifacts and featured enhanced details with noticeable detail enhancement.

For the U-Net approach, the presence of net-like pattern artifacts was particularly challenging. The effectiveness of this technique was further hindered by imperfections in the dataset. The images produced by the I2I model would contain slightly different features, such as modified backgrounds, jewelry, or clothing, among others. CycleGAN approach also produces some artifacts, such as ghost edges, although they occur less frequently and they are only visible upon close inspection and appear at a lower rate.

While our best model works without pairwise training, we recognize the potential value of our novel approach to dataset construction. The presented pipeline and prompt engineering technique allow controllable distribution of perceived race, gender, and age of subjects, paving the path towards diverse, unbiased synthetic datasets. In future work, the study can be continued by extending the exploration of the methodology as well as expanding the dataset by utilizing other generative models as well (e.g., Stable Diffusion 3.5 Large).

**Acknowledgments.** Work financed by: (1) CLARIN ERIC funded by the Polish Minister of Science, no. 2024/WK/01; (2) The European Regional Development Fund, FENG programme, no. FENG.02.04-IP.040004/24; (3) Department of Artificial Intelligence, Wrocław Tech; (4) Polish HPC infrastructure PLGrid (HPC Center: ACK Cyfronet AGH): computer facilities and support within grant no. PLG/2024/017768.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Black Forest Labs: Announcing Black Forest Labs (2024), <https://blackforestlabs.ai/announcing-black-forest-labs/>
2. Black Forest Labs: FLUX.1-dev (2024), <https://huggingface.co/black-forest-labs/FLUX.1-dev>
3. Black Forest Labs: FLUX.1-schnell (2024), <https://huggingface.co/black-forest-labs/FLUX.1-schnell>
4. Borji, A.: Qualitative failures of image generation models and their application in detecting deepfakes. *Image and Vision Computing* **137**, 104771 (2023)
5. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: *International Conference on Learning Representations* (2019)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4690–4699 (2019)
7. Dhariwal, P., Nichol, A.Q.: Diffusion models beat GANs on image synthesis. In: *Advances in Neural Information Processing Systems* (2021)
8. Guo, X., Yang, Y., Ye, C., Lu, S., Xiang, Y., Ma, T.: Accelerating diffusion models via pre-segmentation diffusion sampling for medical image segmentation (2022)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. *Advances in Neural Information Processing Systems* **30** (2017)
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *NIPS* (2015)
11. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems* (2020)
12. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: *International conference on machine learning*. pp. 2790–2799. PMLR (2019)
13. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. In: *International Conference on Learning Representations* (2022)
14. hugovntnr (HuggingFace user): Flux-Schnell-Realism (2024), <https://huggingface.co/hugovntnr/flux-schnell-realism>, accessed: 2025-02-27
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. *CVPR* (2018)
16. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D.: Quantization and training of neural networks for efficient integer-arithmetic-only inference (2017)

17. Jun-Yan, Z., Taesung, P., Phillip, I., Alexei A., E.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. ICCV (2020)
18. Karras, T., Laine, S., Aila, T.: A Style-Based Generator Architecture for Generative Adversarial Networks. IEEE Transactions on Pattern Analysis & Machine Intelligence **43**(12), 4217–4228 (Dec 2021)
19. Krishnanunni, C.G., Bui-Thanh, T.: An Adaptive and Stability-Promoting Layer-wise Training Approach for Sparse Deep Neural Network Architecture (2024)
20. Kwon, G., Ye, J.C.: Diffusion-based Image Translation using disentangled style and content representation. In: The Eleventh International Conference on Learning Representations (2023)
21. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference (2023)
22. Nautiyal, R., Jha, R.S., Kathuria, S., Chanti, Y., Rathor, N., Gupta, M.: Intersection of Artificial Intelligence (AI) in Entertainment Sector. In: 2023 4th International Conference on Smart Electronics and Communication (ICOSEC). pp. 1273–1278. IEEE (2023)
23. OpenAI: DALL-E: Creating images from text (2021)
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
25. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. CVPR (2023)
26. Serengil, S.I., Ozpinar, A.: HyperExtended LightFace: A Facial Attribute Analysis Framework. In: 2021 International Conference on Engineering and Emerging Technologies (ICEET). pp. 1–4. IEEE (2021)
27. Shuttleworth, R., Andreas, J., Torralba, A., Sharma, P.: LoRA vs Full Fine-tuning: An Illusion of Equivalence (2024)
28. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations. In: International Conference on Learning Representations (2021)
29. Wang, J., Chan, K.C., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2555–2563 (2023)
30. Wang, L., Wang, L., Chen, S.: ESA-CycleGAN: Edge feature and self-attention based cycle-consistent generative adversarial network for style transfer. IET Image Processing **16**, 176–190 (2021)
31. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Timotfe, R., Van Gool, L.: Diffi2i: Efficient diffusion model for image-to-image translation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(3), 1578–1593 (2025)
32. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
33. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 2242–2251 (2017)