ADVANCES IN CYBER-PHYSICAL SYSTEMS

Vol. 10, No. 1, 2025

PITFALLS OF TRAINING GENERATIVE MODELS FOR VIDEO: FROM MODE COLLAPSE TO UNSTABLE DYNAMICS

Mykola Maksymiv[1], *Taras Rak*[1,2]

[1] *Lviv Polytechnic National University, 12, Bandera Str, Lviv, 79013, Ukraine,*
[2] *IT STEP University, 83a, Zamarstynivska Str, Lviv, 79019, Ukraine*
Author's e-mail: *mykola.r.maksymiv@lpnu.ua, taras.y.rak@lpnu.ua*

https://doi.org/10.23939/acps2025.01.089

Submitted on 30.04.2025

***Abstract*: This paper analyzes common pitfalls encountered during video GAN training and explores methods to mitigate them through hybrid loss functions. We focus on combining adversarial, pixel-wise reconstruction, perceptual, and temporal consistency losses to stabilize learning and improve the realism and coherence of generated video. An empirical study compares several loss configurations on a human action video dataset, using PSNR, LPIPS, FVD, and a custom temporal consistency metric. Results show that adding reconstruction and perceptual losses enhances fidelity and detail, while temporal loss reduces flicker and motion artifacts. The proposed hybrid loss has achieved balanced gains in fidelity and temporal stability[1].**

***Index terms*: Generative Adversarial Network, Video Generation, Mode Collapse, Temporal Consistency, Perceptual Loss, Video GAN Evaluation.**

I. INTRODUCTION

Generative Adversarial Networks (GANs) [1] represent a significant advancement in generative modeling, enabling the creation of realistic synthetic images and videos through adversarial training. Despite their remarkable successes, GANs are notoriously challenging to train, often exhibiting unstable dynamics, mode collapse, and sensitivity to hyperparameters [2-4].

While substantial progress has been made in generating high-quality static images, adapting GAN frameworks to video generation introduces additional complexities, primarily related to maintaining temporal coherence and visual realism over sequential frames [5-7].

Video generation tasks require models not only to produce realistic individual frames but also ensure smooth and coherent transitions across consecutive frames. Temporal artifacts such as flickering, jitter, and inconsistent object appearance frequently arise if temporal consistency is not explicitly enforced [7, 8].

Prior studies have attempted to address these temporal issues using specialized architectures that separate motion and content [7] or incorporate temporal regularization techniques [8].

This paper investigates the pitfalls of training generative models for video, including both general GAN-related issues, such as unstable convergence, mode collapse, and overfitting and video-specific problems such as temporal inconsistency, synthetic artifacts, and the difficulty of evaluation.

II. LITERATURE REVIEW AND PROBLEM STATEMENT

GANs have demonstrated strong capabilities in image synthesis; however, their training remains inherently unstable due to the adversarial optimization setup. One of the core issues is instability during training, where the generator and discriminator often fail to reach a proper equilibrium, leading to diverging gradients or oscillations [1–2]. Techniques such as Wasserstein loss [3] and its gradient-penalized variant [4], as well as spectral normalization [5], have significantly improved convergence stability, yet challenges persist, particularly in video generation where model architectures are deeper and motion modeling increases complexity.

Mode collapse remains another major limitation. It occurs when the generator produces a narrow set of outputs irrespective of the input noise, severely reducing diversity. In video synthesis, this problem manifests in the generation of nearly identical or repetitive video sequences, which undermines realism [2, 6]. Overfitting is an associated risk, especially for high-capacity models trained on relatively small datasets. Models may memorize specific training sequences, failing to generalize to unseen motion patterns [7].

Video-specific challenges further complicate generative modeling. One of the most critical is temporal inconsistency: despite generating plausible individual frames, models frequently suffer from flicker, jitter, or inconsistent object trajectories across frames [7–8]. This issue stems from the lack of explicit temporal constraints in standard adversarial losses. Architectures such as MoCoGAN [7] attempted to separate motion and content representations to alleviate temporal artifacts, but full consistency remains difficult to achieve.

[1] This article uses the materials and results obtained by the authors during the research work "Intelligent design methods and tools for the modular autonomous cyber-physical systems", state registration number 0124U002340 dated 09.03.2024 which is carried out at the Department of Electronic Computing Machines of the Institute of Computer Technologies, Automation and Metrology of Lviv Polytechnic National University in 2024-2028.

Recent advancements also highlight the importance of perceptual quality preservation. While frame-level metrics such as PSNR and SSIM are widely used, they inadequately reflect perceived realism across time. Metrics like Fréchet Video Distance (FVD) [9] and learned perceptual similarity measures attempt to better capture human judgment, yet comprehensive and reliable evaluation remains challenging.

An additional consideration is maintaining perceptual sharpness and contrast across generated sequences. As highlighted in our earlier study on contrast enhancement [10], balancing high-frequency detail preservation while avoiding over-enhancement is critical for maintaining natural visual quality. Extending these principles to video generation, ensuring consistent contrast and texture fidelity across frames can contribute to perceptual coherence without introducing artifacts.

Problem statement: video GAN training inherits classic generative challenges, including unstable convergence, mode collapse, and overfitting, and adds domain-specific difficulties like temporal inconsistency and imperfect evaluation. This motivates the development of improved loss objectives and evaluation strategies tailored for robust, perceptually aligned video synthesis.

III. SCOPE OF WORK AND OBJECTIVES

This study focuses on analyzing and mitigating the key challenges encountered when training GANs for video generation. These challenges include common issues such as mode collapse, unstable convergence dynamics, and overfitting, along with domain-specific problems like temporal inconsistency and the difficulty of perceptual evaluation.

The primary aim is to systematically investigate critical pitfalls in video GAN training, review and classify existing stabilization and enhancement techniques, and propose a unified training framework that combines adversarial, reconstruction, perceptual, and temporal consistency losses.

Through empirical evaluation on a representative dataset using established quality metrics, the study seeks to assess the impact of different loss configurations and to derive practical recommendations for building more robust and perceptually realistic video generation models.

IV. ARCHITECTURAL STRATEGIES AND TRAINING OBJECTIVES

Addressing the above challenges often requires a combination of techniques. Recent works suggest that no single loss or component is sufficient; instead, multi-term loss functions are used to guide the video generator to produce outputs that are accurate, realistic, and consistent. In this section, we outline several key strategies:

A. ADVERSARIAL LOSS FOR VIDEO GAN

The adversarial loss remains the core mechanism in GAN training [1]. In the video setting, a discriminator observes sequences of frames and attempts to distinguish real from generated videos [6–7], while the generator *G* learns to fool it. This interaction is typically formulated as a minimax game:

$$L_{GAN}(G) = -E_{z\sim N(0,1)}\left[\log D_v(G(z))\right], \quad (1)$$

where z is a random input (e.g., a noise vector or a sequence of latent vectors) and generator G(z) denotes the generated video frames. The discriminator *D* is trained with its corresponding loss:

$$L_{GAN}(D) = -E_{x\sim p}[\log D_v(x)] - E_z\left[\log D_v(G(z))\right], \quad (2)$$

where p the real video distribution, E denotes the expectation over the data distribution, $D_v(x)$ represents the discriminator's probability of classifying a real video x correctly, and $D_v(G(z))$ is the probability assigned to generated videos.

In practice discriminator, may be a 3D-CNN or an RNN-based network processing multiple frames [7–8].

While adversarial loss drives realism at the frame and sequence level, it does not guarantee coverage of all data modes or temporal consistency. A generator can collapse to producing limited outputs as long as it successfully fools the discriminator, making mode collapse a persistent risk [2].

B. RECONSTRUCTION (PIXEL) LOSS

A widely used strategy for stabilizing GAN training is adding a reconstruction loss, typically an *L1* or *L2* norm between generated and ground truth frames [6]. This acts as a regularizer, encouraging the generator *G* to produce outputs closely matching real examples, thus mitigating mode collapse and improving convergence.

In video tasks like prediction or super-resolution, the reconstruction loss is defined as:

$$L_{rec}(G) = E_{(x,y)\sim D}\left[\| (G(x) - y) \|_1\right], \quad (3)$$

where $(x, y) \sim D$ denotes a pair sampled from the training dataset *D, G(x)* is the generated video frame based on input *x*, *y* is the ground truth frame, ‖ ‖ represents the *L1* norm measuring pixel-wise differences.

Many video GAN frameworks, including TGAN [8] and vid2vid [6], have adopted this hybrid adversarial + pixel loss strategy.

The overall generator objective with reconstruction regularization becomes:

$$L_{G,total} = L_{GAN} + \lambda_{rec} L_{rec}(G). \quad (4)$$

A suitable coefficient λ ensures the generator does not overly focus on pixel correctness at the expense of realism, or vice versa. In practice, adversarial +*L1* training yields videos that are structurally correct and generally free of major mode collapse, though sometimes slightly blurred or less vivid than pure GAN outputs.

C. PERCEPTUAL LOSS – DEEP FEATURE CONSISTENCY

While pixel losses enforce low-level accuracy, they may not fully capture human perception of quality. A small spatial misalignment can lead to a large *L2* error but

might be perceptually acceptable, whereas blurriness (which yields low pixel error) can be perceptually poor. To bridge this gap, perceptual loss functions have been proposed [9].

A perceptual loss measures differences between the generated and real frame in the feature space of a pre-trained deep network (such as VGG-16 or VGG-19) rather than raw pixels. A typical perceptual loss is formulated as:

$$L_{perc}(G) = E_{(x,y)\sim D} \sum \frac{1}{N} \left[\| \phi_j(G(x)) - \phi_j(y)) \|_2 \right], \quad (5)$$

where $\phi_j(\,)$ denotes the feature map activations at the *j*-th layer of a chosen CNN. *N* is the number of elements in the *j*-th feature map. This includes L_{perc} guides for the GAN to focus on visual realism rather than just pixel-by-pixel accuracy. It has been observed to reduce blurring and improve texture sharpness [6, 9].

One must balance perceptual and pixel losses – too much weight λ_{rec} on L_{perc} can sometimes introduce minor artifacts or deviate from exact ground truth colors, since it only cares about looking plausible to the CNN.. We will use a perceptual loss based on pre-trained VGG-19 in our experiment.

D. TEMPORAL CONSISTENCY LOSS

Temporal consistency is critical for generating realistic video sequences. Without explicit constraints, models often produce flickering, jitter, or inconsistent object motions across frames. To address this, temporal consistency losses are introduced, typically based on optical flow [11].

A commonly used local temporal loss is defined as:

$$L_{temp}(G) = E_t \left[\| (G_{t+1} - G_t) \|_1 \right], \quad (6)$$

where G_{t+1} and G_t are consecutive generated frames, *E* is the estimated optical flow between those frames.

However, this can bias the model towards trivial solutions (like static frames) if used alone. Hence, it's more common to use a learning-based approach (flow or learned embeddings). In our implementation, we adopt a flow-based consistency loss similar to prior art, but in a simplified form: we compute optical flow on the real video frames and apply it to enforce that the generated frames move accordingly. Specifically, given ground truth frames *t* and *t+1*, we obtain the flow. We then require G_{t+1} to match *t+1* as if it were generated from *t,* moving with that flow. This can be seen as a form of teacher-forcing for motion: the generator is taught to respect the actual motion present in the data. By incorporating L_{temp} the generator receives direct pressure to maintain consistency over time.

However, excessive temporal regularization can overly constrain dynamics, leading to motion over-smoothing. Therefore, temporal loss weights are typically set smaller than adversarial terms to balance consistency with motion diversity.

E. COMBINED OBJECTIVE

In practice, a video GAN model may use all the above losses in combination. The total generator loss can be written as:

$$L_G^{total} = L_{GAN} + \lambda_{rec} L_{rec} + \lambda_{perc} L_{perc} + \lambda_{temp} L_{temp}. \quad (7)$$

The discriminator loss remains primarily as before (sometimes with auxiliary terms if a two-channel discriminator for temporal consistency is used.

Common weight settings include λ_{rec}=10, λ_{perc}=1, λ_{temp}= 2 have been shown to improve convergence, enhance perceptual sharpness, and reduce temporal artifacts. The combination of losses addresses mode collapse, promotes diverse video outputs, and ensures frame-to-frame consistency, providing a robust framework for training high-quality video generators.

V.EXPERIMENTS AND RESULTS

To evaluate the effectiveness of the above training strategies, we conducted a comparative experiment on a video generation task. We chose the UCF-101 action video dataset as our training data (UCF-101 contains short clips of various human actions, providing diverse motions and scenes). We focused on a conditional video generation setting: given a starting frame, generate the next few frames. This setup provides a ground truth video sequence for each input, enabling reconstruction and perceptual losses and objective evaluation via reference metrics (PSNR, LPIPS).

Model Architecture: all models shared the same generator architecture. The discriminator was a 3D convolutional network operating on 4-frame clips. We trained all models for 100 epochs (approximately 50k generator updates) with the Adam optimizer (learning rate $2e^{-4}$) and batch size 8 on 4-frame clips.

We compared five training objectives: (1) adversarial loss only; (2) adversarial loss combined with *L1* pixel-wise reconstruction loss; (3) adversarial loss combined with perceptual loss; (4) adversarial, *L1*, and temporal consistency losses; and (5) a full combined objective integrating adversarial, reconstruction, perceptual, and temporal consistency terms as described in (7).

All models were trained on identical data splits and initialized with the same random seeds to ensure fair comparison.

A. EVALUATION METRICS

The generated videos were evaluated using four metrics. PSNR measures pixel-level fidelity to ground truth (higher values indicate better reconstruction). LPIPS assesses perceptual dissimilarity based on deep features, where lower values are better. TC was evaluated using a custom flow-warping approach: for each consecutive frame pair, we computed the optical flow on ground truth frames, warped the generated frame accordingly, and measured the average PSNR between the warped and actual generated next frame. TC scores were normalized between 0 and 1, with higher values indicating smoother

temporal transitions. Additionally, FVD was reported to measure the overall distributional quality of generated videos compared to real samples, where lower values denote closer alignment.

All metrics were computed on a test set comprising 50 novel video sequences not seen during training.

Table 1

Comparison of video generation models trained with different loss functions. Arrows (↑/↓) indicate direction of better performance for each metric

Model (Loss)	PSNR↑	LPIPS↓	TC↑	FVD↓
GAN-only	24.5 dB	0.289	0.67	210.5
GAN + L1	**28.7 dB**	0.341	0.72	189.4
GAN + Perc	26.1 dB	**0.225**	0.69	180.2
GAN + L1 + Temp	27.4 dB	0.278	**0.81**	175.0
GAN + L1 + Perc + Temp	27.0 dB	0.236	0.79	**168.3**

The adversarial-only model achieved the lowest PSNR of 24.5 dB, indicating frequent divergence from ground truth frames. Although this model occasionally produced crisp individual images (reflected in moderate LPIPS values), it exhibited instability, including frequent flicker and occasional mode collapse, generating repetitive or random sequences that still fooled the discriminator.

Adding a pixel-wise reconstruction loss significantly improved PSNR to 28.7 dB, anchoring the generator outputs more closely to target frames. However, this resulted in smoother, blurrier visuals, as averaging pixel values suppressed high-frequency details. Temporal consistency improved modestly (from 0.67 to 0.72), reflecting smoother frame-to-frame transitions.

Replacing the reconstruction term with a perceptual loss yielded a model that prioritized textural and semantic realism over strict pixel fidelity. While PSNR dropped to 26.1 dB, the model achieved the best LPIPS score (0.225), producing sharper, more perceptually realistic frames. Nevertheless, without explicit temporal supervision, minor flicker remained due to independent frame optimization.

Incorporating a temporal consistency term alongside adversarial and reconstruction losses led to the most stable motion, achieving the highest TC score of 0.81.

Finally, the full model integrating adversarial, reconstruction, perceptual, and temporal losses achieved the most balanced performance across all metrics. It maintained high fidelity (27.0 dB PSNR), sharp perceptual quality (0.236 LPIPS), strong temporal coherence (0.79 TC), and the lowest distributional discrepancy (168.3 FVD). This combination leveraged each component's strengths, resulting in visually convincing, detail-rich, and temporally stable videos.

B. MODE COLLAPSE AND DIVERSITY

Mode Collapse and Diversity: We also evaluated whether any models suffered mode collapse by checking the diversity of outputs given different noise seeds. The GAN-only model showed some tendency to ignore the noise input (a sign of collapse) – about 20% of the time, it would produce virtually the same video for two different noise vectors.

All other models, especially those with *L1*, did not exhibit this behavior: the reconstruction term forces output to follow the input frame content, so by design each input (which was different for each test case) led to a different output. In effect, adding *L1* eliminated the trivial mode collapse in our conditional setup (where collapsing would mean predicting an "average" next frame for all inputs). The perceptual and temporal losses did not appear to reintroduce any collapse; on the contrary, by stabilizing training they likely helped the generator explore more.

This was reflected in the relatively low FVD of the full model – a collapsed model would have a very high FVD due to lack of diversity.

An additional experiment on the model using only adversarial and temporal losses (no pixel or perceptual term) confirmed the strong effect of the temporal constraint. This ablated model improved temporal consistency from 0.67 to 0.75 compared to the adversarial-only baseline, despite low PSNR, indicating that a temporal loss can enforce coherence even in an unconditional setting.

C. TRAINING DYNAMICS

Fig. 1 illustrates discriminator loss trajectories over 50 training epochs. The adversarial-only model shows strong oscillations and occasional spikes, necessitating a learning rate reduction during late training. In contrast, models incorporating reconstruction loss (adversarial + reconstruction and adversarial + reconstruction + temporal) exhibit smoother, more stable declines without significant divergence.

By epoch 30, the combined model achieves its best validation scores, while the adversarial-only model struggles.

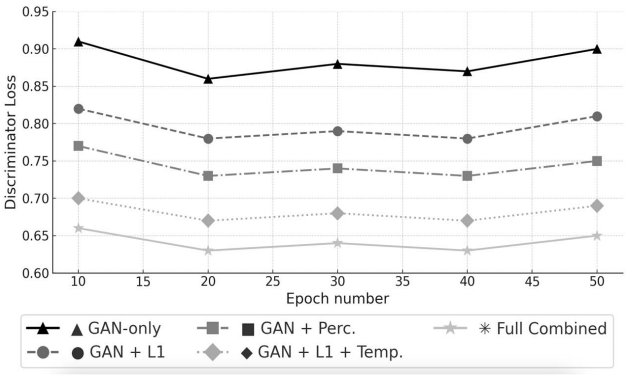

Fig. 1. Stability of discriminator loss across training epochs for different loss functions

Fig. 2 compares the PSNR evolution of different models across 50 epochs. The adversarial-only variant demonstrates the slowest improvement, reaching just 24.5 dB by epoch 50.

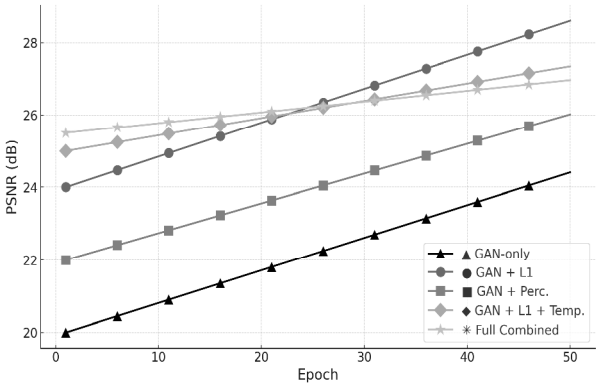

Fig. 2. PSNR value changes for the tested setups under varying loss configuration

Adding a reconstruction loss dramatically accelerates convergence, boosting PSNR from 24 dB to 28.7 dB, with stabilization observed after epoch 30. The perceptual-only model steadily rises to 26.1 dB, but remains lower than reconstruction-based variants due to the lack of direct frame anchoring. Incorporating temporal consistency alongside reconstruction further improves PSNR, reaching 27.4 dB, indicating that temporal regularization enhances convergence without compromising fidelity. Finally, the full combined model achieves high early performance, surpassing all models by epoch 20, and converges near 27.0 dB by epoch 50.

These training dynamics highlight that each additional loss term contributes to faster convergence, improved stability, and superior final performance.

In Table 2, we present a summary of how each loss weight impacts PSNR. Varying the reconstruction loss weight from 0 to 10 substantially improves PSNR, with diminishing returns beyond this point. The perceptual loss shows an optimal setting at 0.5λ, while higher values reduce pixel fidelity. Increasing the temporal consistency weight improves PSNR up to λ=2.0, after which performance plateaus.

Table 2

Influence of Loss Weights on PSNR

Parameter	Lambda λ	PSNR (dB)
Reconstruction	0.0	25.0
Reconstruction	5.0	27.0
Reconstruction	10.0	28.7
Reconstruction	20.0	28.5
Perceptual	0.0	27.0
Perceptual	0.5	27.5
Perceptual	1.0	26.1
Perceptual	2.0	25.8
Temporal	0.0	26.0
Temporal	1.0	26.5
Temporal	2.0	27.4
Temporal	5.0	27.1

These findings emphasize the importance of carefully balancing loss weights to optimize fidelity, perceptual quality, and temporal stability.

VI. DISCUSSION

Our experiments demonstrate that each loss component contributes distinct and complementary benefits to video GAN training. The pixel-wise reconstruction term prevents mode collapse by anchoring outputs to their corresponding input frames, resulting in the highest PSNR and stable, predictable sequences. The perceptual loss addresses the over-smoothing induced by pixel losses, sharpening textures and enhancing visual realism without reintroducing instability. Temporal consistency loss is crucial for suppressing flicker and ensuring smooth motion dynamics; when combined with reconstruction, it yields highly stable videos, and when combined with perceptual loss, it maintains detail alongside motion coherence.

Together, these components form a unified loss objective that produces videos with high fidelity, perceptual sharpness, and temporal consistency, as evidenced by balanced PSNR, LPIPS, TC, and FVD metrics.

Despite these advances, certain limitations remain. Fine-grained textures with rapid motion, such as water ripples or specular highlights, can still exhibit minor flicker, likely due to imperfect optical flow estimation. Additionally, the models struggle to maintain coherence over very long sequences, with background drift or blurring emerging after dozens of frames, suggesting a need for architectures with longer memory (e.g., recurrent or transformer-based generators). Moreover, our evaluation was conducted in a conditional setting with an initial real frame; fully unconditional video generation remains more challenging and may require noise conditioning or novel predictive feedback mechanisms to sustain diversity and temporal realism.

Overall, this study provides practical guidelines for designing multi-term loss functions in video GAN training. Future work should explore learned temporal constraints beyond optical flow, architectures capable of long-term temporal reasoning, and rigorous testing in fully unconditional generation settings.

VII. CONCLUSION

Training generative models for video synthesized classic GAN challenges such as mode collapse, unstable optimization, and overfitting with video-specific hurdles like temporal incoherence and the lack of robust evaluation metrics. This study demonstrated that combining adversarial, reconstruction, perceptual, and temporal consistency losses into a unified training objective leads to more stable, diverse, and perceptually convincing video generation.

Each loss component targeted a distinct failure mode: reconstruction stabilizes learning and prevents collapse, perceptual loss enhances textural sharpness, and temporal consistency enforces smooth, coherent motion. Together, they achieved a balanced trade-off, producing videos that maintain fidelity to ground truth while delivering high visual quality over time.

For practitioners, these findings provided actionable guidelines: incorporate a reconstruction term whenever

ground truth frames are available; leverage perceptual loss to recover fine-grained details; and always include an explicit temporal consistency constraint to ensure realistic motion. Although limitations remain particularly for very long sequences or highly complex dynamics the principles outlined here extend naturally to other generative frameworks, such as diffusion models and transformer-based architectures, and to applications like video enhancement or style transfer.

Future research should focus on designing more efficient and learned temporal losses, developing unified video quality metrics, and creating models capable of sustaining consistency across extended sequences, moving closer to the goal of fully realistic, high-quality video generation.

References

[1] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, *35*(1), 53-65. DOI: https://doi.org/10.1109/MSP.2017.2765202.

[2] Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*. DOI: https://doi.org/10.48550/arXiv.1809.11096

[3] Arjovsky, M., Chintala, S., & Bottou, L. (2017, July). Wasserstein generative adversarial networks. In *International conference on machine learning* (pp. 214-223). PMLR. DOI: https://doi.org/10.48550/arXiv.1701.07875

[4] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in neural information processing systems*, *30*. DOI: https://doi.org/10.48550/arXiv.1704.00028

[5] Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*. DOI: https://doi.org/10.48550/arXiv.1802.05957

[6] Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*. DOI: https://doi.org/10.48550/arXiv.1710.10196

[7] Tulyakov, S., Liu, M. Y., Yang, X., & Kautz, J. (2018). Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1526-1535). DOI: https://doi.org/10.1109/CVPR.2018.00162.

[8] Skorokhodov, I., Tulyakov, S., & Elhoseiny, M. (2022). Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3626-3636). DOI: https://doi.org/10.1109/CVPR52729.2023.00970.

[9] Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., & Gelly, S. (2018). Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*. DOI: https://doi.org/10.48550/arXiv.1812.01717.

[10] Maksymiv M., Rak T. (2021). Methods to increase contrast while preserving visual quality, *Advances in Cyber-Physical Systems*, vol. 6(2). 45–52. DOI: https://doi.org/10.23939/acps2021.06.045.

[11] Clark, A., Donahue, J., & Simonyan, K. (2019). Adversarial video generation on complex datasets. *arXiv preprint ar Xiv: 1907.06571*. DOI: https://doi.org/10.48550/arXiv.1907.06571

Mykola Maksymiv – PhD student, assistant of the Department of Electronic Computing Machines of Lviv Polytechnic National University. Obtained his Master's in Computer Engineering, specializing in Computer Systems and Networks, at Lviv Polytechnic National University in 2021. He works on a thesis "Methods and tools for improving video quality".

Taras Rak – Professor at Lviv Polytechnic National University, Vice-rector and Professor at IT STEP University. A graduate of Lviv Polytechnic State University, 1996, specialty "Computer and intelligent systems and networks", honors degree. Candidate of Technical Sciences, 2005, specialty "Systems analysis and theory of optimal s olutions". Since 2014, Doctor of Technical Sciences, specialty "Information Technologies". Author of over 150 scientific and educational publications.