Exploiting Temporal Relationships: Enhancing GAN Training through Adaptive Temporal Augmentations

Anonymous Authors[1]

Abstract

This paper addresses the challenge of training high-quality generative adversarial networks (GANs) with limited data. I first review three relevant papers that propose novel techniques to improve GAN training and image generation using limited datasets. Building upon the insights gained from these papers, I propose a new approach called temporal augmentation, which enhances the performance of GAN models by incorporating temporal dynamics into the training process. I conduct a series of experiments using a modified StyleGAN2 ADA framework and a dataset of GIFs featuring human faces. The results demonstrate that temporal augmentation leads to marginally improved model performance, highlighting the potential of this technique to enhance the training of GANs on limited datasets that already have a temporal component.

1. Introduction

Generative adversarial networks (GANs) have shown great promise in generating realistic images [1]. However, training GANs with limited data remains a challenge. Acquiring large and diverse datasets for training can be costly and impractical, limiting the applicability of GANs in various domains. To address this problem, I explore three papers that propose innovative approaches to improve GAN training and image generation using limited data.

The first paper focuses on training GANs with limited data by leveraging transfer learning and data augmentation techniques [1]. The second paper introduces a novel framework that reparameterizes the latent generative space as a mixture model to capture image modality diversity [2]. The third paper presents a method for text-driven image manipulation using StyleGAN and CLIP [3].

In my extended work, I aim to introduce a novel augmentation approach: temporal augmentation. By incorporating temporal dynamics into the augmentation process, the GAN model can hopefully become more resilient on datasets containing temporal information than with other augmentation strategies alone. This not only aligns with the objective of the original paper in enabling GANs to perform well with limited datasets, but also extends the benefits to harness the inherent structure present in temporal data. This new augmentation technique seeks to move beyond the agnostic replication of identical training material, and instead facilitates the creation of a more diverse and robust data-feed, contributing to the advancement of machine learning models in scenarios involving temporal data.

If the implementation proves successful, it would suggest that the temporal character of specific datasets could be leveraged to potentially create more robust AI models, and that incorporating temporal augmentations within the ADA framework could further enhance the utilization of limited datasets by GANs.

2. Review of Paper to Implement (Paper 1)

2.1. Storyline

Introduction The paper ”Training Generative Adversarial Networks with Limited Data” proposes an adaptive discriminator augmentation mechanism to improve the training process and achieve better results when training GANs with limited data.

High-Level Motivation/Problem The paper addresses the need to train high-quality generative models with limited data. GANs have shown promise in generating realistic data, but their performance is often hindered by the lack of large and diverse training datasets. Acquiring such datasets can be costly and impractical, limiting the application of GANs in various fields. The paper aims to make training high-quality generative models with small custom datasets more feasible, thereby expanding the use of GANs in different research fields.

Prior Work on the Problem Prior research has explored transfer learning and data augmentation techniques to address the challenge of training GANs with limited data. The paper ”TransferGAN: A Generative Model for Small Data Transfer Learning” introduces TransferGAN, which leverages transfer learning to improve GAN performance with limited data. The paper ”Data Augmentation Generative Adversarial Networks” introduces DAGANs, which use data augmentation strategies to enhance GAN training with limited data.

Research Gap Despite previous efforts, there is still a research gap in stabilizing GAN training and achieving improved results with limited data. Existing methods often struggle with overfitting, where the discriminator becomes too good at distinguishing between real and generated data. This leads to training divergence and poor performance of the generator. Addressing this overfitting issue and finding effective ways to stabilize training with limited data is the key research gap that this paper aims to fill.

Contributions The main contributions of the paper are:

1. An adaptive discriminator augmentation mechanism to prevent overfitting in GANs trained with limited data.
2. A diverse set of augmentations that do not affect the generated data.
3. Demonstration of the method’s effectiveness on different datasets, achieving comparable results to existing methods with significantly fewer training images.
4. Improved performance on the widely used CIFAR-10 dataset.

2.2. Proposed Solution: Adaptive Discriminator Augmentation Mechanism

The proposed solution is an adaptive discriminator augmentation mechanism that applies a diverse set of augmentations to the training data seen by the discriminator. This mechanism prevents overfitting and improves the stability of GAN training with limited data. The augmentations are carefully designed to not affect the generated samples.

Mathematically, the adaptive discriminator augmentation mechanism can be represented as follows: The discriminator input is denoted by x_{train}, representing the training data with augmentations.

Augmentations are defined as $A = \{a_1, a_2, \ldots, a_n\}$, which is the set of augmentations applied.

The probability of applying an augmentation is denoted by p, and it is controlled by hyperparameter tuning.

During training, the discriminator receives augmented training data x_{train}, where each sample undergoes a sequence of augmentations chosen from the set A based on the hyperparameter p. This ensures that the discriminator sees a diverse range of augmented training examples, preventing overfitting to specific augmentation patterns. This process essentially functions as the discriminator’s ”goggles”, enabling it to see a diverse range of strategically augmented training examples. It is this diversity, created by the adaptive application of augmentation patterns, that prevents the discriminator from overfitting to any specific pattern.

Experimental Results Extensive experiments were conducted on various datasets, such as FFHQ and LSUN CAT, to evaluate the effectiveness of the proposed mechanism. The results showed significant improvements in the quality and diversity of generated samples compared to baselines and other methods.

2.3. Claims and Evidence

Claim 1: Improved Stability of GAN Training The proposed mechanism improves the stability of GAN training compared to previous methods.

Evidence 1: Experimental results on the FFHQ dataset showed smoother and faster convergence compared to baselines and alternative methods (see Figure 1 of the paper) [1].

Claim 2: Robustness to Hyperparameters The proposed method is robust to the choice of hyperparameters.

Evidence 2: The augmentation probability remained stable and adjusted effectively during training, demonstrating the method’s robustness to hyperparameter choice (see Figure 5c of the paper) [1].

Claim 3: Superior Results with Limited Data The proposed method achieves superior results with limited data compared to baseline methods.

Evidence 3: The Fréchet inception distance (FID) scores consistently showed the superiority of the proposed method compared to baselines and balanced consistency regularization (bCR) methods (see Figure 7 of the paper) [1].

2.4. Critique and Discussion

The paper provides a thorough study of augmentations and their potential leakage, and the experimental results demonstrate how the adaptive nature of the augmentation strength improves model performance. The experimental results provide strong support for the effectiveness of the proposed method, outperforming alternatives, even with a limited number of training images. It is particularly noteworthy that this includes the CIFAR-10 benchmark.

However, a more detailed comparison with similar augmentation mechanisms discovered by other research groups would have strengthened the evaluation. Clarifying similarities and differences with these approaches would enhance understanding of the state-of-the-art in training GANs with limited data. The clear description of methodology, experimental configurations, and impact of hyperparameters indicate a robust approach, but future work could delve deeper into these aspects.

3. Review of Paper 2 (DeLiGAN)

3.1. Storyline

Introduction The paper ”DeLiGAN: Generative Adversarial Networks for Diverse and Limited Data” proposes the DeLiGAN framework, which reparameterizes the latent generative space as a mixture model. This modification enables the generation of diverse and realistic images using limited amounts of training data.

High-Level Motivation/Problem The motivation is to generate diverse images with limited training data, which is often challenging in real-world applications. Existing GAN-based approaches require large datasets to capture image modality diversity. The authors aim to address this limitation and enable the generation of diverse images with limited data, benefiting domains like object recognition, image synthesis, and remote sensing.

Prior Work on the Problem Previous research has focused on improving GAN performance using transfer learning, data augmentation, and regularization techniques. However, these approaches struggle with limited data and fail to capture image modality diversity effectively.

Research Gap The research gap lies in generating diverse images with limited data. Existing approaches either lack diversity or require impractical amounts of training data. This paper aims to fill this gap by proposing a novel framework.

Contributions The main contributions of the paper are:

1. The DeLiGAN framework, which reparameterizes the latent generative space as a mixture model to capture image modality diversity.
2. Demonstration of DeLiGAN’s effectiveness in generating diverse images across different modalities.
3. Introduction of a modified ”inception-score” to quantitatively measure intra-class diversity of generated samples.

3.2. Proposed Solution: DeLiGAN Framework

The proposed DeLiGAN framework reparameterizes the latent generative space as a mixture model to capture image modality diversity, even with limited training data.

Mathematically, the DeLiGAN framework can be represented as follows:

1. Latent Space Reparameterization: The latent space is represented as a mixture of Gaussians model:

$$p(z) = \sum_{i=1}^{N} \phi_i g(z|\mu_i, \Sigma_i)$$

   where N is the number of Gaussian components, ϕ_i are the mixture weights, and $g(z|\mu_i, \Sigma_i)$ represents the probability of the sample z in the i-th Gaussian distribution with mean μ_i and covariance matrix Σ_i.

2. Training Procedure: The generator and discriminator networks are trained adversarially in the standard GAN framework. The generator learns to generate realistic samples from a random input vector z sampled from the latent space, while the discriminator aims to distinguish between real and generated samples, providing feedback to the generator for improvement.

3. Diversity Enhancement: The reparameterization of the latent space as a mixture model allows the generator to capture the diverse modes of the image distribution. By incorporating multiple Gaussian components, the model can generate diverse and realistic images even with limited training data.

The DeLiGAN framework offers a flexible and effective approach to handle limited data scenarios while maintaining diversity in the generated samples.

3.3. Claims and Evidence

Claim 1: Improved Diversity in Generated Samples The DeLiGAN framework enables the generation of diverse samples even with limited training data.

Evidence 1: Experiments on different image modalities, including handwritten digits, photo objects, and hand-drawn sketches, demonstrated that DeLiGAN generates samples with higher diversity compared to baseline GAN models. Visual comparisons in Figure 5 of the paper support this claim [2].

Claim 2: Effective Generation with Limited Data DeLiGAN achieves effective generation of diverse samples even with limited training data.

Evidence 2: Experiments on the CIFAR-10 dataset, with limited data, showed that DeLiGAN achieved higher modified ”inception-score” values compared to baseline GAN models, indicating effective generation of diverse samples (see Table 1 of the paper) [2].

Claim 3: Stable Training Process The DeLiGAN framework stabilizes the training process compared to baseline GAN models.

Evidence 3: Visual comparisons in Figure 6 of the paper [2] demonstrate that DeLiGAN's training process is more stable and leads to improved convergence compared to baseline GAN models.

3.4. Critique and Discussion

The DeLiGAN framework addresses the challenge of generating diverse images with limited data and demonstrates promising results. The paper presents a clear narrative, discussing motivation, prior work, research gap, and contributions.

The evidence provided, including visual comparisons and modified "inception-score" values, supports the claims made by the authors. Experiments on different datasets and limited data scenarios provide robust empirical evidence for the effectiveness of the DeLiGAN framework.

However, a more detailed comparison with related augmentation mechanisms from other research papers would have strengthened the evaluation. Exploring similarities and differences with these approaches could enhance understanding of the state-of-the-art in generating diverse images with limited data.

Overall, the paper presents a well-structured and well-supported research proposal. The DeLiGAN framework shows promise in addressing the challenges of generating diverse images with limited data, and the evidence presented validates its effectiveness.

4. Review of Paper 3 (StyleCLIP)

4.1. Storyline

Introduction The paper "StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery" introduces a method for text-driven image manipulation using StyleGAN and CLIP. It addresses the challenge of discovering meaningful latent manipulations in StyleGAN by leveraging the joint language-image representation learned by CLIP. The proposed method allows users to manipulate various visual attributes of images by providing text prompts.

High-Level Motivation/Problem The motivation is to enable intuitive and versatile image manipulation techniques without manual effort or large annotated datasets. Existing methods for semantic control discovery often require manual examination or domain-specific data. The authors aim to fill this research gap by leveraging the CLIP model's ability to understand visual concepts expressed in natural language, enabling more intuitive and efficient text-driven image manipulations.

Prior Work on the Problem Previous research has explored text-guided image generation and manipulation using conditional GANs, attention mechanisms, and additional supervision. However, these methods have limitations in control and training requirements. The proposed StyleCLIP approach builds upon CLIP's language-image representation to provide a more versatile and intuitive text-driven manipulation method.

Research Gap The research gap lies in combining the generative power of StyleGAN with the language-image representations learned by CLIP. Previous methods lack fine-grained control or require manual effort or annotated data. The StyleCLIP approach fills this gap by leveraging the strengths of StyleGAN and CLIP to enable semantic image manipulations using text prompts.

Contributions The main contributions of the paper are:

1. A text-guided optimization scheme using CLIP to modify the latent vector of a StyleGAN image, enabling versatile text-driven image manipulations.

2. A latent mapper that infers a manipulation step in the latent space of StyleGAN, providing faster and more stable text-driven image manipulations.

3. A method for mapping text prompts into input-agnostic directions in StyleGAN's style space, enabling interactive and fine-grained text-driven image manipulations.

4.2. Proposed Solution: StyleCLIP

The proposed StyleCLIP approach combines StyleGAN and CLIP for text-driven image manipulations. It offers three methods: latent optimization, latent mapper, and global directions.

4.2.1. Latent Optimization

Latent optimization modifies the latent vector of a StyleGAN image by minimizing a loss function. The loss function combines CLIP loss, L2 distance in latent space, and an identity loss. The optimization problem is solved using gradient descent.

4.2.2. Latent Mapper

The latent mapper trains a mapping network to infer a manipulation step in the latent space of StyleGAN based on a text prompt. The mapping network consists of three fully-connected networks, each responsible for a different layer of StyleGAN.

4.2.3. Global Directions

Global directions map a text prompt into an input-agnostic direction in StyleGAN's style space. This is achieved by assessing the relevance of each style channel to the target attribute. The manipulation direction is determined based on a threshold parameter, allowing fine-grained control over manipulation strength and disentanglement.

4.3. Claims and Evidence

Claim 1: Versatile Image Manipulation StyleCLIP enables versatile text-driven image manipulations with fine-grained control over visual attributes.

Evidence 1: Visual examples in Figure 7 of the paper [3] demonstrate a wide range of semantic manipulations achieved using latent optimization, latent mapper, and global directions. These examples showcase the versatility and fine-grained control of StyleCLIP in responding to text prompts and producing visually coherent image manipulations.

Claim 2: Faster and More Stable Manipulations The latent mapper method in StyleCLIP allows for faster and more stable text-driven image manipulations compared to latent optimization.

Evidence 2: The similarity of manipulation directions inferred by the latent mapper for different input images demonstrates stability and consistency across inputs. This evidences the advantage of the latent mapper in providing faster and more stable image manipulations.

Claim 3: Fine-Grained and Disentangled Manipulations Global directions in StyleCLIP enable fine-grained and disentangled image manipulations by mapping text prompts into input-agnostic directions in StyleGAN's style space.

Evidence 3: Image manipulations along global text-driven manipulation directions demonstrate fine-grained changes in visual attributes while preserving other attributes. The degree of disentanglement is controlled by a threshold parameter, allowing users to achieve desired levels of manipulation strength and disentanglement (see Figure 6 of the original paper) [3].

4.4. Critique and Discussion

While addressing the challenge of discovering meaningful latent manipulations without extensive manual effort or annotated data, one potential limitation is the reliance on pretrained StyleGAN and CLIP models, which may limit manipulations to within the domain of the pretrained generator. Additionally, achieving drastic manipulations in visually diverse datasets may be challenging. Further comparisons with similar augmentation mechanisms proposed by other research groups could enhance the evaluation and understanding of the state-of-the-art in text-driven image manipulation.

5. Implementation

a) Implementation Motivation

While the current data augmentation techniques proposed by the original NVidia paper [1] have proven valuable in enhancing the performance of GANs on limited datasets, they come with certain weaknesses. These methods often generate augmented data that is too closely aligned with the original dataset, leading to limited diversification. Additionally, their effectiveness can vary, with certain augmentations proving less effective than others, as exemplified by the comparison between cutout and blitting in the original paper [1]. Moreover, there is a risk of introducing issues such as overly high p-values, affecting the reconstruction of the original dataset.

In addition, the proposed adaptive discriminator augmentation method has been predominantly limited to raw image datasets with no inherent correlation between samples. A potential avenue for improvement lies in extending these augmentation strategies to leverage temporal data sources. In numerous areas, such as medical scans captured at different times, or frames from a video, images are not isolated frames but rather sequences captured over time. This presents an opportunity to enhance model robustness by tapping into the temporal relationships within the original video sources by creating a new kind of ”temporal” augmentation, and measuring its success within the ADA framework proposed by NVidia [1].

b) Implementation Plan and Setup

What is a Temporal Augmentation:

Building upon the NVIDIA approach, where image augmentations are strategically employed to prevent overfitting without influencing generated images, temporal augmentation introduces a nuanced and non-agnostic element into the sequence of frames.

Similar to the existing augmentations outlined in the NVIDIA paper, which transform the source image through methods like color adjustments, rotation, pixel blitting, or cutouts, a temporal augmentation introduces a temporal traversal aspect. In this context, the augmentation involves skipping frames forward in the temporal sequence. The underlying concept is to generate a new frame that retains most of the qualities of the original frame but with enough differences to be meaningful (see Figure 1).

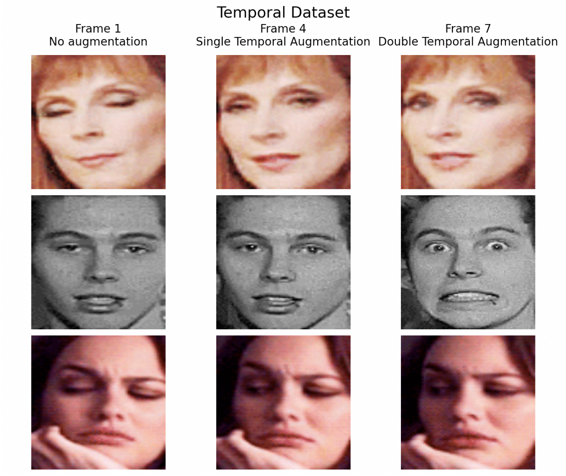


Figure 1. Temporal augmentations performed on 3 random images of the TGIF dataset. Demonstrates how each successive frame retains most of the qualities of the previous frame, but as the frames advance, the information in the face image changes due to movement or change in expression.

Significance of this Method:

This departure from agnosticism is pivotal. Unlike traditional agnostic augmentations, these temporal shifts are inherently tied to the specifics of the temporal sequence, breaking away from the reversibility characteristic observed in other augmentations. In contrast to non-agnostic augmentations like cutout, which removes information from the data, the temporal augmentation approach avoids such data reduction. Cutout, while effective, has the drawback of potentially discarding critical information, introducing a weakness in scenarios where retaining all original features is crucial.

The non-agnostic nature of temporal augmentation, while not reversible, preserves all original data, addressing the limitation of information loss present in cutout. This tradeoff could allow for enhanced model robustness in temporal scenarios, as the discriminator is compelled to discern both static features and temporal relationships within the data.

Implementation of this Method:

In my implementation, I introduced two distinct temporal augmentations, each designed to advance the frame by three frames. However, the combined effect of both augmentations skips the frame forward by six frames. I decided to set the combined probability for any temporal augmentation to be 0.8, as this was shown to be the augmentation probability value that resulted in the minimum FID scores for the FFHQ-2k and FFHQ-10k datasets in the NVidia paper [1], and since I am trying to generate human faces with datasets around those sizes, I determined this number to be reasonable for my implementation.

To achieve an overall temporal augmentation probability of 0.8, both individual probabilities for a 3-frame augmentation (P_1 and P_2) are set to 0.553. This ensures that the combined probability ($P_{\text{combined}} = P_1 \times P_2 = 0.306$) and the sum of individual probabilities align with the desired overall probability. Consequently, with a 0.306 probability, the augmentation skips six frames, and with a 0.553 probability, it skips three frames. This approach collectively maintains the intended 0.8 probability of a temporal augmentation during training.

Crucially, these temporal shifts don't occur within the dataset itself, as outlined in the original paper, but rather before the data is passed into the discriminator, much like the methodology explained in the NVIDIA paper for standard image augmentations. However, there is also a slight weakness of the method: unlike other augmentation methods, since this relies solely on the original data to be performed, it cannot be applied to images created by the generator. This could potentially limit the efficacy of the method, which will be noted in the results.

Dataset:

For the temporal dataset, I curated a collection of 4800 frames of video data featuring human faces extracted from GIFs sourced from the Tumblr GIF (TGIF) dataset (https://raingo.github.io/TGIF-Release/), a collaborative effort by Yahoo Research and the University of Rochester. Specifically, there are 1600 unique GIFs, from which I processed the 1st, 4th, and 7th frames in order to avoid situations where the GIFs are "slow-paced", or contain immediate frames where there is no noticeable difference from frame to frame.

To identify facial features within these frames, I employed the MTCNN (Multi-task Cascaded Convolutional Networks) algorithm originally developed by David Sandberg. Then, using the OpenCV library, I focused on the detected faces, cropped them to isolate the facial regions, and resized the resulting images to a standardized 256x256 resolution. This preprocessing ensures that the dataset comprises high-quality facial images ready for integration into the GAN training process (see Figure 2).

In addition, each processed image is labeled according to its originating frame within the GIF sequence. This labeling becomes a crucial aspect in the subsequent GAN code implementation, allowing for the incorporation of temporal information during training.

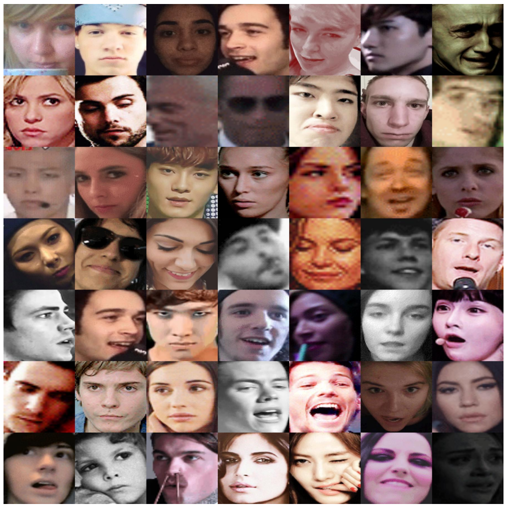


Figure 2. Random examples of pre-processed 256x256 images taken from the temporal dataset, sourced from the 1st frames of GIFs originally in the TGIF dataset.

Pre-trained Model:

To streamline reproducibility, I will start with a semi-pretrained LSUN Cats 256x256 model. This decision is motivated by the desire to facilitate result replication while accommodating time constraints. Using a model for cats ensures that the model's characteristics are not already biased towards human faces, which would otherwise make it difficult to evaluate the success of the new augmentation methods. While starting from an object dataset like CIFAR would not be suitable due to the dissimilarity between object characteristics and faces, leveraging a cat model provides a balance—sharing fundamental features with human faces while still presenting distinctive qualities.

Experiments:

I will conduct three unqiue experiments to help assess the impact of temporal augmentations on training a StyleGAN2 ADA model. The Transfer Learning Base for each experiment will be the LSUN Cats 256x256 dataset, and the models will be trained for 200kimg on the preprocessed and filtered Tumblr GIF (TGIF) dataset. Under ideal circumstances, this number would be around 10000kimg+, however, due to time constraints, it has been reduced.

All three experiment types will be performed on two dataset sizes: 1600 unique GIFs and 5000 unique GIFs, resulting in a total of six experimental runs.

The three experiment types are as follows:

Control Set 1 – Standard Augmentations applied for ADA (Blit + Geom)

Dataset: The first frame of each GIF in the dataset.

In this experiment, the augmentation method that will be used involves the application of standard pixel blitting and geometric augmentations during the training process. "Blit" augmentations refer to pixel-level transformations such as flipping, rotation, and translation. On the other hand, "geom" augmentations encompass general geometric transformations like scaling, rotation, anisotropic scaling, and fractional translation. These augmentations were identified as the two most successful types in the Nvidia paper and will serve as a benchmark for the other experiments. The dataset composition comprises one face frame from each unique GIF.

Control Set 2 – Standard Augmentations applied for ADA (Blit + Geom)

Dataset: The first **three** frames of each GIF in the dataset.

This mirrors Control Set 1, except that the temporal augmentations are applied directly to the dataset before training, as opposed to during the training process. Consequently, the model will be exposed to 3x the data from the start. This experiment is necessary to compare to the experimental set to ensure that any advantages gained from temporal augmentation are not merely a result of increased initial data exposure but instead stem from the adaptive augmentation mechanism itself.

Experimental Set – Temporal and Standard Augmentations applied for ADA (Temporal + Blit + Geom)

Dataset: The first frame of each GIF in the dataset.

The augmentation method will involve simultaneously applying temporal, pixel blitting, and geometric augmentations during the training process. Similar to Control Set 1, the dataset will be restricted to the first frame of each GIF. Performance improvements will be compared to Control Set 1 to determine if adding the temporal augmentation alongside the standard augmentations enhances the model's abilities, rather than solely benefiting from an increase in the quantity of data like Control Set 2.

Evaluation Metrics:

I will measure the results of each experiment using Fréchet Inception Distance (FID) scores.

The FID measures the similarity between the generated samples and the target dataset in terms of both the distribution of the samples and the perceptual similarity. The score can be calculated using the formula below, where a lower FID score indicates more similarity:

$$\text{FID} = ||\mu_{\text{real}} - \mu_{\text{fake}}||^2 + \text{Tr}(C_{\text{real}} + C_{\text{fake}} - 2(C_{\text{real}} \cdot C_{\text{fake}})^{0.5})$$

In addition, the Generator and Discriminator testing losses were also monitored and recorded throughout the experiments, which represent the respective performance metrics indicating how well the generator is synthesizing realistic samples and how effectively the discriminator is distinguishing between real and generated data during the evaluation phase.

Priority of Implementation Efforts:

1. Reproducing baseline experiments with the control set featuring standard augmentations (Control Set 1).
2. Generating the temporal dataset with the correct frame labels.
3. Evaluating the control set with temporal augmentations applied directly to the starting data (Control Set 2).
4. Implementing temporal augmentation capabilities within the existing GAN framework.
5. Executing experiments with the simultaneous application of temporal and standard augmentations (Experimental Set).
6. Analyzing and comparing results using FID scores to draw conclusions about the effectiveness of temporal augmentations in enhancing the training of StyleGAN2 ADA on limited datasets.

c) Implementation Details

Code Base:

The implementation will be based on the StyleGAN2 ADA Pytorch implementation by Nvidia Labs, with optimizations for Google Colab by GitHub user @dvshultz. This codebase provides a solid foundation for training and evaluating GAN models based on the paper's ADA method.

Temporal Augmentation Implementation:

Temporal augmentations are not inherently supported in the existing GAN framework, as they require additional information beyond the original image. To address this limitation, I developed new code to enable the GAN to account for the temporal component of images. This was achieved by passing the image ID and frame number through the image label during training. By reading the image label, the model could identify the specific image being processed and retrieve the next frame's image data from the dataset. This functionality was seamlessly integrated into the model's code, with the flexibility to toggle it on and off using a configuration flag.

Dataset Implementation:

The implementation of the temporal augmentation code, along with the dataset generation code, was done from scratch. The dataset generation involved downloading GIFs, extracting individual frames, and utilizing the MTCNN (Multi-task Cascaded Convolutional Networks) algorithm from the mtcnn Python package. This implementation of MTCNN is credited to David Sandberg, known for his contribution to FaceNet's MTCNN. The OpenCV library was then used to detect faces, crop around them, and resize the resulting images to 256x256.

FID Score Implementation:

The FID (Fréchet Inception Distance) scores, crucial for evaluating the model's performance, were calculated using the fid-score pip package developed by Rayyan Akhtar. This alternative was adopted due to challenges faced with the metric calculator provided by Nvidia, ensuring accurate and reliable evaluation of the generated images.

System Requirements:

The custom version of the StyleGAN2 code was ran on Google Colab using Pytorch 1.9.0+cu111 and torchvision 0.10.0+cu111 on a Tesla V100-SXM2-16GB GPU.

d) Implementation Details

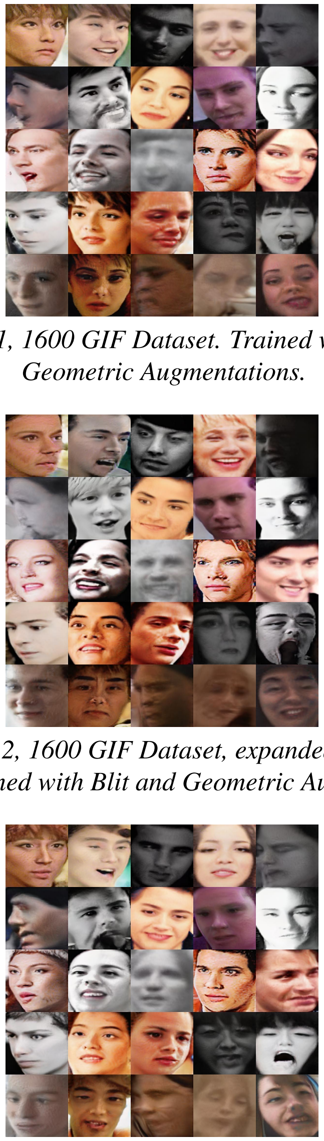
FID Scores:

Table 1. FID Scores for Different Experiment Sets

Experiment	**1600 GIFs**	**5000 GIFs**
Control Set 1	44.25	32.69
Control Set 2	50.04	35.77
Experimental Set	**40.15**	**31.64**

The FID scores in Table 1 reveal that the experiments incorporating temporal augmentations yielded slightly lower FID scores than the initial experiment without them. This suggests that introducing temporal dynamics contributes to a subtle yet discernible improvement in image diversity, as indicated by the FID metric. Conversely, the second experiment, relying solely on data augmentation, displayed significantly worse performance, potentially pointing to an overfitting scenario. The elevated test loss further supports this interpretation, indicating that augmenting the data without the adaptive augmentation mechanism led to a less robust model. The relationship between these scores also remained consistent between the 1600 GIF and 5000 GIF dataset sizes, indicating that the temporal augmentation had the same positive effect across different limited dataset sizes.

Image Output:



Control Set 1, 1600 GIF Dataset. Trained with Blit and Geometric Augmentations.

Control Set 2, 1600 GIF Dataset, expanded with extra frames. Trained with Blit and Geometric Augmentations.

Experimental Set, 1600 GIF Dataset. Trained with Temporal, Blit and Geometric Augmentations.

Figure 3. Sample generated images from each experiment to visually demonstrate the impact of temporal augmentations on image diversity.

While the generated images from all three experiments as shown in Figure 3 appear nearly identical to the untrained eye, the subtle improvement in FID scores with temporal augmentations suggests that these changes, while not visually striking, have a positive impact on the underlying data distribution. The nuanced quality enhancement, although not immediately perceivable, aligns with the goal of GANs to generate realistic and diverse images.

Loss Values:

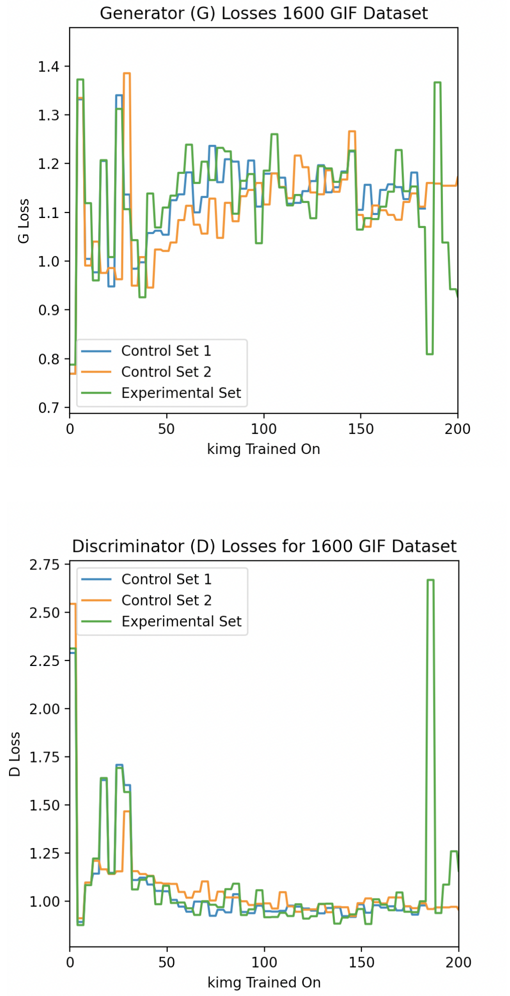


Figure 4. The generator and discriminator losses over the 200kimg training iterations for each experiment on the 1600 GIFs dataset.

Analyzing the loss values of the generator and discriminator (Figure 4) shed light on some the GAN's training dynamics. The initial increase in losses during the early iterations is a typical behavior when transitioning from a pre-trained model on a different dataset (LSUN Cats) to the target dataset (human faces). The subsequent decline in losses, especially in the discriminator model, indicates that the model adapted to the new dataset, but there was still ample room for improvement. The consistent loss post-initial iterations in the generator model hints at the potential for further refinement with an extended training period. Furthermore, there was no significant difference in loss value behavior between experiments.

Expectations and Hypotheses:

The overall alignment of results with expectations reinforces the understanding that the small dataset size and limited training iterations led to high FID scores. While the generator may not directly benefit from added augmentation, the discernible improvement in FID scores with temporal augmentations emphasizes their potential in increasing image diversity, even within the constraints of the limited dataset. The reason why the FID scores were likely not any lower is most likely due to the fact that the generator itself cannot benefit from the added augmentation, and it is only applied to the real images right before they're passed into the discriminator. However, the slightly lower FID score compared to Control Set 1 is still an indication that having the augmentation applied with a certain probability to the real images still helped increase image diversity, and was indeed a better outcome than feeding in all the frames at once as part of the original dataset itself (Control Set 2). Furthermore, the decline in losses post-initial iterations supports the hypothesis that continued training could lead to a more refined and robust model, with even better FID scores.

Discussion

These results, overall, indicate potential in exploiting the temporal nature of a dataset to enhance the performance of GAN models via the implementation of temporal augmentation within an existing adaptive augmentation framework.

This approach could be further beneficial in a different type of GAN architecture, such as the bCR used as a reference in the Nvidia paper [1]. This is because this approach only needs to perform the augmentation once during a single back-and-forth iteration between the generator and discriminator, and it also allows for comparison between the unaugmented and augmented real images. In the context of the temporal augmentation, this might not be perceived as a weakness, considering that the augmented image remains part of the valid data distribution, even if augmented data leaks into the generated data. Although bCR may greatly benefit from it, it's seen as an obsolete technique, which is why I instead pursued it using ADA.

Thus, this experimental implementation suggests that the simultaneous application of temporal and standard augmentations could potentially enhance the utilization of limited datasets in GANs. The broader implications may extend to leveraging the temporal nature of video data to improve the quality of image datasets for training AI models *in general*. However, it is important to note that further validation is required to substantiate this hypothesis.

Conclusion

In my implementation, I introduced a novel approach for leveraging temporal data augmentation as a method to enhance the performance of GAN models, primarily when working with limited datasets. This technique, building upon the widely accepted NVidia ADA framework, marks a change from traditional data augmentation methods and opens up new avenues for further exploration and refinement.

Interestingly, the results indicated that temporal augmentation could lead to a subtle yet discernible improvement in image diversity, as evidenced by the lower FID scores. This finding highlights the potential of such an approach in increasing the range and versatility of GAN-based AI models, particularly in scenarios involving temporal data sources such as video frames. In the medical field, temporal augmentation could improve the analysis of sequential medical scans, resulting in earlier disease detection and better patient outcomes. For autonomous vehicles, it could enhance the generation and interpretation of temporal data, such as traffic changes or weather conditions, leading to more reliable self-driving systems.

However, the trade-off of this method is its non-agnostic nature. Unlike typical augmentations, temporal shifts are tied to the specifics of the temporal sequence and lose the reversibility characteristic seen in other augmentations. This presents unique challenges that future work could seek to address, exploring ways to enhance the reversibility of temporal augmentations while maintaining their potential benefits.

Furthermore, while the temporal augmentation method showed promise, it may be relatively constrained when used in conjunction with the ADA framework. The technique might hold greater potential when used with other GAN architectures, like binary cross-entropy with reject (bCR). Future research could explore this avenue, evaluating the effectiveness of temporal augmentation in differing GAN architectures.

Moreover, an interesting extension of this study would be to test the temporal augmentation technique on a larger scale. The current implementation was somewhat limited by dataset size and training iterations, which might have restricted the ultimate potential of the technique. More extensive research with larger datasets and increased training iterations could provide a more comprehensive understanding of the technique's strengths and weaknesses, as well as its broader applicability.

In conclusion, the temporal augmentation technique proposed in this study represents an exciting idea in the field of GANs. The ability to leverage inherent temporal information in datasets to enhance model performance suggests potential for the development of more sophisticated and robust machine learning models. However, as with all new techniques, further research is necessary to fully understand its implications, adapt to its unique challenges, and maximally exploit its potential benefits.

References

[1] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, Timo Aila, ”Training Generative Adversarial Networks with Limited Data,” *NeurIPS 2020*. https://proceedings.neurips.cc/paper_files/paper/2020/file/8d30aa96e72440759f74bd2306c1fa3d-Paper.pdf

[2] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, R. Venkatesh Babu, ”DeLiGAN: Generative Adversarial Networks for Diverse and Limited Data,” Video Analytics Lab, CDS, Indian Institute of Science, Bangalore, INDIA 560012. https://arxiv.org/pdf/1706.02071.pdf

[3] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, Dani Lischinski, ”StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery,” Hebrew University of Jerusalem, Tel-Aviv University, Adobe Research. https://openaccess.thecvf.com/content/ICCV2021/papers/Patashnik_StyleCLIP_Text-Driven_Manipulation_of_StyleGAN_Imagery_ICCV_2021_paper.pdf