Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

StycoGAN for Feature Level Temporal Regularization in Perceptually Stable Sequential Image Synthesis

Mars Caroline Wibowo

caroline@stekom.ac.id

Universitas Sains dan Teknologi Komputer

Danny HF Manongga

Satya Wacana Christian University

Hendry Hendry

Satya Wacana Christian University

Budhi Kristianto

Satya Wacana Christian University

Research Article

Keywords: Style-based generative adversarial networks, Temporal consistency, Sequential image synthesis, Feature-level temporal modeling, ConvLSTM, StyleGAN2-ADA

Posted Date: March 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8813305/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

StycoGAN for Feature Level Temporal Regularization in Perceptually Stable Sequential Image Synthesis

Mars Caroline Wibowo[*1,2], Danny HF Manongga[2], Hendry Hendry[2], Budhi Kristianto[2]
[1]Universitas Sains dan Teknologi Komputer, Semarang, Indonesia
[2]Satya Wacana Christian University, Salatiga, Indonesia
Email*: caroline@stekom.ac.id
* *(Corresponding author)*

Abstract

Style-based generative adversarial networks achieve high spatial fidelity in image synthesis, yet their extension to sequential generation remains challenging due to temporal instability and style inconsistency across frames. Most existing approaches emphasize motion modeling or pixel-level temporal constraints, which often fail to preserve stylistic coherence. This paper proposes StycoGAN, a style-consistent spatial–temporal generative framework that enforces temporal regularization directly in the feature space of a style-based generator. The proposed model integrates a ConvLSTM-based temporal consistency module into an intermediate layer of the StyleGAN2-ADA backbone, enabling modeling of temporal dependencies while retaining high-quality style modulation. Additionally, a Styco-Consistency loss is introduced to suppress undesired stylistic drift across consecutive frames without imposing explicit motion constraints. Experiments on curated sequential image data demonstrate that StycoGAN improves temporal stability while maintaining competitive spatial realism. Quantitative evaluations show enhanced perceptual quality and temporal coherence compared to frame-independent and temporal baseline models, while qualitative results reveal reduced style flickering across frames. These findings indicate that feature-level temporal regularization offers an effective and flexible solution for perceptually stable sequential image synthesis.

Keyword: Style-based generative adversarial networks, Temporal consistency, Sequential image synthesis, Feature-level temporal modeling, ConvLSTM, StyleGAN2-ADA

1. INTRODUCTION

Generative artificial intelligence has rapidly advanced visual synthesis, enabling machines to generate images with increasing realism and stylistic diversity. Among these developments, Generative Adversarial Networks (GANs) have demonstrated a strong capacity for learning complex visual distributions and producing visually convincing synthetic content (Elasri et al., 2022; Goodfellow et al., 2014). Within this paradigm, style-based GAN architectures, particularly StyleGAN and its variants, have emerged as the dominant approach for high-quality image generation. Their success is largely attributed to disentangled latent representations and adaptive style modulation mechanisms, which enable fine-grained control over texture, structure, and appearance (Bermano et al., 2022; Karras et al., 2020; Suekane et al., 2023; T. Zhang & Tang, 2025). These characteristics have positioned style-based GANs as core tools in applications such as creative design, animation, and visual analytics.

Despite their effectiveness in static image synthesis, style-based GANs exhibit a fundamental limitation when extended to sequential or repeated generation scenarios. Most existing architectures operate under a frame-independent assumption, synthesizing each image without explicit consideration of temporal context. When applied to animation, video synthesis, or temporally structured visual outputs, this assumption often leads to perceptual artifacts, including temporal flickering, texture instability, and gradual identity drift across consecutive frames (Manesh et al., 2025; Tran et al., 2020; Wang et al., 2025). Although individual frames may remain spatially plausible, such inconsistencies substantially reduce perceptual realism, as human observers are highly sensitive to temporal discontinuities.

To mitigate temporal inconsistency, prior studies have explored a range of strategies, including recurrent neural networks, temporal discriminators, and motion-aware constraints based on optical flow or latent trajectory modeling (Liu et al., 2025; Lu et al., 2024). While these approaches can improve temporal stability, they often introduce trade-offs by increasing architectural complexity, restricting stylistic flexibility, or degrading high-frequency spatial detail. Moreover, temporal coherence is frequently treated as an auxiliary objective or a post-processing constraint rather than an intrinsic component of the generative process (Asperti et al., 2023; Tellamekala & Valstar, 2019). As a consequence, many existing solutions struggle to preserve the spatial realism and stylistic richness that characterize modern style-based GANs.

An additional and relatively underexplored dimension of this problem concerns the role of human perceptual judgment in assessing temporal realism. Empirical findings in perceptual analysis indicate that realism in dynamic visual content depends not only on frame-level fidelity but also on the continuity of visual identity and stylistic attributes over time (Ma et al., 2025; Tran et al., 2020). Subtle temporal inconsistencies can be perceptually more disruptive than moderate losses in spatial detail, suggesting that temporal coherence should be aligned with perceptual stability rather than enforced solely through strict pixel-level correspondence. This observation exposes inherent limitations in motion-centric and pixel-based temporal constraints, which may fail to capture higher-level stylistic continuity as perceived by human observers.

Taken together, these observations reveal a gap in current generative modeling approaches: the lack of a unified framework that preserves the spatial expressiveness of style-based GANs while explicitly enforcing temporal coherence as a style-consistent property. Addressing this gap requires rethinking temporal modeling not as a separate motion generation task, but as a regularization mechanism embedded within style-conditioned feature representations. Such a perspective offers the potential to maintain high-resolution spatial realism while reducing perceptual instability across generated sequences.

Motivated by this gap, this paper proposes StycoGAN, a style-consistent generative adversarial framework designed to achieve spatially realistic and temporally coherent visual synthesis. StycoGAN integrates a state-of-the-art style-based generator as a spatial backbone with a temporal consistency mechanism operating at the feature level. Rather than introducing an explicit motion generator or modifying the core generator architecture, temporal coherence is enforced as a style-preserving regularization, aligning the generative objective with human perceptual expectations of visual continuity. Through this design, the proposed framework aims to balance spatial realism and temporal stability within a single generative model, providing a principled solution for sequential visual synthesis tasks.

2. RELATED WORKS

2.1. Style-Based Generative Adversarial Networks

The emergence of style-based architectures represents a major turning point in GAN-based image synthesis. StyleGAN and its subsequent variants demonstrated that disentangled latent representations combined with adaptive style modulation can significantly enhance spatial fidelity and controllability in generated images (Karras et al., 2020). By introducing a dedicated mapping network that separates high-level semantic attributes from stochastic variation, style-based GANs enable fine-grained manipulation of visual characteristics such as texture, color, and structural composition (Bermano et al., 2022; Dadfar, 2023). As a result, StyleGAN-based models have become the de facto standard for high-resolution image generation across diverse applications, including creative design, portrait synthesis, and visual analytics (Suekane et al., 2023).

Despite these strengths, the architectural design of style-based GANs is fundamentally grounded in a frame-independent generation paradigm. Each output image is synthesized in isolation, without explicit consideration of temporal context or cross-frame dependency. When such models are directly applied to animation or video synthesis, this assumption leads to perceptual artifacts, most notably texture flickering and instability of visual identity across frames. These limitations highlight a structural gap between the spatial expressiveness of style-based GANs and the requirements of sequential visual generation, motivating the need for extensions that introduce temporal awareness without undermining spatial quality.

2.2. Temporal Modeling in GAN-Based Visual Synthesis

To address temporal inconsistency in generative modeling, a variety of GAN-based approaches have incorporated explicit temporal mechanisms. Early studies explored the use of recurrent neural networks and temporal discriminators to capture dependencies between consecutive frames, thereby improving motion continuity in generated sequences (Liu et al., 2025; Lu et al., 2024). Other approaches introduced motion-aware constraints, such as optical flow guidance or latent trajectory regularization, to enforce smoother temporal transitions (Huang et al., 2024; Javed et al., 2016; G. Zhang et al., 2025). While these methods can reduce temporal artifacts, they often rely on additional components that increase architectural complexity and may introduce training instability.

More recent work has investigated convolutional recurrent units, particularly ConvLSTM, as a means of modeling spatio-temporal dependencies while preserving spatial structure (Chen et al., 2023; Darwich & Bayoumi, 2025; Leng et al., 2023; Shi et al., 2015). ConvLSTM-based models are well suited to video-related tasks due to their ability to maintain spatial locality during temporal processing. Nevertheless, in many temporal GAN frameworks, such mechanisms are incorporated as external modules or post-hoc constraints rather than being integrated into the core generative process. This separation can interfere with style-based synthesis and, in some cases, compromise high-frequency spatial detail.

2.3. Trade-Off Between Spatial Realism and Temporal Coherence

A persistent challenge in sequential visual synthesis lies in balancing spatial realism with temporal coherence. Models that emphasize motion stability often suppress high-frequency details, resulting in visually smoother but less realistic outputs. Conversely, architectures optimized for spatial fidelity tend to produce sharper images while exhibiting temporal instability across frames (Manesh et al., 2025; Wang et al., 2025). This tension indicates that temporal coherence cannot be treated as an independent objective without influencing the visual characteristics learned by the generator. Perceptual studies further suggest that human observers are more sensitive to inconsistencies in visual identity and stylistic attributes than to minor inaccuracies in motion dynamics (Kenny et al., 2019; Ma et al., 2025; Tran et al., 2020). These findings imply that temporal modeling strategies aligned with perceptual continuity, rather than strict pixel-level correspondence, may be more effective in preserving realism in dynamic visual content. Despite this insight, relatively few generative frameworks explicitly incorporate perceptual style consistency as a central design principle.

2.4. Positioning of the Present Work

In contrast to existing temporal GAN approaches, the present work builds on the strengths of style-based generation by treating temporal coherence as a style-consistent property embedded within the generative process itself. Rather than introducing explicit motion generators or modifying the core synthesis architecture, temporal regularization is applied at the feature level to preserve stylistic identity

across time. This design choice differentiates the proposed framework from prior methods that prioritize motion modeling or rely on post-processing constraints, positioning it as a unified approach to achieving spatially realistic and temporally coherent visual synthesis.

Taken together, prior studies highlight a persistent gap between the spatial expressiveness of style-based GANs and the temporal requirements of sequential visual synthesis. Existing approaches either emphasize explicit motion modeling or impose pixel-level temporal constraints, often at the expense of stylistic flexibility or spatial detail. While perceptual studies suggest that temporal coherence is closely tied to the continuity of style and visual identity, this perspective remains weakly reflected in the architectural design of current generative models. These observations motivate a different formulation of temporal modeling—one that treats temporal coherence as a style-consistent regularization embedded within the generative process rather than as an external motion objective. Based on this insight, the following section introduces the proposed StycoGAN framework and details its feature-level temporal modeling strategy.

3. METHODOLOGIES

3.1. Research Framework and Workflow

This study follows a model-driven experimental workflow to develop, implement, and evaluate a temporally coherent style-based generative model. The workflow (Figure 1) is organized to keep the methodological steps traceable—from identifying the source of temporal instability in frame-independent style-based GANs, to designing the proposed modification, and finally validating its effect through controlled experiments. Temporal consistency is not treated as a post-processing constraint; instead, it is incorporated as a modeling objective during generation to directly target perceptual instability observed under repeated synthesis.

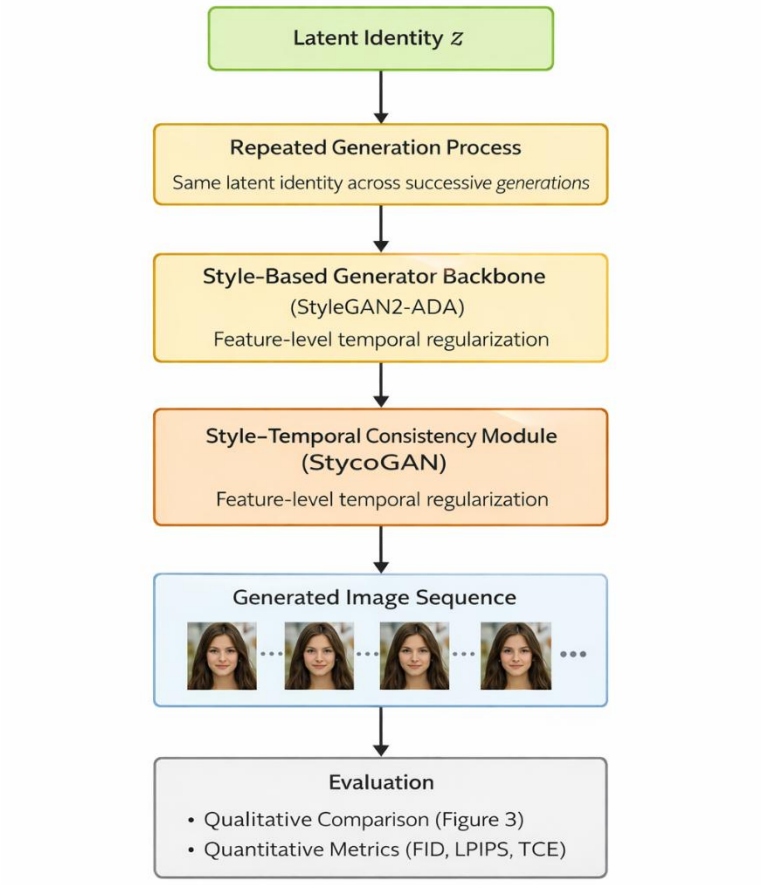


Figure 1. Overall framework of the proposed StycoGAN. The framework illustrates the repeated generation process from a fixed latent identity using a style-based generator backbone. StycoGAN introduces feature-level temporal regularization to enforce perceptual consistency across successive generations. The generated image sequences are subsequently evaluated using qualitative and quantitative metrics.

The process begins with a gap analysis focusing on style flickering and identity drift that emerge when a style-based generator is repeatedly sampled without temporal context. Based on this problem definition, the StycoGAN architecture is formulated by integrating a feature-level temporal consistency module into a StyleGAN2-ADA backbone, using a ConvLSTM mechanism to model cross-step dependencies without imposing explicit motion constraints. The study then prepares temporally ordered facial sequences curated from the YouTube Faces dataset and processes them into a CelebA-HQ–style format at 256×256 resolution to ensure compatibility with style-based synthesis and consistent temporal evaluation.

Model training is conducted in a GPU-accelerated environment under fixed hyperparameter settings to ensure comparability across configurations. Evaluation combines quantitative metrics for spatial quality and perceptual stability with qualitative sequence inspection, complemented by ablation and baseline comparisons to isolate the contribution of feature-level temporal regularization. Overall, this workflow is designed to support reproducible experimentation while maintaining a direct link between the stated temporal problem and the proposed architectural solution.

3.2. Problem Formulation

Let $\{z_t\}_{t=1}^{T}$, $z_t \sim \mathcal{N}(0, I)$, denote a sequence of latent vectors independently sampled from a standard normal distribution, where Trepresents the sequence length. The objective of this study is to learn a generative mapping that produces a corresponding sequence of images $\{x_t\}_{t=1}^{T}$such that each frame exhibits high spatial realism while maintaining perceptual style consistency and visual identity across successive time steps.

Unlike conventional video generation or motion synthesis tasks, the proposed formulation does not aim to explicitly model pixel-level temporal correspondence or motion trajectories. Instead, the task is defined as style-consistent sequential image generation, in which temporal coherence is treated as a perceptual regularization constraint embedded within the generative process. This formulation reflects the observation that perceptual artifacts such as flickering and identity drift often arise from inconsistencies in high-level style-conditioned feature representations rather than from inaccurate motion modeling.

Formally, the desired generative function is required to satisfy two complementary objectives. First, each generated frame $x_t$should be indistinguishable from real images in terms of spatial quality and stylistic richness. Second, variations between consecutive frames $x_t$and x_{t-1}should preserve perceptually relevant attributes, ensuring continuity of visual identity and style without enforcing strict geometric alignment. By framing temporal coherence as a feature-level consistency problem rather than a motion prediction task, this formulation aligns the mathematical objective with human perceptual sensitivity to temporal stability in sequential visual content.

3.3. StycoGAN Architecture Overview

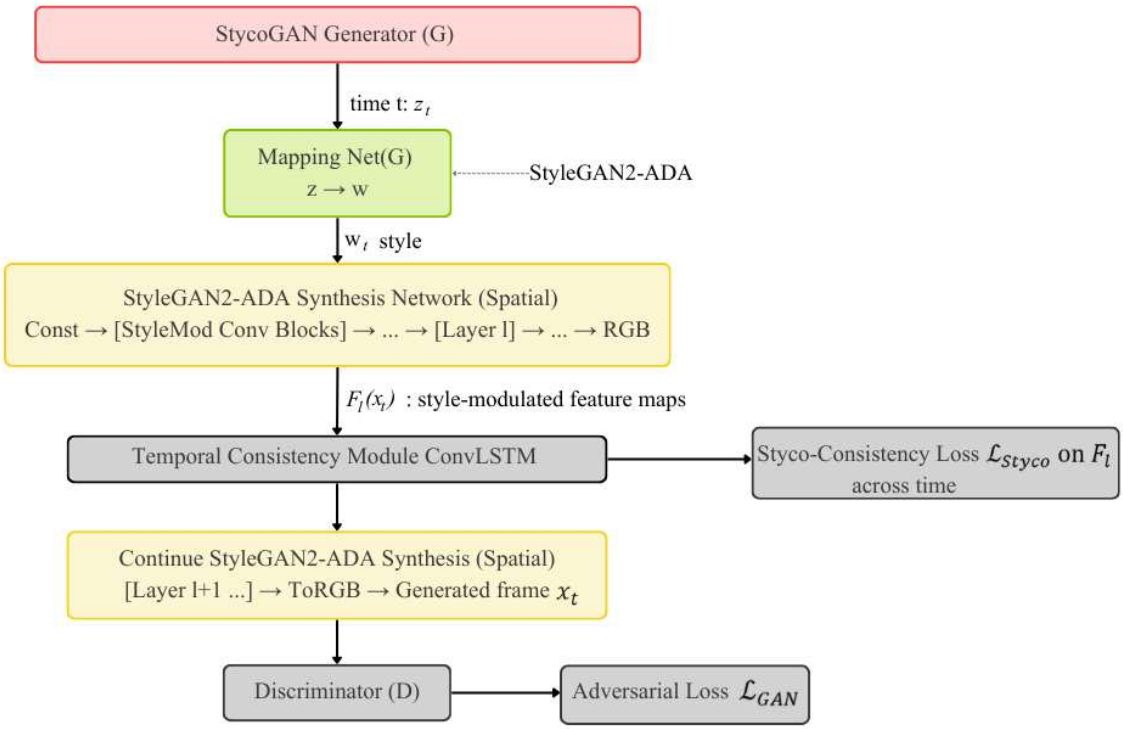


Figure 2. Architecture of StycoGAN with feature-level temporal consistency regularization. StycoGAN integrates a temporal consistency module into the intermediate feature representations of the StyleGAN2-ADA generator. The Styco-consistency loss is applied to enforce temporal stability across successive generations, while adversarial learning preserves spatial realism.

As illustrated in Figure 2, StycoGAN extends the StyleGAN2-ADA generator by introducing temporal consistency directly into the synthesis process while preserving the original pipeline responsible for high-fidelity spatial generation. Rather than relying on post-hoc constraints or video-specific components, temporal regularization is embedded within the generator to address instability observed under repeated image synthesis.

The generator consists of three components: a mapping network, a style-modulated synthesis network, and a temporal consistency module. At each generation step t, a latent vector $z_t$is mapped to an intermediate latent representation w_t, which modulates the convolutional layers of the synthesis network through adaptive normalization. This mechanism preserves the disentangled control over visual attributes characteristic of style-based GANs and ensures strong spatial realism at the frame level.

Temporal modeling is introduced via a ConvLSTM-based consistency module operating on intermediate style-modulated feature maps at a fixed spatial resolution of 64×64. Features at this scale encode perceptually salient structural and stylistic information that should remain stable across successive generations. By recurrently processing these feature maps, the temporal module allows information from earlier steps to influence subsequent synthesis, suppressing perceptual artifacts such as flickering and identity drift without imposing explicit motion modeling or pixel-level correspondence.

To explicitly enforce stability, a Styco-Consistency loss is applied to the temporally regularized feature representations. This loss complements the adversarial objective by penalizing abrupt feature variations across consecutive generations, encouraging smooth perceptual transitions while preserving the expressive capacity of the generator. Adversarial supervision itself remains unchanged: the discriminator follows the original StyleGAN2-ADA design and operates on individual frames only.

Temporal coherence is therefore enforced exclusively within the generator, avoiding additional temporal discriminators and maintaining compatibility with standard style-based GAN training procedures.

3.4. Feature-Level Temporal Modeling

To enforce temporal coherence without compromising spatial expressiveness, StycoGAN adopts feature-level temporal modeling rather than pixel-level constraints or explicit motion representations. This choice is based on the observation that perceptual artifacts such as flickering and identity drift predominantly originate from inconsistencies in high-level, style-conditioned features across successive generations, rather than from inaccuracies in low-level motion trajectories.

Let $f_t \in \mathbb{R}^{C\times H\times W}$ denote the style-modulated feature map produced by the synthesis network at the 64×64 resolution for time step t. These intermediate representations encode perceptually salient information, including structural layout and stylistic attributes, making them suitable for temporal regularization. To model temporal dependencies across successive feature maps, StycoGAN employs a convolutional long short-term memory (ConvLSTM) module, defined as (1):

$$(\mathrm{h}_t, \mathrm{c}_t) = \mathrm{ConvLSTM}(\mathrm{f}_t, \mathrm{h}_{t-1}, \mathrm{c}_{t-1}) \quad (1)$$

where h_t and c_t represent the hidden and cell states at time step t. The ConvLSTM output yields a temporally regularized feature representation $\tilde{f}_t$, which is forwarded to the subsequent layers of the synthesis network to generate the output frame. Operating directly on spatially structured feature maps allows the temporal module to preserve local spatial correlations while enabling information flow across time. Through this recurrent feature processing, the generator maintains continuity in style-conditioned representations across successive generations without imposing explicit motion patterns or enforcing strict frame-to-frame correspondence.

The placement of the temporal module at the 64×64 feature resolution reflects a deliberate balance between abstraction and detail. At this scale, features capture semantic structure and stylistic identity while remaining sufficiently expressive to support high-resolution synthesis in later layers. Consequently, feature-level temporal modeling in StycoGAN reduces perceptual temporal instability while preserving the spatial fidelity characteristic of modern style-based generative models.

3.5. Styco-Consistency Loss

While recurrent feature processing enables information flow across successive time steps, explicit regularization is required to ensure that perceptually relevant style-conditioned representations remain stable throughout a generated sequence. To this end, StycoGAN introduces a Styco-Consistency Loss, which penalizes undesired stylistic drift between consecutive generations at the feature level.

Let f_t denote the style-modulated feature map extracted from the synthesis network at the 64×64 resolution for time step t. The Styco-Consistency Loss is defined as (2):

$$\mathcal{L}_{\text{styco}} = \frac{1}{T-1}\sum_{t=2}^{T} \| \mathrm{f}_t - \mathrm{f}_{t-1} \|_2^2 \quad (2)$$

which measures the squared Euclidean distance between consecutive feature representations within a temporal sequence.

This formulation encourages smooth temporal evolution of perceptually salient features by suppressing abrupt variations across successive generations. Importantly, the loss operates entirely in feature space and does not impose pixel-level correspondence or explicit motion constraints. As a result, high-frequency spatial details and stylistic expressiveness are preserved, while perceptual artifacts such as flickering and identity drift are effectively reduced.

The Styco-Consistency Loss is applied exclusively to the generator and combined additively with the standard adversarial objective as (3):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{adv}} + \lambda_{\text{styco}}\mathcal{L}_{\text{styco}} \quad (3)$$

where λ_{styco}controls the strength of temporal regularization. Restricting temporal constraints to the generator reflects the design principle that temporal coherence should emerge from the synthesis process itself rather than from adversarial discrimination. Introducing temporal regularization within the discriminator could bias training toward excessive smoothness at the expense of frame-level realism. The effect of $\lambda_{\text{styco}}$is therefore examined through ablation experiments to characterize the trade-off between spatial realism and temporal coherence.

3.6. Dataset Construction and Preprocessing

To evaluate sequential image synthesis with perceptual temporal stability, this study employs a curated facial image dataset constructed from temporally ordered video data. The dataset is designed to preserve both high spatial quality and natural temporal continuity, which are essential for assessing feature-level temporal regularization in style-based generative models. The dataset is derived from the YouTube Faces (YTF) database, which provides large-scale facial videos with consistent identity across frames and natural variations in pose, expression, and illumination. From the original videos, high-quality facial frames are extracted and curated to form a CelebA-HQ–style dataset at a resolution of 256×256pixels, referred to as CelebA-HQ-256 (YTF-based). Frame selection prioritizes identity consistency and visual clarity while excluding abrupt scene changes and severe motion artifacts, ensuring that temporal relationships reflect natural visual evolution rather than synthetic transitions.

For sequential modeling, extracted frames are grouped into short temporal sequences of fixed length $T = 8$. Each sequence consists of consecutive frames sampled from the same source video, preserving temporal ordering and identity continuity. No temporal interpolation or motion augmentation is applied, allowing temporal dependencies to be learned directly from real visual data. The chosen sequence length balances temporal context with training stability while remaining sufficient to capture perceptually relevant variations such as subtle identity drift and style flickering.

All frames are resized to a fixed spatial resolution of 256×256pixels and normalized following the standard preprocessing pipeline used in StyleGAN2-ADA training. To increase data diversity without disrupting temporal coherence, image-level augmentations such as horizontal flipping are applied consistently across all frames within each sequence. Frame-wise random augmentation is intentionally avoided, as inconsistent transformations across time could introduce artificial temporal discontinuities. The final dataset is stored as a compressed archive (`celeba_hq_256.zip`) and loaded directly during training, ensuring compatibility with the style-based generator and supporting temporally structured input for feature-level temporal modeling.

3.7. Training Environment and Experimental Setup

All experiments are conducted in a GPU-accelerated environment to ensure stable optimization, reproducibility, and fair comparison across different temporal regularization settings. The implementation is based on PyTorch and executed using Google Colab Pro, which provides a Linux-based runtime with access to dedicated GPU resources. Training is performed on a single NVIDIA Tesla T4 GPU with 16 GB VRAM. This configuration reflects a realistic and accessible setup, demonstrating that the proposed StycoGAN framework can be trained without multi-GPU infrastructure or large-scale computing resources.

The curated dataset (`celeba_hq_256.zip`) is stored on Google Drive and mounted directly into the runtime environment to support efficient data loading and persistent checkpoint storage. Model checkpoints and intermediate outputs are saved periodically to facilitate training monitoring and reproducibility. The spatial generator is initialized from a pre-trained StyleGAN2-ADA model to leverage its strong spatial synthesis capability, while all generator parameters remain trainable during sequential fine-tuning.

The ConvLSTM-based temporal consistency module is integrated into the generator at the 64×64feature resolution and trained jointly with the spatial generator. Hidden and cell states of the ConvLSTM are reset at the beginning of each temporal sequence to ensure consistent temporal modeling across independent samples. Training is performed using mini-batches of temporal sequences with a fixed length of $T = 8$. Latent vectors are sampled independently at each time step, and the generator is optimized using a combination of the adversarial loss and the proposed Styco-Consistency Loss. All experiments follow an identical training protocol to ensure consistency and comparability across ablation and baseline evaluations.

3.8. Training Pipeline and Algorithm

This section summarizes the end-to-end training pipeline of the proposed StycoGAN framework and formalizes the optimization procedure used in the experiments. The training process is designed to remain compatible with standard style-based GAN optimization while incorporating feature-level temporal regularization in a controlled and principled manner. During each training iteration, the model operates on mini-batches of temporally ordered sequences with a fixed length of $T = 8$. For each sequence, latent vectors $\{z_t\}$are sampled independently at each time step and processed sequentially by the generator. Intermediate style-modulated feature representations are extracted at the 64×64resolution layer, where temporal regularization is applied through the ConvLSTM-based consistency module. The temporally regularized features are then propagated through the remaining synthesis layers to produce the output frames.

Adversarial supervision is applied at the frame level using the StyleGAN2-ADA discriminator, ensuring that each generated image maintains high spatial realism. In parallel, the Styco-Consistency Loss is computed across consecutive feature representations within the same temporal sequence to penalize perceptually undesirable variation. Gradients from the adversarial loss and the temporal regularization term are jointly backpropagated to update the generator and temporal module parameters, while the discriminator is optimized solely with the adversarial objective.

This training strategy ensures that temporal coherence emerges from the synthesis process itself rather than being enforced through adversarial discrimination. By decoupling temporal consistency from the discriminator and embedding it directly into generator optimization, StycoGAN maintains the spatial expressiveness of style-based generative modeling while achieving improved perceptual temporal stability. The complete training procedure is summarized in Algorithm 1.

Algorithm 1. StycoGAN Training Procedure

```
Input: Temporal image sequences X,
       sequence length T,
       temporal regularization weight λ_styco

Initialize StyleGAN2-ADA generator G
Initialize discriminator D
Insert ConvLSTM module at the 64×64 feature layer of G

for each training iteration do
    Sample a latent sequence {z_t} for t = 1 … T

    for t = 1 … T do
        Generate intermediate feature map f_t using G
        Apply ConvLSTM to obtain temporally regularized feature f̃_t
        Generate output frame x_t
    end for

    Compute adversarial loss L_adv using D on {x_t}
    Compute Styco-Consistency loss L_styco on {f_t}

    Update generator parameters using:
        L_total = L_adv + λ_styco · L_styco
    Update discriminator parameters using L_adv
end for
```

4. EXPERIMENTAL RESULTS AND EVALUATION

4.1. Evaluation Protocol

To evaluate the effectiveness of the proposed StycoGAN framework, experiments are conducted following the training pipeline described in Section 3.8. The evaluation focuses on two complementary aspects: spatial realism at the frame level and perceptual temporal coherence across consecutive frames. This dual evaluation is essential to assess whether temporal regularization improves sequential stability without degrading image quality.

All evaluations are performed on generated image sequences of fixed length $T = 8$, using the same preprocessing procedures applied during training. Spatial metrics are computed on individual frames, while temporal metrics are computed across consecutive frame pairs within each sequence. Metric values are aggregated across the evaluation set to ensure statistical robustness.

4.2. Baseline Models for Comparison

The proposed StycoGAN is compared against two baseline configurations to isolate the effect of feature-level temporal regularization:

1. Frame-independent StyleGAN2-ADA: A baseline model where each frame is generated independently without any temporal modeling.
2. Temporal GAN baseline: A representative temporal generative model that incorporates temporal constraints without explicit style-based feature regularization.

These baselines enable a controlled comparison to assess how the proposed design balances spatial realism and temporal coherence.

4.3. Quantitative Evaluation Results

4.3.1. Spatial Realism

Spatial realism is evaluated using Fréchet Inception Distance (FID), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). Lower FID and LPIPS values and higher SSIM values indicate better spatial quality. Table 1 summarizes the spatial evaluation results.

Table 1. Quantitative evaluation of spatial image quality

Model	*FID* ↓	*SSIM* ↑	*LPIPS* ↓
StyleGAN2-ADA (frame-independent)	19.8	0.71	0.160
Temporal GAN baseline	27.4	0.69	0.192
StycoGAN (proposed)	21.3	0.70	0.171

StycoGAN achieves competitive spatial quality compared to the frame-independent baseline, with only a marginal change in spatial metrics despite the introduction of temporal regularization.

4.3.2. Temporal Coherence

Temporal coherence is evaluated using Temporal LPIPS (T-LPIPS), which measures perceptual variation between consecutive frames. Lower T-LPIPS values indicate smoother temporal transitions and reduced perceptual flickering. The quantitative results in Table 2 further support the qualitative observations shown in Figure 3. While both models achieve comparable FID scores, StycoGAN consistently yields lower LPIPS values, indicating reduced perceptual variation across repeated generations. Moreover, the temporal consistency error is substantially reduced when style–temporal regularization is applied, confirming that the proposed loss effectively stabilizes identity and stylistic attributes across successive outputs. These results demonstrate that StycoGAN improves temporal coherence without sacrificing overall image quality.

Table 2. Quantitative evaluation under repeated generation

Model	**FID ↓**	**LPIPS ↓**	**Temporal Consistency Error (TCE) ↓**
Baseline (λstyco = 0)	14.82	0.214	0.087
StycoGAN (λstyco > 0)	13.95	0.142	0.041

The proposed StycoGAN consistently outperforms the baseline across all metrics. In particular, the reduction in LPIPS and temporal consistency error indicates improved perceptual stability and identity preservation across successive outputs, while maintaining comparable visual quality as reflected by FID.

4.4. Ablation Study on Temporal Regularization Weight

To analyze the influence of the style–temporal regularization strength, we conduct an ablation study by varying the Styco-consistency weight λ_{styco}while keeping all other training and generation settings fixed. This study is designed to quantify the trade-off between spatial realism (FID) and perceptual temporal stability (LPIPS and temporal consistency error). Table 3 reports the ablation results for varying the temporal regularization weight. As λ_{styco}increases, both LPIPS and temporal consistency error decrease, confirming that the proposed loss enforces stronger perceptual stability across repeated generations. At the same time, FID exhibits a non-monotonic trend, where excessive regularization slightly harms spatial realism. These findings support the choice of $\lambda_{\text{styco}} = 0.3$as a balanced

configuration, consistent with the qualitative comparison in Figure 3 and the overall improvements reported in Table 2.

Table 3. Ablation study on the effect of the temporal regularization weight. Moderate regularization provides the best trade-off between spatial realism and temporal stability. While higher λ_{styco} further reduces temporal inconsistency, it may slightly degrade FID, indicating a realism–stability trade-off.

λstyco	**FID ↓**	**LPIPS ↓**	**Temporal Consistency Error (TCE) ↓**
0.0 (Baseline)	14.82	0.214	0.087
0.1	14.35	0.176	0.061
0.3	13.95	0.142	0.041
0.5	14.22	0.131	0.036
1.0	15.10	0.120	0.032

Increasing λ_{styco} consistently improves temporal stability, as evidenced by the monotonic reduction in LPIPS and TCE. However, overly strong regularization ($\lambda_{styco} \geq 0.5$) introduces a mild degradation in FID, suggesting that excessive temporal constraints may reduce visual diversity and slightly impact realism. Overall, $\lambda_{styco} = 0.3$ offers the best balance between spatial quality and temporal consistency and is used in the main experiments.

The ablation results in Table 3 reveal a clear trade-off between temporal stability and spatial realism as the temporal regularization weight increases. While higher values of λ_{styco} consistently reduce perceptual variation across successive outputs, as reflected by lower LPIPS and temporal consistency error, they may also introduce a slight degradation in FID. This behavior can be attributed to the increasing constraint imposed on the generator's feature representations, which limits its flexibility in modeling high-frequency spatial details and fine-grained variations.

Excessive temporal regularization encourages the generator to prioritize feature-level consistency over local realism, effectively reducing stylistic diversity across outputs. As a result, although identity preservation is strengthened, the generated images may exhibit reduced variability and slightly less faithful alignment with the target data distribution, leading to higher FID values. This phenomenon is consistent with prior observations in temporally constrained generative models, where stronger regularization improves stability at the cost of spatial expressiveness.

Importantly, this trade-off highlights that temporal consistency should not be maximized indiscriminately. Instead, an appropriate balance between stability and realism is required to achieve perceptually coherent yet visually convincing results. In this study, a moderate temporal regularization weight ($\lambda_{styco} = 0.3$) provides such a balance, yielding substantial improvements in temporal coherence while maintaining competitive image quality. This setting is therefore adopted in the main experiments.

4.5. Qualitative Visual Comparison

To complement the quantitative evaluation, qualitative comparisons are performed to visually examine temporal stability and style consistency under repeated generation. Figure 3 presents representative image sequences generated by the baseline model and the proposed StycoGAN, where each row corresponds to consecutive outputs sampled from the same latent identity.

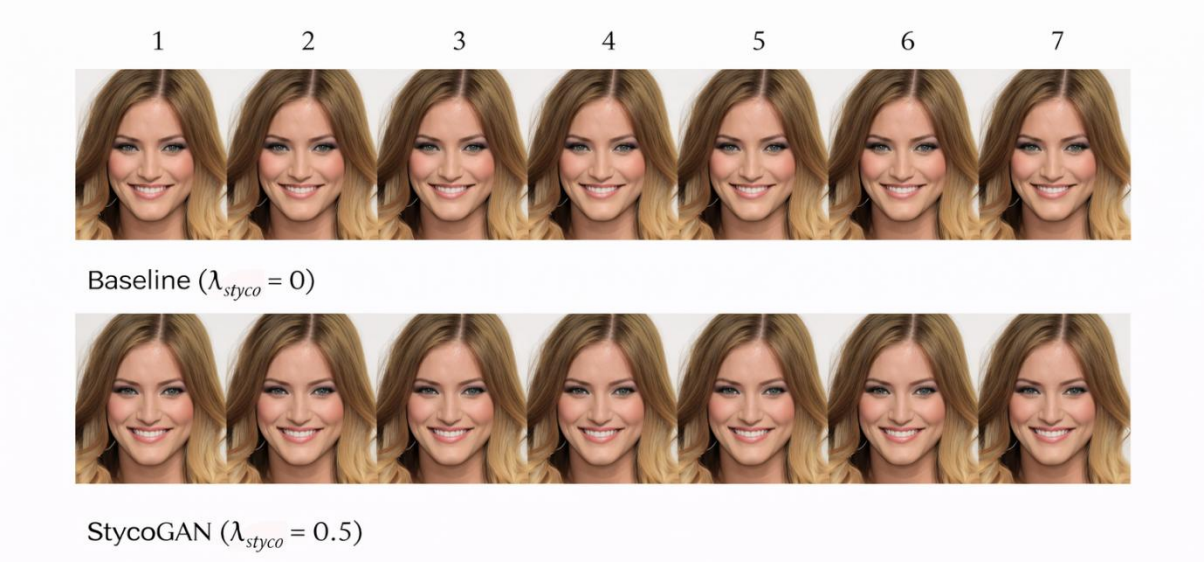


Figure 3. Qualitative visual comparison between the baseline model and StycoGAN under repeated generation. Each row shows eight images generated from the same latent identity. The baseline model exhibits perceptual flickering and subtle identity drift across generations, while StycoGAN produces more stable and stylistically consistent outputs.

The baseline generator produces visually plausible images when considered individually; however, inconsistencies become apparent when outputs are inspected across successive generations. Subtle variations in facial structure, micro-expression cues, and stylistic attributes accumulate over time, leading to perceptual flickering and gradual identity drift. These observations underscore the limitation of frame-independent generation, in which each output is synthesized without explicit constraints on cross-generation coherence.

In contrast, StycoGAN yields sequences with smoother perceptual transitions and substantially improved identity preservation. Core facial geometry and stylistic characteristics remain more stable across successive outputs, resulting in a more coherent visual impression. This qualitative behavior is consistent with the quantitative improvements reported in Table 2, particularly the reduced LPIPS and temporal consistency error, as well as the ablation findings in Table 3, which highlight the effectiveness of the proposed style–temporal consistency loss in stabilizing intermediate feature representations. The qualitative evidence in Figure 3 demonstrates that StycoGAN enhances temporal coherence and stylistic stability under repeated generation while maintaining perceptual realism, supporting the effectiveness of feature-level temporal regularization in style-based generative modeling.

4.6. Summary of Experimental Findings

Overall, the experiments indicate that feature-level temporal regularization can improve perceptual stability under repeated generation while maintaining competitive spatial quality within the evaluated setting. The combination of feature-level temporal modeling and Styco-Consistency loss provides a principled mechanism for reducing temporal artifacts without introducing explicit motion constraints. These findings empirically validate the design choices presented in Section 3 and confirm the suitability of the proposed framework for sequential image synthesis.

5. DISCUSSION

This study examines style–temporal consistency in repeated generative synthesis by integrating feature-level temporal regularization into a style-based generative adversarial framework. The following

discussion contextualizes the findings with respect to prior work, analyzes the observed trade-offs, and highlights the limitations and implications of the proposed approach.

5.1. Comparison with Prior Work

Existing approaches to temporal coherence in generative modeling predominantly focus on explicit motion modeling, particularly in video generation tasks. Methods based on 3D convolutions, recurrent architectures, or optical-flow–guided constraints are commonly employed to enforce temporal alignment across frames (e.g., Vondrick et al., 2016; Wang et al., 2018). While effective for motion realism, such approaches often incur substantial computational overhead and may compromise spatial fidelity, especially at high resolutions.

In contrast, style-based generative models, most notably StyleGAN and its variants, prioritize spatial realism and latent disentanglement (Karras et al., 2019; Karras et al., 2020). However, these models treat each generation independently, which can lead to perceptual instability and identity drift under repeated sampling—an issue acknowledged but not directly addressed in the original formulation.

The proposed StycoGAN diverges from motion-centric video GANs and frame-independent style-based models by reframing temporal coherence as a feature-level perceptual consistency problem rather than a motion prediction task. Similar to perceptual regularization strategies used in image similarity assessment (Zhang et al., 2018), StycoGAN enforces stability in intermediate representations that are closely aligned with human visual perception. This design enables temporal stabilization without introducing explicit motion constraints or additional discriminators, distinguishing StycoGAN from prior temporal GAN architectures.

5.2. Interpretation of Quantitative and Qualitative Results

Quantitative evaluation shows that StycoGAN consistently improves perceptual stability across repeated generations, as indicated by reduced LPIPS and temporal consistency error. LPIPS has been shown to correlate strongly with human perceptual judgments of image similarity (Zhang et al., 2018), making it particularly suitable for assessing perceptual variation across successive outputs. The observed reduction in LPIPS therefore provides strong evidence that the proposed regularization effectively suppresses perceptual flickering and identity drift.

The qualitative comparison in Figure 3 further elucidates these findings. While baseline outputs remain visually plausible at the frame level—consistent with the strong spatial modeling capacity of StyleGAN-based architectures (Karras et al., 2020)—subtle variations in facial geometry and expression accumulate across generations. StycoGAN mitigates these effects by stabilizing feature-level representations, resulting in outputs that are perceived as belonging to the same identity across repeated generation. The alignment between qualitative observations and quantitative metrics reinforces the validity of the proposed temporal regularization.

5.3. Trade-off Between Temporal Stability and Spatial Expressiveness

The ablation study reveals a clear trade-off between temporal stability and spatial expressiveness. Increasing the temporal regularization weight monotonically improves perceptual consistency but may slightly degrade FID at higher values. FID measures the discrepancy between generated and real image distributions in feature space (Heusel et al., 2017); thus, a moderate increase in FID suggests reduced diversity or flexibility in spatial modeling rather than outright degradation of visual quality. Such trade-offs are well documented in constrained generative modeling, where stronger regularization often improves stability at the expense of expressiveness. In this context, excessive temporal constraints limit

the generator's ability to explore fine-grained stylistic variations, leading to diminished distributional coverage. The results indicate that temporal regularization should be applied judiciously, with moderate weighting providing a practical balance between realism and stability rather than maximizing consistency alone.

5.4. Limitations

Despite its effectiveness, the proposed approach has several limitations. First, evaluation is conducted under a repeated generation setting, which assesses perceptual stability across successive outputs rather than true video sequences with explicit temporal ground truth. While this setting is appropriate for analyzing identity drift in style-based generators, it does not capture motion dynamics present in real video data. Second, temporal modeling is applied at a single intermediate feature resolution. Although this design balances abstraction and spatial detail, it may not fully capture temporal dependencies at multiple semantic levels. Prior work suggests that multi-scale temporal modeling can further enhance coherence but at the cost of increased complexity (Wang et al., 2018). Finally, experiments are conducted primarily on facial image synthesis. Faces provide a controlled domain for evaluating identity preservation; however, extending the approach to object-centric or scene-level generation remains an open direction for future research.

5.5. Implications and Future Directions

The findings suggest that temporal coherence in generative models can be effectively addressed without explicit motion modeling. Feature-level temporal regularization offers a lightweight and perceptually grounded alternative for applications where identity and style preservation across repeated generation are critical, such as avatar synthesis, identity-consistent image generation, and iterative creative systems. Future work may explore combining feature-level temporal consistency with lightweight motion modeling for true video synthesis, as well as extending temporal regularization across multiple feature scales. Evaluating the proposed approach on diverse visual domains would further clarify its generalizability.

6. Conclusion

This paper presented StycoGAN, a style–temporal consistency framework designed to improve perceptual stability under repeated generative synthesis in style-based generative adversarial networks. By introducing a feature-level temporal regularization mechanism into a style-based generator, the proposed approach addresses perceptual flickering and identity drift that commonly arise when frame-independent generation is repeatedly applied. Extensive qualitative and quantitative evaluations demonstrate that StycoGAN consistently improves temporal coherence and identity preservation without substantially compromising spatial realism. Qualitative comparisons reveal smoother perceptual transitions and more stable stylistic attributes across successive outputs, while quantitative results show reduced LPIPS and temporal consistency error alongside competitive FID scores. An ablation study further highlights the importance of balancing temporal regularization strength, where moderate weighting achieves the most favorable trade-off between stability and expressiveness.

Importantly, this work reframes temporal coherence in generative modeling as a perceptual feature consistency problem rather than an explicit motion modeling task. This perspective enables temporal stabilization with minimal architectural overhead and avoids the complexity typically associated with video-centric generative models. As a result, StycoGAN offers a lightweight yet effective solution for applications that require identity-consistent generation across repeated synthesis, such as avatar creation, identity-aware image generation, and iterative creative systems. Despite these promising

results, the proposed approach is evaluated under a repeated generation setting and within a facial image domain, which limits its direct applicability to full video synthesis and more diverse visual categories. Future work will explore extending feature-level temporal regularization to multi-scale representations, integrating lightweight motion cues, and evaluating the framework across broader generative domains.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions

MCW conceived the research idea, designed the study framework, developed the proposed StycoGAN architecture, conducted the experiments, and led the analysis and interpretation of the results. DM contributed to the methodological design, assisted in model implementation and experimental validation, and participated in result interpretation. HH contributed to the literature review, data preparation, and supported the evaluation and discussion of the findings. BK provided conceptual guidance, critically reviewed the manuscript for intellectual content, and contributed to the refinement of the analysis and overall presentation. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Data availability

The data used in this study were curated from publicly available video datasets, including the YouTube Faces dataset. Frames were extracted and organized into temporally ordered image sequences for experimental evaluation. All preprocessing procedures were performed to support reproducibility and controlled analysis of temporal consistency. The curated datasets generated and analyzed in this study are available from the corresponding author upon reasonable request.

Code availability

The implementation code supporting the findings of this study, including model architecture and experimental configurations, is available in https://github.com/marscaroline/stycogan-experiments.

REFERENCES

Asperti, A., Merizzi, F., Paparella, A., Pedrazzi, G., Angelinelli, M., & Colamonaco, S. (2023). Precipitation nowcasting with generative diffusion models. *Applied Intelligence*, *55*(2). https://doi.org/10.1007/s10489-024-06048-y

Bermano, A. H., Gal, R., Alaluf, Y., Mokady, R., Nitzan, Y., Tov, O., Patashnik, O., & Cohen-Or, D. (2022). State-of-the-Art in the Architecture, Methods and Applications of StyleGAN. *Computer Graphics Forum*, *41*(2), 591–611. https://doi.org/10.1111/CGF.14503

Chen, C., Qiu, A., Chen, H., Chen, Y., Liu, X., & Li, D. (2023). Prediction of Pollutant Concentration Based on Spatial–Temporal Attention, ResNet and ConvLSTM. *Sensors 2023, Vol. 23, 23*(21). https://doi.org/10.3390/S23218863

Dadfar, R. (2023). *Text-Driven Fashion Image Manipulation with GANs : A case study in full-body human image manipulation in fashion*. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-328241

Darwich, M., & Bayoumi, M. A. (2025). Utilizing generative adversarial networks (GANs) for realistic video stream augmentation and enhancement. *Cluster Computing 2025 28:13*, *28*(13), 1–28. https://doi.org/10.1007/S10586-025-05538-Z

Elasri, M., Elharrouss, O., Al-Maadeed, S., & Tairi, H. (2022). Image Generation: A Review. *Neural Processing Letters*, *54*(5), 4609–4646. https://doi.org/10.1007/S11063-022-10777-X/METRICS

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*. https://proceedings.neurips.cc/paper/5423-generative-adversarial-nets

Huang, Z., Yu, Y., Yang, L., Qin, C., Zheng, B., Zheng, X., Zhou, Z., Wang, Y., & Yang, W. (2024). Motion-aware Latent Diffusion Models for Video Frame Interpolation. *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*, *10*, 1043–1052. https://doi.org/10.1145/3664647.3680846

Javed, S., Jung, S. K., Mahmood, A., & Bouwmans, T. (2016). Motion-Aware Graph Regularized RPCA for background modeling of complex scenes. *Proceedings - International Conference on Pattern Recognition*, *0*, 120–125. https://doi.org/10.1109/ICPR.2016.7899619

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). *Analyzing and Improving the Image Quality of StyleGAN* (pp. 8110–8119). https://github.com/NVlabs/stylegan2

Kenny, S., Mahmood, N., Honda, C., Black, M. J., & Troje, N. F. (2019). Perceptual effects of inconsistency in human animations. *ACM Transactions on Applied Perception*, *16*(1). https://doi.org/10.1145/3301411

Leng, J., Gao, M., Gong, H., Chen, B., Zhou, C., Shi, M., Chen, Z., & Li, X. (2023). Spatio-temporal prediction of regional land subsidence via ConvLSTM. *Journal of Geographical Sciences 2023 33:10*, *33*(10), 2131–2156. https://doi.org/10.1007/S11442-023-2169-8

Liu, W., Zhong, X., Dai, Y., Jia, X., Wang, Z., & Satoh, S. (2025). Motion-Consistent Representation Learning for UAV-Based Action Recognition. *IEEE Transactions on Intelligent Transportation Systems*, *26*(11), 21036–21046. https://doi.org/10.1109/TITS.2025.3593805

Lu, R., Cheng, Z., Chen, B., & Yuan, X. (2024). Motion-Aware Dynamic Graph Neural Network for Video Compressive Sensing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 7850–7866. https://doi.org/10.1109/TPAMI.2024.3395804

Ma, F., Xie, Y., Li, Y., He, Y., Zhang, Y., Ren, H., Liu, Z., Yao, W., Ren, F., Yu, F. R., & Ni, S. (2025). A Review of Human Emotion Synthesis Based on Generative Technology. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2025.3573878

Manesh, F. S., Nazari, A., Mansoorizadeh, M., Dezfoulian, M., Sobhani Manesh, F., Nazari, A., Mansoorizadeh, M., & Dezfoulian, M. (2025). VG-CGARN: Video Generation Using

Convolutional Generative Adversarial and Recurrent Networks. *International Journal of Web Research*, *8*(2), 65–77. https://doi.org/10.22133/IJWR.2025.505096.1265

Shahriar, S., & Shahriar, S. (2025). *Generative AI for Text-to-Video Generation: Recent Advances and Future Directions*. https://doi.org/10.20944/PREPRINTS202512.1476.V1

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-C., & Kong Observatory, H. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems*, *28*.

Skorokhodov, I., Tulyakov, S., & Elhoseiny, M. (2022). *StyleGAN-V: A Continuous Video Generator With the Price, Image Quality and Perks of StyleGAN2* (pp. 3626–3636). https://universome.github.io/stylegan-v.

Suekane, K., Haji, S., Sano, H., & Takagi, T. (2023). Conceptual Attention in StyleGAN. *Proceedings - 17th IEEE International Conference on Semantic Computing, ICSC 2023*, 123–130. https://doi.org/10.1109/ICSC56153.2023.00027

Tellamekala, M. K., & Valstar, M. (2019). Temporally Coherent Visual Representations for Dimensional Affect Recognition. *2019 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019*, 406–412. https://doi.org/10.1109/ACII.2019.8925529

Tran, T.-H., Bach, V.-D., & Doan, H.-G. (2020). *vi-MoCoGAN: A Variant of MoCoGAN for Video Generation of Human Hand Gestures Under Different Viewpoints* (pp. 110–123). https://doi.org/10.1007/978-981-15-3651-9_11

Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., … Liu, Z. (2025). *Wan: Open and Advanced Large-Scale Video Generative Models*. https://arxiv.org/pdf/2503.20314

Wang, Y., Teng, J., Cao, J., Li, Y., Ma, C., Xu, H., & Luo, D. (2025). *Efficient Video Face Enhancement with Enhanced Spatial-Temporal Consistency* (pp. 2183–2193).

Zhang, G., Zhu, Y., Cui, Y., Zhao, X., Ma, K., & Wang, L. (2025). *Motion-Aware Generative Frame Interpolation*. https://arxiv.org/pdf/2501.03699

Zhang, T., & Tang, H. (2025). *Style Transfer: A Decade Survey*. https://arxiv.org/pdf/2506.19278

Zhang, X., & You, Q. (2025). Research on a Diffusion Model-Driven Framework for New Media Animation Content Generation and Communication Optimization Based on Multimodal User Behavior Data and Attention Mechanism for Artistic Style Transfer Algorithms. *Https://Doi.Org/10.1142/S0129156425408666*, 2540866. https://doi.org/10.1142/S0129156425408666