Generating Synthetic Data via Augmentations for Improved Facial Resemblance in DreamBooth and InstantID

Koray Ulusan
University of Tuebingen
koray.ulusan@student.uni-tuebingen.de

Benjamin Kiefer
LOOKOUT
University of Tuebingen
benjamin@lookout.team

Abstract

The personalization of Stable Diffusion for generating professional portraits from amateur photographs is a burgeoning area, with applications in various downstream contexts. This paper investigates the impact of augmentations on improving facial resemblance when using two prominent personalization techniques: DreamBooth *and* InstantID*. Through a series of experiments with diverse subject datasets, we assessed the effectiveness of various augmentation strategies on the generated headshots' fidelity to the original subject. We introduce* FaceDistance*, a wrapper around* FaceNet*, to rank the generations based on facial similarity, which aided in our assessment. Ultimately, this research provides insights into the role of augmentations in enhancing facial resemblance in* SDXL*-generated portraits, informing strategies for their effective deployment in downstream applications.*

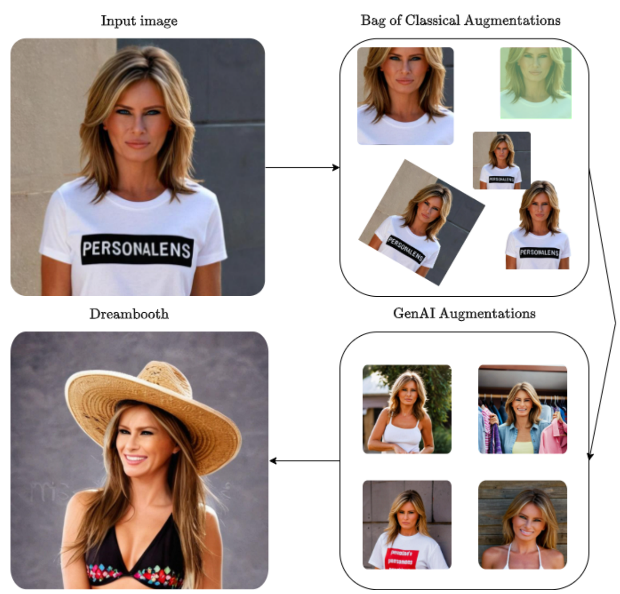

Figure 1. Pipeline for creating personalized images based on synthetically generated images through classical and GenAI-based augmentations for better downstream resemblance in DreamBooth-generated images.

1. Introduction

Personalized text-to-image generation has gained traction with the rise of models like Stable Diffusion (SD). However, training SD on small, user-specific datasets presents challenges, such as identity retention, overfitting, and artifact generation. Augmentation techniques are widely used in deep learning to improve generalization, but their role in personalized text-to-image generation is underexplored.

In this work, we analyze the effect of augmentations on personalized SD models trained with few-shot and zero-shot methods, particularly DreamBooth and InstantID. We investigate how different augmentations impact model performance and whether they enhance the realism and consistency of generated images.

In particular, we analyze both classical and generative augmentation strategies to bridge the gap between limited real data and high-fidelity synthetic outputs. By refining facial features and preserving identity through targeted GenAI-based augmentations, such as InstantID, we aim to improve the applicability of personalized generation in scenarios where synthetic data must closely mirror real-world characteristics. We analyze under which conditions we can ensure that "GenAI outputs improve GenAI outputs", avoiding a data quality collapse, providing best practices and heuristics.

Our contributions include:

- Analysis of classical augmentation techniques such as flipping, cropping, color enhancement, and background modifications.
- Using InstantID as a fast way of enhancing the user-specific dataset using the diffusion model itself.
- We conduct a survey to evaluate how white-collar work-

ers perceive personalized generations from DreamBooth and InstantID under various augmentation strategies.

2. Background and Related Work

In this section, we present the foundational concepts and prior research relevant to our work on augmentation techniques for few-shot personalization in diffusion models.

2.1. Text-to-Image Diffusion Models

Text-to-image diffusion models generate high-quality images from natural language descriptions by gradually denoising random Gaussian noise guided by text embeddings [12, 15]. Stable Diffusion [15] employs a latent diffusion approach that operates in a compressed latent space rather than pixel space, reducing computational requirements while maintaining generative quality.

2.2. Subject-Driven Image Generation

Subject-driven image generation creates images featuring specific subjects with high fidelity while maintaining their identity across contexts [16]. Key approaches include:

DreamBooth [16] fine-tunes the U-Net of Stable Diffusion using 3-5 images of a specific subject. It preserves the semantic prior through class-specific prior preservation loss and uses a rare token with weak prior to refer to the subject with the prompt format "a [V] [class noun]".

InstantID [20] is a zero-shot method that combines facial feature extraction with text conditioning. It extracts five key facial landmarks to condition the position and orientation of the generated face, providing greater control over the output.

We standardize our experiments using the same SDXL model for both techniques to ensure fair methodological comparison.

2.3. Image Augmentation Techniques

2.3.1. Classical Image Augmentations

Classical image augmentation techniques include geometric transformations (flipping, rotation, scaling, cropping), photometric adjustments (brightness, contrast, saturation, hue), and noise injections (Gaussian, salt-and-pepper). These predefined transformations maintain semantic integrity while introducing controlled diversity to expand limited training datasets.

2.3.2. Augmentations in Diffusion Models

Data augmentation enhances diffusion model performance while reducing computational demands [19]. Key approaches include mixing-based augmentations that interpolate between existing samples [9] and consistency regularization techniques that enforce invariance to specific transformations [8, 11]. Our work investigates these techniques specifically for few-shot personalization applications.

2.4. Face Processing Approaches

FaceNet [17] maps facial images to a 128-dimensional embedding space where similar faces are positioned closely together. The standard pipeline uses MTCNN [22] for face detection before embedding generation, with cosine distance metrics for similarity assessment [18].

Alternative approaches include faceswapping methods [6, 10] and augmented reality techniques for virtual try-on applications [7]. While these provide real-time capabilities, they often lack the flexibility and integration capabilities of diffusion-based approaches.

Our research builds on these foundations to investigate how strategic data augmentation can improve few-shot personalization in diffusion models, focusing on identity preservation and recontextualization.

3. Methodology

We use augmentations across various Subject Datasets to see if there is an overall improvement in generated pictures.

3.1. Subject Datasets.

Our dataset consists of 3 to 15 images per participant, with $n = 10$ participants. To maintain a naturalistic data collection process, we instructed them: *"Can you send me portrait/selfie-style photos of your face in different places? The more different places, the better."* By avoiding rigid guidelines, we ensured that the collected images reflect realistic user behavior. As a result, our findings are well-aligned with real-world data distributions, enhancing the transferability and applicability of our results.

Our dataset exhibits a diverse range of environmental conditions, facial orientations, and image qualities, ensuring variability that mirrors real-world scenarios. The images encompass different backgrounds, lighting conditions, and subject behaviors, contributing to a dataset that is both representative and robust. For instance, some images feature cluttered or irregular backgrounds (e.g., *Baker-Zoe*, *Bottle-Hugo*), while others are captured in controlled environments (*Biometric-Kora*). Variation in gaze direction is also present, with *Doctor-Nina* not looking at the camera, while *3D-Gary* includes dynamic head movements extracted from a video. Additionally, differences in personal appearance and accessories are observed, such as *Farmer-Lisa* wearing a helmet and *Staircase-Judy* wearing makeup. Lighting conditions range from well-lit (*Vacation-Anna*) to suboptimal (*2024-Kora*), further enhancing the dataset's realism. These characteristics make our dataset a valuable resource for evaluating model performance under unconstrained, real-world conditions.

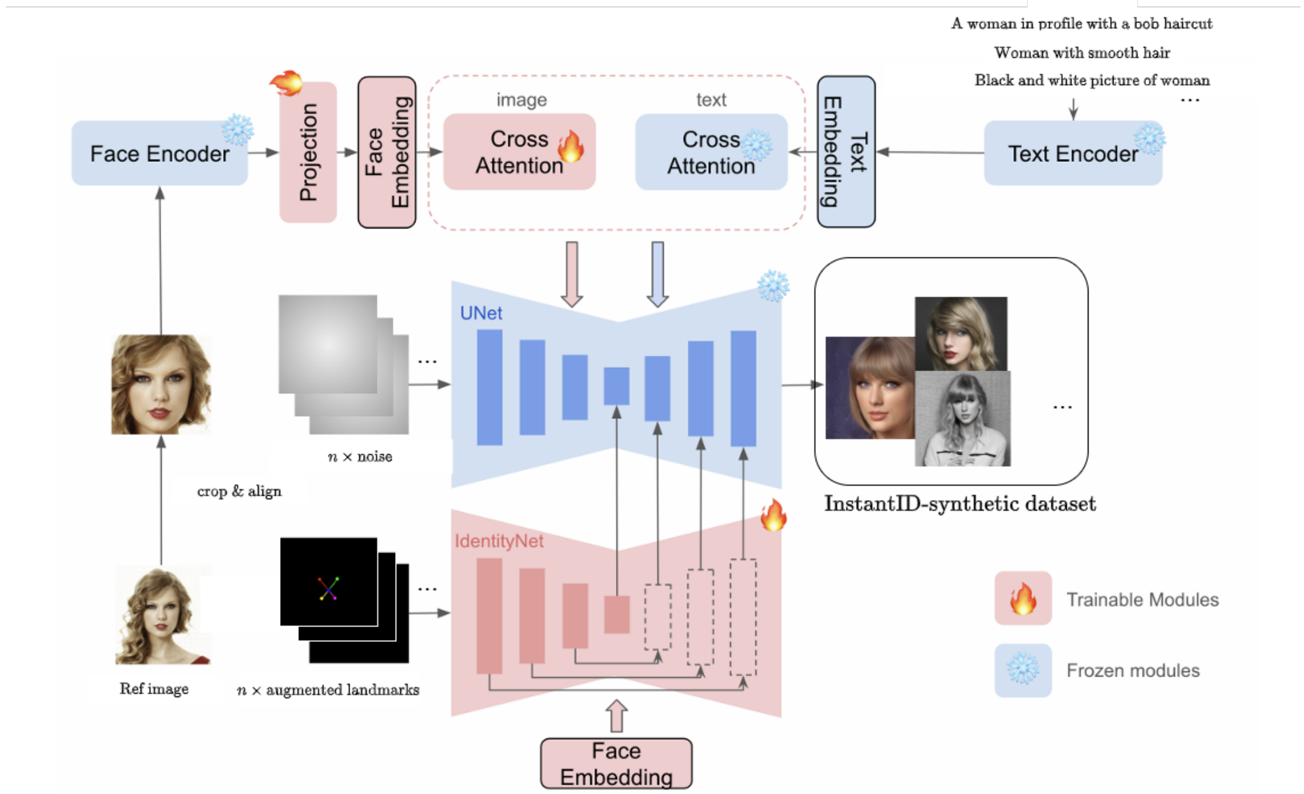

Figure 2. Pipeline for creating Gen-AI augmented personalized data via InstantID. Based on one or more input images of a person, we run it through the InstantID pipeline but with augmented landmarks and prompts. The landmarks are taken from the input image and slightly perturbed for good resemblance. The collected synthetic dataset is then further used for downstream DreamBooth training. Figure modified from [20].

3.2. Dataset Augmentations

We apply augmentations individually to evaluate each technique's performance improvement independently.

Classical Augmentations Standard techniques include: (i) *Random Horizontal Flip* with $p \in \{0, \frac{1}{2}, 1\}$, and (ii) *Color Jitter* varying brightness, contrast, saturation $(\pm 5, \pm 15)$ and hue $(\pm 5°)$.

Background Augmentation We use U^2-Net [14] for subject isolation, testing both base and human segmentation models. Backgrounds include *flat colors*, *patterns* from Wikimedia [5], and Flickr Places.

Blending Techniques We separately evaluate *Alpha Blending* and *Poison Blending* through both automated and manual techniques.

Resizing Methods We compare: (i) downsampling then upsampling, (ii) upsampling only, and (iii) original dimensions. Methods include ESRGAN [21], Lanczos, and Bilinear.

Cropping Strategies Five approaches: (i) SDXL dimensions [13], (ii) automated center cropping to 1MP, (iii) downsample-then-crop to 1MP at various aspect ratios, (iv) manual eight-variation cropping, and (v) MTCNN face-based cropping.

Color Adjustment Adobe Lightroom auto-adjustment enhances visual quality.

Generative Augmentation Using InstantID, we generate new subject images with prompts from `dolphin 2.2.1 - Mistral 7B` [3] and varied facial landmarks (Figure 2).

3.3. Hardware, Software, and Hyperparamters

All experiments were conducted on a single NVIDIA GeForce RTX 3090 with 24GB VRAM. We use `sd-scripts`[4] for DreamBooth and `ComfyUI_InstantID`[1] for InstantID experiments, inheriting all bias in their pipeline, if any.

Results of DreamBooth finetuning a diffusion model (DM) greatly depends on the DMs ability of generating images. We use `RealVisXL_V4.0` [2], which is a community finetune of SDXL for realistic image generation.

Default prompt we use "A professional headshot of a *subject* wearing a suit standing in a well-lit studio, DSLR" as the default prompt. Empirical evidence suggests including the gender as *man* and *woman* gives generated images gender characteristics based on western culture, which we preferred.

For DreamBooth, we use "a $[V]$ [man|woman]" where $[V]$ denotes the rare token. InstantID doesn't have a special prompt and work with any text. LLM generated prompts are useful in both.

3.4. FaceDistance Metric

For a subject image dataset, we calculate their embedding using FaceNet [17], which maps similar faces to similar locations on a hypersphere. Then, we calculate the mean face vector v_{real}. For a given generated face, we measure its embedding vectors cosine distance to v_{real}.

FaceDistance is a useful technique for distinguishing between "good" and "bad" generations. This can be used to rank generated images based on their similarity (lower is better). It can be used to discard the largest n% of distances to improve personalization pipelines.

For our subject datasets, the mean cosine distance of v_{real} to real images is $\bar{v}_{within\ real} \approx 0.13$. We notice $\max(v_{within\ real}) = 0.22$ and $\min(v_{within\ real}) = 0.05$.

4. Experiment Results

We try to achieve higher facial similarity via DreamBooth and InstantID using highlighted augmentations.

Despite selecting a realistic image generation model, achieving photorealistic generation of an individual's face remains challenging without imposing strict constraints on the subject images. We have relaxed many of these constraints to enhance usability, as expecting an average user to compile a dataset of themselves without understanding the underlying image generation techniques presents a significant challenge. Ensuring high subject fidelity is crucial for these methods to be effective in downstream applications, as humans are highly sensitive to variations in facial features compared to textures.

One major issue with datasets without great constraints is that the images is not a good representation of the person. It can be compared to having difficulty recognizing a person in real life whom you only saw in photographs. We observe this phenomenon for small datasets with $\text{size} \leq 3$. In these cases, the generated images is a good reflection of the dataset (if someone doesn't know them in real life, they are likely to claim that these pictures are good. Otherwise the generated images are not a good representation of the real person).

4.1. DreamBooth

We configure our hyperparameters such that recontextualization capabilities can be sacrificed for high facial fidelity. Identity preservation is hard in DreamBooth. so we rather overfit to achieve high subject fidelity and have limited freedom in generations.

The common theme in augmentations is that if the augmented image has any kind of artifact/anomaly, then the rare token will be associated with it. The supporting observations are (i) When background is replaced with a geometric pattern (from wikimedia patterns), the model will focus on learning the pattern than the subject (ii) When image is upscaled with ESRGAN, the texture ESRGAN introduces say present in generations (iii) the masks generated with U^2-Net is not pixel-perfect. and results in a mix around hair/air boundary. This mix becomes associated with the subject. The human segmentation models training data was not highly accurate around hairs but was better in identifying body parts. The base model is performs better around hair and was overall better. The robustness of human segmentation model is not needed. (iii) any kind of *color jitter* is visible in generated images. For example the saturation change of 0.1 is clearly present in generations. (iv) using Adobe Lightroom as a preprocessing step resulted in better color graded generations compared to non-preprocessed datasets. (v) datasets with low contrast (e.g. exclusively Polaroid pictures) resulted in copying the photography style/lighting from the pictures — though this can be attributed to our hyperparameter configuration.

Because of the low recontextualization capabilities, backgrounds becomes highly associated with the rare token. Replacing the background with **Pastel Colors** and **Rainbow Colors** led to eccentric and often unrealistic images, with the latter occasionally generating pictures without subjects. **Gray** offered the highest resemblance to the subject, while **Dark Gray** caused the model to disassociate the subject from its context. Because of problems with U^2-Net, **Light Gray** background outperformed **Dark Gray**, especially in bright environments.**Wikimedia Patterns** slowed down learning and degraded the image quality across all generations. Lastly, **Studio Backdrops** introduced irregularities that reduced the quality of the generated images which can be thought as similar to Wikimedia Patterns because backdrops has patterns.

Random Horizontal Flip slowed learning due to face asymmetry, which confused facial features. **Random Rotation** caused distorted images and introduced black padding bars, which also can be seen in augmented subject images. **Color Jitter** led to undesirable results, as brightness, contrast, saturation, and hue changes were linked to rare tokens, causing erratic generations.

Both **Alpha Blending** and **Poison Blending** are discouraged, as they require careful manual processing to achieve

(a) Real Images

(b) DreamBooth results with classical augmentations (crop, resize, and color)

(c) DreamBooth results with GenAI augmentations & without classical augmentations

Figure 3. Example improvement of including Instantid generated images in the subject dataset *Vacation-Anna*. The model is prompted with *default prompt* with batchsize 4. The results are **not** cherry-picked to resemble the downstream application use. Although (b) is visually more interesting, the method in (c) is more consistent across many subject datasets.

good results. These techniques are not straightforward to apply and can lead to undesirable artifacts if not handled properly.

Images around **1 Megapixel** performed best, providing a balanced resolution for high-quality generation. **Upscaling with ESRGAN** introduced visible artifacts, especially around facial features. **Upscaling with Lanczos** was effective, particularly when starting from larger images. However, if the initial dataset contained low-resolution images, the generated images exhibited facial blurring due to the nature of the Lanczos algorithm. The difference between bicubic and Lanczos was negligible. **Downscaling** resulted in lower-quality generations compared to using original-sized images. It should be noted that our testing output resolution was 1024×1024.

InstantID Augmentation Datasets augmented with InstantID yield clearly superior performance. The added images need to be diverse (i.e., generated with various text conditioning and different keypoint images). Since we trade recontextualization abilities for increased facial similarity, generating the same person in similar contexts is beneficial. DreamBooth achieves similar facial similarity compared to InstantID but allows for greater control. The rigidity caused by the keypoint images is eliminated. However, this method is more computationally expensive than raw InstantID. Additionally, achieving proper prompt diversity can be challenging. I prefer InstantID over DreamBooth.

The ratio of real to InstantID-generated images depends entirely on the diversity of the generated images. One rule of thumb is that no single concept should comprise more than 25% of the dataset. For example, if images labeled as "a [V] man in a library" exceed 25%, DreamBooth training will associate the rare token with the concept. This results in a final DreamBooth model that is unusable due to a complete loss of recontextualization ability caused by overfitting.

Since InstantID generations are highly realistic, one can generate additional images with it to better represent the subject during DreamBooth training. We use the same diffusion model for both InstantID and DreamBooth to integrate the subject more effectively into the model without altering the subject's context. This ensures that the dataset distribution remains closer to the diffusion model's generation space.

4.1.1. FaceDistance

We tried to select the "best" DreamBooth checkpoint by generating images of "a [V] man" in different contexts for all checkpoints and ranking them using *FaceDistance*. This method was able to discard obviously bad checkpoints (e.g. anomalies in generations, unable to generate the subject, divergence) but is not able to rank "good enough" checkpoints within themselves. (Figure 4 The FaceNet manifold isn't sensitive to very similar looking people. For a given hyperparameter configuration, a few tests show when the model will be converged to its best state (usually between 3k and 6k steps) and since FaceDistance isn't able to differentiate betweent them, FaceDistance isn't a useful tool for this purpose.

Despite these challenges, FaceDistance appears to be

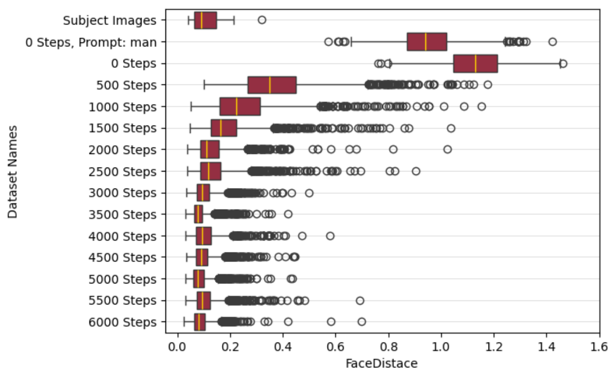

Figure 4. FaceDistance Distributions of 2000 Samples from Different Saved Dreambooth finetunes of SDXL Real. *Closeup-Kora* is used. The KDE for each looks like a normal distribution.

functioning for loosely ranking generated images. This improves the user experience.

4.2. InstantID

The effectiveness of InstantID is highly dependent on the quality and characteristics of the provided reference images.

4.2.1. Face Embedding

We conducted experiments to determine the optimal number of reference images that balances usability and facial similarity. Our findings confirm those of [20], demonstrating that using multiple reference images results in increased facial similarity. When only one reference image is provided, the generated face is heavily influenced by the specific appearance captured in that single image. We attribute this limitation to insufficient information being extracted by the Face Encoder from a single perspective. Our analysis indicates that four reference images provide satisfactory results in most cases, with diminishing returns observed beyond this number. Since reference images are cropped and aligned before being processed by the face encoder, users have considerable flexibility in selecting images without compromising model performance.

4.2.2. Landmarks Image

We observe that facial landmarks exert strong conditioning influence, often rendering text prompts ineffective for controlling the subject's position. The generated image consistently replicates the face placement, orientation, and size specified by the provided keypoints, due to the five-point landmark system employed.

For practical applications, users frequently struggle to understand how face positioning in the landmark image transfers to the generated output. This communication challenge often results in user dissatisfaction with generated images, despite the issue stemming from suboptimal conditioning input.

To address this limitation, we propose two solutions: *2-shot generation* and *face replacement*.

In **2-shot generation**, we collect subject reference images $(s_1, \ldots, s_n)$ and a separate image representing the desired pose and composition s_{kpts}. These are used as reference images and the keypoints image, respectively. While the resulting output s_{out} is generally satisfactory, using facial landmarks from one person to generate another reduces facial similarity due to structural differences in the five keypoints (eyes, nose, mouth). We hypothesize this stems from imbalanced conditioning weights. Performance improves when replacing s_{kpts} with a previously generated image of the subject, yielding better facial similarity while maintaining compositional control.

In **face replacement**, users interact with a simple tool to manipulate (move/rotate/resize) their cropped face on a canvas matching the diffusion model's output dimensions. This approach eliminates the similarity issues caused by using another person's facial landmarks. However, the method performs poorly when none of the reference images show the subject facing the camera (deviations > 30 degrees). User satisfaction was higher with this approach compared to *2-shot generation*, which we attribute to increased interactivity and faster generation times.

4.2.3. Augmentations

Due to InstantID's architectural design, rotational and shape-altering augmentations proved ineffective. Background replacement and similar context-modifying augmentations degraded similarity because the resulting artifacts fall outside the distribution of images encountered during training by the model provided in [20]. The trained model demonstrates robustness to meaningful color adjustments, rendering color modifications unnecessary for well-lit scenes. For low-resolution images, traditional upscaling methods (Lanczos/bicubic) performed adequately, while neural network-based upscaling introduced novel artifacts unseen during training, resulting in reduced quality.

5. Survey

We conducted the survey to evaluate the viability of AI-generated portraits for professional use and to compare the performance of DreamBooth and InstantID in generating realistic headshots. 97 white-collar workers from diverse professional backgrounds participated in the online survey. Numerical data can be found in Suppl. 11 and questionary can be found in Suppl. 12.

Overall Performance of Generated Portraits Portraits generated by DreamBooth and InstantID performed similarly across multiple aspects, including overall quality, facial detail clarity, identity preservation, perceived level of editing, and background quality. Using high-quality subject datasets led to slightly better results in most categories, except for "Editing," where participants indicated familiarity and acceptance of traditional Photoshop-enhanced portraits.

Method Preferences A slightly higher percentage of participants (4%) preferred the standardized portraits from InstantID over the more flexible outputs of DreamBooth. InstantID was often perceived as more professional, likely due to its consistent "Photoshopped look," which resonated with a broader audience. Open-ended responses highlighted diverse preferences, with participants emphasizing factors such as lighting, pose, angle, expression, detail, color, and background.

Facial Similarity DreamBooth demonstrated superior facial similarity between real images individuals and their generated portraits compared to InstantID. More participants identified InstantID images as depicting a different person than the reference. DreamBooth consistently maintained a higher level of facial similarity across both high- and low-quality subject datasets.

Noticing AI Generations Most white-collar workers struggled to identify AI-generated headshots when not explicitly prompted, often focusing on well-known but absent flaws commonly associated with AI generation. Among a subset of participants ($n = 77$) who regularly notice AI-generated images in daily life, the generated portraits blended well with conventional studio photographs. However, participants who actively use AI for image creation ($n = 29$) demonstrated better identification skills. This group was more likely to recognize DreamBooth images as AI-generated, possibly due to DreamBooth's popularity, while InstantID generations, being more niche, had a near $50/50$ chance of being identified as AI.

6. Discussion

Our experiments offer insights into augmentation strategies for improving facial resemblance in personalized text-to-image generation using DreamBooth and InstantID. While classical augmentations are common in deep learning, applying them to few-shot personalization can yield undesirable results. Geometric transformations like flipping and rotation disrupted learning due to face asymmetry and artifacts. Color jittering caused erratic generations by associating color shifts with DreamBooth's rare token. Background augmentations with U^2-Net introduced segmentation imperfections, especially around hair, which the model learned. Replacing backgrounds with patterns or studio backdrops also degraded image quality. However, auto color adjustment with Adobe Lightroom improved color grading.

Generative augmentation via InstantID proved more effective for enhancing facial similarity in DreamBooth training. By generating diverse synthetic images with varied prompts and facial landmarks, we enriched the dataset with realistic examples, aligning it with the diffusion model's space. However, maintaining a balance between real and InstantID-generated images is crucial to avoid overfitting and loss of recontextualization.

FaceDistance provided a quantitative measure of facial similarity but became less useful for hyperparameter tuning once a certain fidelity level was reached. A user survey among white-collar workers showed that both DreamBooth and InstantID performed similarly in quality, clarity, identity preservation, editing, and background. A slight preference emerged for the "Photoshopped look" of InstantID portraits. While DreamBooth achieved better facial similarity, many participants struggled to distinguish AI-generated images from real ones, particularly those unfamiliar with AI tools. Users actively engaged in AI image creation were more likely to identify DreamBooth images as synthetic, possibly due to its higher popularity.

InstantID's effectiveness depends on reference image quality and diversity. Using multiple references (around four) improved similarity by enriching information for the Face Encoder. Facial landmarks strongly influenced pose and composition, sometimes overriding text prompts. We explored 2-shot generation and interactive face replacement to enhance control, with the latter showing higher user satisfaction. Rotational and shape-altering augmentations were ineffective, and background modifications reduced similarity. Traditional upscaling worked well for low-resolution images, whereas neural network-based upscaling introduced artifacts.

7. Limitations

A key limitation is that InstantID-based augmentation reduces realism in generated images. While DreamBooth remains more flexible for personalized generation, InstantID-enhanced datasets still outperform unaugmented ones. Given the baseline model's photorealism constraints, using generative augmentation to refine its training data is a practical approach.

8. Conclusion

This study examined augmentation strategies for improving facial resemblance in personalized image generation using DreamBooth and InstantID. Classical augmentations can introduce artifacts that degrade facial fidelity, requiring careful application.

We found generative augmentation with InstantID to be highly effective for improving DreamBooth training. Creating diverse, realistic synthetic images while maintaining a balanced ratio with real data prevents overfitting.

User surveys confirmed that both DreamBooth and InstantID produce high-quality, professional-looking headshots, often indistinguishable from real photos. While DreamBooth excels in facial similarity, InstantID's consistent output appears more polished.

For practical use, employing multiple reference images enhances facial information capture. Improving control over pose and composition through landmarks is crucial, with interactive face replacement showing promise.

Overall, our findings provide insights into augmentation strategies for personalized image generation, guiding their application in tasks requiring high facial fidelity. Future work should explore advanced generative augmentation techniques and better user control over InstantID outputs.

References

[1] ComfyUI_InstantID, 2025. [Online; accessed 30. Mar. 2025]. 3

[2] SG161222/RealVisXL_V4.0 · Hugging Face, 2025. [Online; accessed 30. Mar. 2025]. 3

[3] TheBloke/dolphin-2.2.1-mistral-7B-GGUF · Hugging Face, 2025. [Online; accessed 30. Mar. 2025]. 3

[4] sd-scripts, 2025. [Online; accessed 30. Mar. 2025]. 3

[5] Category:Patterns - Wikimedia Commons, 2025. [Online; accessed 31. Mar. 2025]. 3

[6] Aditya Agarwal, Bipasha Sen, Rudrabha Mukhopadhyay, Vinay Namboodiri, and C. V. Jawahar. FaceOff: A Video-to-Video Face Swapping System. *arXiv*, 2022. 2

[7] Ting-Yu Chang and Seretsi Khabane Lekena. GlamTry: Advancing Virtual Try-On for High-End Accessories. *arXiv*, 2024. 2

[8] Khawar Islam and Naveed Akhtar. Context-guided Responsible Data Augmentation with Diffusion Models. *arXiv*, 2025. 2

[9] Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. DiffuseMix: Label-Preserving Data Augmentation with Diffusion Models. *arXiv*, 2024. 2

[10] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping. *arXiv*, 2019. 2

[11] Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. MagicMix: Semantic Mixing with Diffusion Models. *arXiv*, 2022. 2

[12] Alex Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. *arXiv*, 2021. 2

[13] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv*, 2023. 3

[14] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaiane, and Martin Jagersand. U^2-Net: Going Deeper with Nested U-Structure for Salient Object Detection. *arXiv*, 2020. 3

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models, 2022. [Online; accessed 31. Mar. 2025]. 2

[16] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *arXiv*, 2022. 2

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. *arXiv*, 2015. 2, 4

[18] Sefik Serengil and Alper Özpınar. A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules. *Bilişim Teknolojileri Dergisi*, 17(2):95–107, 2024. 2

[19] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective Data Augmentation With Diffusion Models. *arXiv*, 2023. 2

[20] Qixun Wang, Xu Bai, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. InstantID: Zero-shot Identity-Preserving Generation in Seconds. *arXiv*, 2024. 2, 3, 6

[21] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, and Chao Dong. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. *arXiv*, 2018. 3

[22] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. *arXiv*, 2016. 2