CROSS-DOMAIN FACE CLASSIFICATION:
TRANSFER BETWEEN REAL AND DIFFUSION-GENERATED IMAGES

AUTHORS

Gerardo Gómez Argüelles
Oliver Tausendschön

June 2025

INTRODUCTION

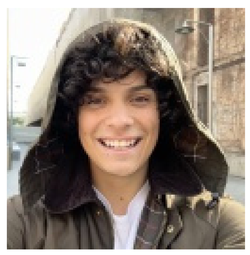
Face classification models typically rely on real image datasets, but the emergence of diffusion models raises a question: how effectively do these models transfer between real and synthetic domains? While recent work has applied diffusion models for augmentation, their use for facial classification remains largely unexplored. (See references at the end.)

In this project, we conduct the first comprehensive bidirectional analysis of cross-domain face classification, evaluating transfer performance between real photographs and Stable Diffusion-generated images. Our findings reveal significant domain gaps and surprising insights about when synthetic data helps or hurts classification performance.

DATASET

To build a dataset, we recorded short portrait-style videos of three individuals in varied settings and lighting conditions. Each video was processed by extracting frames at a fixed interval using a custom Python script. This ensured both quantity and diversity without requiring manual photo collection.

Using this approach, we gathered approximately 5,500 facial images per person, resulting in a total dataset of around 16,500 images across three classes: Gerardo, Oliver, and Timothy. All images were preprocessed to standardize size and aspect ratio, and each was labelled to enable conditional generation. An example can be found below.


METHODOLOGY

To augment our dataset, we apply a Stable Diffusion 2.1 model, a latent diffusion architecture that generates images by iteratively denoising a latent representation, conditioned on a text prompt. Rather than training from scratch — which is computationally expensive and infeasable for us — we adopt a parameter-efficient fine-tuning strategy using LoRA (Low-Rank Adaptation).

We fine-tune only a small number of trainable weights injected into the UNet's attention layers in the end, while keeping the base model the same. This enables effective adaptation to our dataset of three individuals with minimal resource overhead.

Each image in the dataset is paired with a label and the model is trained to reconstruct noisy latent inputs given these. During training, only LoRA weights are updated using a mean squared error loss on the predicted noise. This allows us to generate synthetic face images guided by text, which we later use to augment the training set for classification.

Initially, we also implemented a conditional DDPM Diffusion model (another class of diffusion models) but it turned out to not generalize very well. It made more sense to opt for Stable Diffusion as it provides higher quality more general images. the models applied can be found in huggingface (DDPM Diffusion, Stable Diffusion) and the corresponding code in Github.

To evaluate cross-domain transfer between real and synthetic data, we trained a ResNet-18 classifier for face classification. The model achieves near-perfect accuracy (~100%) on real data, indicating that the three individuals have sufficiently distinct facial features for reliable classification. However, our primary focus is not baseline performance but rather cross-domain transfer capabilities. We designed five key experiments: (1) Real→Real as baseline, (2) Real→Diffusion to test forward transfer, (3) Diffusion→Diffusion to establish synthetic baseline, (4) Diffusion→Real to test reverse transfer, and (5) Combined→Diffusion to evaluate data mixing effects.

The model is trained using cross-entropy loss with Adam optimizer, and performance is evaluated using classification accuracy, F1-score, and confidence metrics. Data augmentation includes random horizontal flips, rotation, color jittering, and random grayscale conversion (30% probability) to test color dependency. This comprehensive evaluation framework allows us to quantify domain gaps and transfer effectiveness in both directions.

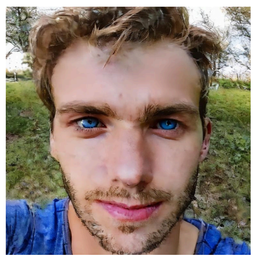
STRATEGY

To explore the effectiveness of synthetic images for data augmentation, we designed a multi-stage strategy centered on prompt-controlled image generation and comparative classifier training: Each of the three individuals in our dataset was synthetically created with a series of prompts that aimed at capturing different levels of deatils/complexity. In total, we generate 50 images per class and promt, totaling 250 images per class.

- Gerardo
  - "a selfie of a young man named Gerardo, curly black hair"
  - "a selfie of a young man named Gerardo"
  - "a selfie of a person Gerardo"
  - "a picture of a person Gerardo"
  - "a selfie of a young man named Gerardo, curly black hair, brown eyes, mexican looking"
- Oliver
  - "a selfie of a young man named Oliver, blonde hair"
  - "a selfie of a young man named Oliver"
  - "a selfie of a person Oliver"
  - "a picture of a person Oliver"
  - "a selfie of a young man named Oliver, blonde hair, blue eyes, european looking"
- Timothy
  - "a selfie of a young man named Timothy, long brown hair, bearded"
  - "a selfie of a young man named Timothy"
  - "a selfie of a person Timothy"
  - "a picture of a person Timothy"
  - "a selfie of a young man named Timothy, long brown hair, bearded, brown eyes, european looking"

An example of generated images can be found below.





These prompt variations were used with the fine-tuned Stable Diffusion model to generate different versions of class-specific images. This systematic approach allows us to analyze how different prompt complexities affect synthetic image quality and classifier performance.

Each configuration isolates different aspects of cross-domain transfer:

Train Dataset	Test Dataset	Objective
Real	Real	Establishes baseline performance on the target domain
Synthetic	Real	Reveals whether synthetic images provide meaningful signal
Real + Synthetic	Real	Evaluates whether a model trained only on generated data generalizes to real-world inputs — a domain adaptation test
Sythetic	Synthetic	Tests within-domain performance on synthetic data
Real + Synthetic	Synthetic	Assesses whether combining domains improves synthetic performance

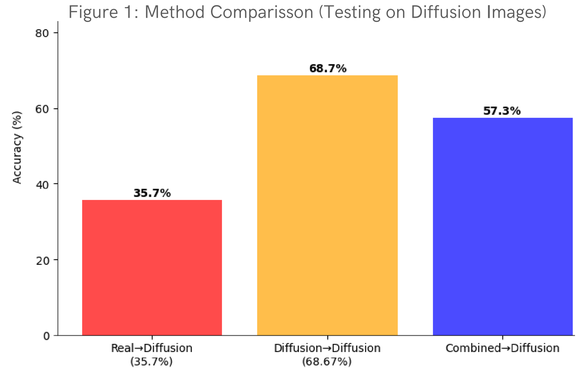
RESULTS/FINDINGS

Our comprehensive cross-domain evaluation reveals significant domain gaps and surprising transfer patterns between real and synthetic face images.

Cross-Domain Transfer Performance

Training → Testing	Accuracy	Key Finding
Real → Real	100%	Perfect baseline
Real → Diffusion	35.70%	Severe forward transfer drop
Diffusion → Diffusion	68.70%	Reasonable synthetic baseline
Diffusion → Real	48.00%	Better reverse transfer
Combined → Diffusion	57.30%	Domain interference (-11.4%)

Figure 1: Method Comparisson (Testing on Diffusion Images)



Key Results:

1. Asymmetric Domain Gap: The 12.3% difference between reverse (48.0%) and forward (35.7%) transfer suggests diffusion images contain more transferable features than initially expected.

2. Negative Transfer Effect: Combining real and synthetic data hurts performance compared to diffusion-only training, demonstrating domain interference rather than complementary learning. (Figure 1)

3. Person-Specific Recognition: Oliver achieved highest recognition (57%) while Gerardo proved most challenging (40.9%), indicating individual variation in diffusion generation quality.

4. Prompt Engineering Impact: Simple prompts often outperform complex ones - "a picture of a person [Name]" (40.0%) vs detailed descriptions, suggesting specificity bias can hurt generation. (Figure 3)

Figure 2: Diffusion→Real individual performance breakdown

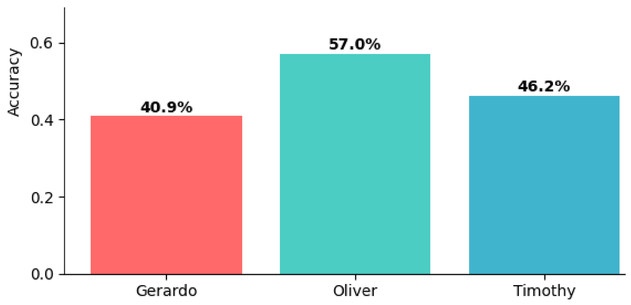
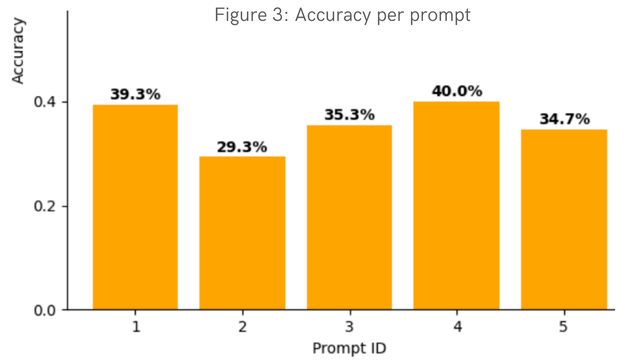


Figure 3: Accuracy per prompt



CONCLUSION

This study fully explores how well face recognition models work when trained on real photos and tested on AI-generated ones—and vice versa. It reveals important findings about how useful synthetic images really are.

Key Findings: We found that models trained on fake (diffusion) images perform better on real images (48.0%) than the other way around (35.7%). Surprisingly, mixing real and fake images during training actually makes performance worse (57.3%) compared to just using real images alone (68.7%). This challenges the common belief that more data diversity always helps.

Implications: These results suggest that it's usually better to train a model on one type of data (real or synthetic) rather than mixing both, especially when you want good performance across different types.

Novel Contributions: We establish classifier accuracy as an objective evaluation metric for diffusion model quality, provide practical prompt engineering guidelines (simple prompts outperform complex ones due to specificity bias), and quantify person-specific generation quality variations.

Future Work: Using a more diverse dataset could provide more stable results, as our training (real) images are taken from video frames, most of which are similar and might be generating overfitting.

References:

Akrout et al., Diffusion-based Data Augmentation for Skin Disease Classification (2023)
Sousa et al., Data Augmentation in Earth Observation: A Diffusion Model Approach (2024)
Wang & Chen, Diff-II: Diffusion-based Augmentation for Data-Scarce Classification (2024)