UDC 004.9

MULTI-SCALE TEMPORAL GAN-BASED METHOD FOR HIGH-RESOLUTION AND MOTION STABLE VIDEO ENHANCEMENT

Maksymiv M. R. – Postgraduate student, Assistant of the Department of Electronic Computing Machines of the Lviv Polytechnic National University, Lviv, Ukraine.

Rak T. Y. – Dr. Sc., Associate Professor, Professor at IT STEP University, and Professor of the Department of Electronic Computing Machines at Lviv Polytechnic National University, Lviv, Ukraine.

ABSTRACT

Context. The problem of improving the quality of video images is relevant in many areas, including video analytics, film production, telemedicine and surveillance systems. Traditional video processing methods often lead to loss of details, blurring and artifacts, especially when working with fast movements. The use of generative neural networks allows you to preserve textural features and improve the consistency between frames, however, existing methods have shortcomings in maintaining temporal stability and the quality of detail restoration.

Objective. The goal of the study is the process of generating and improving video images using deep generative neural networks. The purpose of the work is to develop and study MST-GAN (Multi-Scale Temporal GAN), which allows you to preserve both spatial and temporal consistency of the video, using multi-scale feature alignment, optical flow regularization and a temporal discriminator.

Method. A new method based on the GAN architecture is proposed, which includes: multi-scale feature alignment (MSFA), which corrects shifts between neighboring frames at different levels of detail; a residual feature boosting module to restore lost details after alignment; optical flow regularization, which minimizes sudden changes in motion and prevents artifacts; a temporal discriminator that learns to evaluate the sequence of frames, providing a consistent video without flickering and distortion.

Results. An experimental study of the proposed method was conducted on a set of different data and compared with other modern analogues by the metrics SSIM, PSNR and LPIPS. As a result, values were obtained that show that the proposed method outperforms existing methods, providing better frame detail and more stable transitions between them.

Conclusions. The proposed method provides improved video quality by combining detail recovery accuracy and temporal frame consistency.

KEYWORDS: video enhancement, deep neural networks, generative adversarial networks, multiscale smoothing, temporal discriminator, motion stabilization.

ABBREVIATIONS

GAN is a Generative Adversarial Network;

VSR is a Video Super-Resolution;

MST-GAN is a Multi-Scale Temporal Generative Adversarial Network;

MSFA is a Multi-Scale Feature Alignment;

OF is a Optical Flow;

PSNR is a Peak Signal-to-Noise Ratio;

SSIM is a Structural Similarity Index;

LPIPS is a Learned Perceptual Image Patch Similarity;

VFI is a Video Frame Interpolation;

BM3D is a denoising method implementation on Python;

Noise2Noise is a GAN-based denoising method;

RAFT is a Recurrent All-Pairs Field Transforms (Opti-cal Flow Model);

DAIN is a Depth-Aware Video Frame Interpolation Network;

PDE is a Partial Differential Equation;

SRCNN is a Super-Resolution Convolutional Neural Network;

ESPCN is an Efficient Sub-Pixel Convolutional Network;

VSR-DUF is a Video Super-Resolution with Dynamic Upsampling Filters;

RBPN is a Recurrent Back-Projection Network;

TimeWarpGAN is a GAN-based model for improving temporal consistency in video enhancement;

ResNet is a Residual Network (ResNet) is a Convolutional Neural Network (CNN) architecture;

VGG is a Very Deep Convolutional Networks;

PyTorch is an open source machine learning framework for Python;

vCPU is a virtual Central processing unit;

GPU is a graphical processing unit;

Adam optimizer is an adaptive moment stochastic gradient descent method.

NOMENCLATURE

x,y are the image indexes;

I_t is an input video frame t before enhancement;

$\overline{I_t}$ is an enhanced output of particular t frame;

F_t is a multi-scale feature extracted value of frame t;

D_t is a temporal discriminator for video coherence;

u_{i} is a mean insensitive of image x;

σ_X is an image X variance;

σ_{XY} is a covariance of images X and Y;

k is a stabilizing constant;

$L_{entropy}$ is an entropy calculation to measure the level of noise within an image;

$p(x, y)$ is a probability distribution of pixel intensities in the frame;

DOI 10.15588/1607-3274-2025-3-9

L_{PDE} is a Partial Differential Equation (PDE) constraint

V is a speed of pixel in horizontal x and vertical y axis;

I is a mage intensity in spaces x and y, and over time t;

∇V is a velocity gradient that control the smoothness of the flow;

ε is a feature extraction function;

$L_{MST-GAN}$ is a total loss function for model training;

L_{LPIPS} is a Learned Perceptual Image Patch Similarity (LPIPS) metric;

L_{GAN} is a modified adversarial loss for the Temporal Discriminator;

W_t is a warping function based on optical flow of frame t;

$F_{aligned}$ is a motion-aligned feature map from MSFA;

$R(F^t_{aligned})$ is a predicted residual correction of aligned feature-extracted frame t;

$D_t(I_t)$ is a probability that the real triplet of t frame with siblings is authentic;

$D_t(\overline{I_t})$ is a probability that the generated sequence is synthetic;

λ_1 is a coefficient that controls adversarial learning strength, encouraging realism;

λ_2 is a coefficient that ensures smooth motion transitions, penalizing flickering;

λ_3 is a coefficient that prevents noise accumulation, ensuring clean video quality.

L_{L1} is a loss function of pre-trained generator;

MSE is a mean squared error between the generated and ground truth images;

MAX is a maximum value of pixel.

INTRODUCTION

Video content has become a crucial part of our daily lives, from entertainment to education and advertising communication. However, poor video quality can significantly reduce the viewers' experience and engagement with the content.

Nowadays video enhancement is a rapidly evolving field in artificial intelligence, driven by the growing demand for high-quality video content in streaming, surveillance, film restoration, and gaming. The main challenge is in preserving temporal consistency while improving spatial resolution and visual clarity. Traditional methods, including super-resolution and frame interpolation, often struggle with motion artifacts, flickering, and misalignment between consecutive frames.

A significant breakthrough in video generation has been the adoption of Generative Adversarial Networks (GANs). Since their introduction, GANs have demonstrated remarkable success in synthesizing high-resolution images and enhancing video sequences. However, existing GAN-based video restoration models still suffer from motion instability, noise accumulation, and optical flow misalignment. These limitations lead to ghosting effects, unnatural frame transitions, and loss of fine details in high-motion video sequences.

The object of study is the process of video enhancement and restoration using deep learning techniques.

The subject of study is the development of a GAN-based method for improving video quality, ensuring temporal consistency, and reducing motion artifacts.

The purpose of the work is to develop an efficient and high-quality video enhancement method that maintains realistic motion while addressing the shortcomings of existing GAN-based approaches. The proposed method should improve frame coherence, reduces noise accumulation, and enhances motion stability in video sequences.

1 PROBLEM STATEMENT

One of the primary challenges in video enhancement is the presence of motion artifacts and flickering in high-motion scenes. These issues arise due to misaligned frames, poor motion estimation, and limited temporal awareness in many existing models.

Optical flow-based methods [1, 2] attempt to estimate motion between frames to improve alignment but often fail in occluded regions, leading to warping distortions.

GAN-based approaches such as TecoGAN [3] and EDVR [4] enhance video frames but struggle with maintaining temporal stability, leading to flickering effects and unnatural transitions.

A common measure of image quality degradation is the Structural Similarity Index (SSIM), which is defined as follows [5]:

$$SSIM(I,Y) = \frac{2u_x u_y + k_1}{u_x^2 + u_y^2 + k_1} * \frac{2\sigma_{XY} + k_2}{\sigma_X^2 + \sigma_Y^2 + k_2}. \quad (1)$$

A lower SSIM score between consecutive frames indicates a lack of temporal stability, leading to visible flickering.

Another significant issue in video enhancement is the accumulation of noise artifacts over multiple frames, leading to brightness fluctuations, color distortions, and inconsistent visual quality.

Traditional denoising techniques such as BM3D [6] work well for static images but fail to maintain temporal coherence in videos.

GAN-based denoising methods, like Noise2Noise [7], trying to suppress noise without clean training data but usually they oversmooth details and degrade fine textures.

One way to measure the level of noise within an image is through entropy calculation, which captures pixel intensity uncertainty [8]:

DOI 10.15588/1607-3274-2025-3-9

$$L_{entropy} = \sum_{x,y} p(x,y) \log p(x,y). \quad (2)$$

A higher entropy value correlates with more unpredictable noise patterns, which require effective suppression.

Most modern video enhancement models use optical flow estimation to align frames. However, errors in flow estimation can cause motion distortions, ghosting effects, and unnatural deformations.

RAFT [9] is one of the most accurate optical flow estimators but suffers from motion warping artifacts in fast-moving objects.

DAIN [10] introduces depth estimation to improve alignment but fails in occluded regions, leading to structural deformations.

Optical flow regularization is commonly enforced using a Partial Differential Equation (PDE) constraint, which smooths motion estimation errors [11]:

$$L_{PDE} = \left| \frac{\partial I}{\partial x} V_x + \frac{\partial I}{\partial y} V_y + \frac{\partial I}{\partial t} \right|^2 + \lambda \left(\left| \nabla V_x \right|^2 + \left| \nabla V_y \right|^2 \right). \quad (3)$$

Minimizing L_{PDE} ensures more stable motion estimation, reducing frame distortions in high-speed video sequences.

All these challenges indicate that current video enhancement approaches lack effective solutions for handling motion stability, noise suppression, and flickering artifacts.

2 REVIEW OF THE LITERATURE

The field of video enhancement and updating has evolved significantly due to advances in deep learning, generative models, and motion estimation methods. Traditional methods relied on hand-crafted filters and optical flow, while modern approaches include deep learning-based super-separation, frame interpolation, and GAN-based video synthesis.

Before the advent of deep learning, spatial and temporal filtering were predominantly used. Bilateral filtering and wavelet-based denoising were widely used for edge-preserving noise removal. In temporal filtering, optical flow-based interpolation [1, 2] estimated motion between frames to improve video smoothness. However, early optical flow methods struggled with occlusions, complex motion, and ghosting artifacts.

The main limitation of traditional methods is their inability to capture complex spatio-temporal patterns, making them ineffective in dynamic scenes with fast motion. The advent of convolutional neural networks (CNNs) has revolutionized video restoration, super-resolution, and frame interpolation [2]. Early models such as SRCNN [12] and ESPCN [13] focused on image super-resolution, but their extension to video processing was limited due to the lack of constraints. Later advances such as VSR-DUF [14] and RBPN [15] introduced multi-frame aggregation, where information from neighboring frames was used to improve resolution. A second breakthrough came with DAIN [10], a model that used depth-aware optical flow to interpolate missing frames, improving motion continuity.

Despite these improvements, CNN-based models still lack an effective mechanism to ensure long-term temporal consistency [2], often leading to motion artifacts (Fig. 1) and flickering.

Figure 1 – Example of motion distortions caused by optical flow failures in CNN-based VSR models

GANs have emerged as the dominant approach for realistic video enhancement, particularly in super-resolution, denoising, and frame interpolation. GAN-based models consist (Fig. 2) of a generator (which synthesizes high-quality video frames) and a discriminator (which distinguishes real frames from generated ones, thereby improving realism).

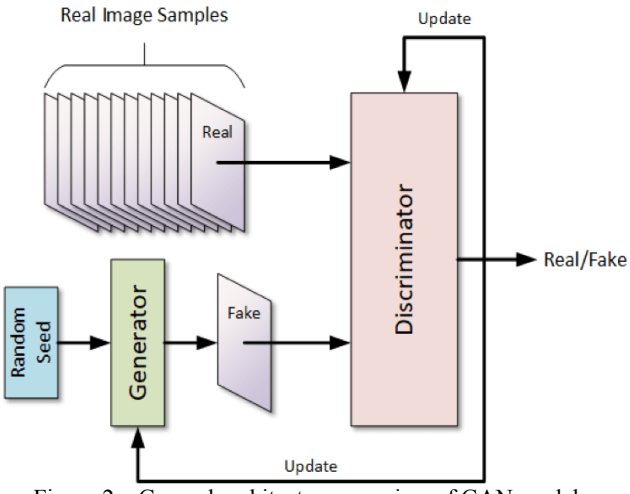

Figure 2 – General architecture overview of GAN models

One of the earliest GAN-based video models was TecoGAN [3], which introduced a temporal adversarial loss to enforce smooth frame transitions. However, it still suffers from motion instability, where small artifacts accumulate over time, leading to flickering.

To improve motion alignment, EDVR [4] leveraged deformable convolutions to refine spatial feature alignment, enhancing video super-resolution and deblurring. However, EDVR lacks explicit temporal

DOI 10.15588/1607-3274-2025-3-9

constraints, causing motion inconsistencies in high-speed sequences.

Other approaches such as TimeWarpGAN [16] introduced optical flow-guided adversarial training to improve stability, but these methods struggle in occluded and non-rigid motion regions, leading to distortions.

Based on this, the following key points can be highlighted:

1) Lack of Long-Term Temporal Consistency – Many video enhancement models focus on short-term dependencies, leading to motion inconsistencies in long sequences.

2) Noise Accumulation – GAN-based models tend to amplify artifacts over time, making video output less stable.

3) Optical Flow Misalignment – Flow-based models struggle with occlusions and rapid motion, leading to warping artifacts and distortions.

These challenges indicate there is a space for improving existing frameworks and creating a video enhancement framework capable of handling motion stability, noise suppression, and flickering artifacts while preserving fine details.

3 MATERIALS AND METHODS

Video enhancement is a challenging task that requires a balance between spatial quality improvement and temporal stability to ensure smooth transitions between frames. Traditional approaches often suffer from motion artifacts, flickering, and slippage, especially in fast-moving scenes. Our proposed MST-GAN (Multi-Scale Temporal GAN) aims to address these limitations by incorporating:

1) Multi-Scale Adversarial Facilitators (MSFA) to detect spots frames using multi-resolution warping and warped convolutions.

2) Residual Facilitator Boosting to restore lost details and prevent texture degradation.

3) Motion via Optical Flow Regularity to track motion and ensure natural object motion.

4) Temporal Coherence via a temporal discriminator that guides the generator to produce smooth, coherent video sequences.

Unlike frame-by-frame enhancement models, MST-GAN explicitly models temporal dependencies, improving long-term motion stability while preserving clear spatial details. MST-GAN takes in a sequence of three consecutive frames I_{t-1}, I_t, I_{t+1} and predicts an enhanced frame $\overline{I_t}$. Each module in the generator is interconnected, ensuring a progressive refinement process:

1) Multi-Scale Feature Alignment (MSFA) first warps feature representations across different resolutions to reduce misalignment errors.

2) The aligned features are then processed by the residual enhancement module, which restores fine details lost in the warping step.

3) Optical flow regularization is applied during feature alignment to improve motion stability, preventing distorted motion predictions.

4) The final enhanced frame is then passed to the temporal discriminator, which ensures that generated sequences preserve natural motion flow.

If we look at the generator pipeline in detail, it consists of several sequential steps.

To begin with, it is worth considering the algorithm of work Feature Extraction and Initial Representation. Each input frame I_{t-1}, I_t, I_{t+1} is passed through a shared feature extractor that outputs multi-scale feature maps:

$$F_{t-1}, F_t, F_{t+1} = \varepsilon(I_{t-1}, I_t, I_{t+1}). \quad (4)$$

*F*eature extractor ε in formula (4) can be implemented in different ways. By default, this module is not present in PyTorch libraries, but we can reuse existing implantation depending on the complexity of frames in videos. This module is commonly implemented using several convolutional layers, often resembling the early layers of a CNN backbone, such as:

ResNet-based feature extraction (ResNet-50, ResNet-101) [18].

VGG-like convolutional feature maps (used in perceptual losses) [17].

Custom-designed lightweight CNN blocks (e.g., ESPCN, RBPN) [3, 4].

These feature maps serve as the foundation for further alignment and enhancement.

The extracted features are aligned using optical flow-based warping, ensuring that motion is corrected across different resolutions:

$$F_{aligned} = W_t(F_t) + W_{t+1}(F_{t+1}).$$

Deformable convolutions further refine alignment by allowing the network to dynamically adjust receptive fields based on motion variations.

Warping often leads to loss of fine details. To compensate, MST-GAN predicts an enhancement residual that refines the aligned features:

$$\overline{I_t} = I_t + R(F^t_{aligned}).$$

This prevents over-smoothing while ensuring that textural details are preserved. Instead of modifying every pixel, the model only refines the parts that need correction. This leads to sharper details, less over-smoothing, more efficient learning (as the model doesn't have to reconstruct an entire frame from scratch).

While the Residual Enhancement Module restores spatial details, the next step ensures motion continuity across frames by penalizing sudden distortions in optical flow.

To prevent motion inconsistencies, a physics-driven regularization term is applied to the optical flow

DOI 10.15588/1607-3274-2025-3-9

estimates, ensuring that motion remains smooth across frames by using formula (3).

This term means sudden motion changes, enforcing temporal stability. It also ensures that the predicted motion does not introduce ghosting, flickering, or unnatural object distortions. Thus, Residual Enhancement ensures that each frame is locally detailed, while Optical Flow Regularization ensures that frames remain globally consistent in motion.

While the generator is responsible for improving the first frames, the temporal discriminator D_t plays a major role in ensuring smooth motion transitions and preventing measurement artifacts. Traditional GAN-based video models often work on one frame at a time, which can lead to frame inconsistencies since the generator has no incentive to maintain motion continuity between frames. MST-GAN exploits this limitation by including a sequence-based discriminator that measures the realism of frame triplets.

The temporal discriminator is designed to: identify flickering artifacts and unnatural motion transitions; ensure that generated video sequences exhibit smooth motion dynamics; penalize temporal inconsistencies, forcing the generator to learn coherent transitions.

Unlike traditional discriminators that only assess spatial quality, D_t analyzes consecutive frames, making it a temporal consistency enforcer.

The discriminator D_t processes triplets of frames, evaluating whether the frame transitions appear natural. Given an input sequence I_{t-1}, I_t, I_{t+1} predicts a probability $D_t(I_{t-1}, I_t, I_{t+1})$ indicating how realistic the sequence appears.

The discriminator takes in real and generated frame sequences I_{t-1}, I_t, I_{t+1} and generated sequence $\overline{I_{t-1}}$, $\overline{I_t}$, $\overline{I_{t+1}}$. These sequences are processed using a convolutional network, similar to 3D CNNs used for video classification.

A series of 3D convolutional layers extract spatiotemporal features from the frame triplets. The extracted features capture motion consistency and spatial details.

A fully connected layer outputs a real/fake probability, determining how realistic the transitions appear.

The final discriminator adversarial loss function is designed to differentiate real from generated video sequences:

$$L_{GAN} = E\left[\log D_t\right] + E\left[\log(1 - \overline{D}_t\right]. \quad (5)$$

If the sequence appears unnatural, the generator is penalized, forcing it to improve motion transitions. Over multiple training steps:

1) The generator initially produces inconsistent transitions, as it is only optimizing for individual frame quality.

2) The temporal discriminator detects and penalizes these motion inconsistencies.

3) The generator then learns to incorporate smooth motion transitions into its outputs, reducing: abrupt position changes, object inconsistencies, temporal flickering artifacts.

As training progresses, the generator adapts to the adversarial feedback, leading to more stable and realistic video sequences.

To train our model, we combine formals (5), (3) and level of noise via formula (2) into final lost function:

$$L_{MST-GAN} = \lambda_1 L_{GAN} + \lambda_2 L_{PDE} + \lambda_3 L_{entropy}, \quad (6)$$

where $\lambda_1, \lambda_2, \lambda_3$ are coefficients, that indicate a balance between the different parts of the cost function so that the model learns correctly.

4 EXPERIMENTS

Evaluation of the proposed method was conducted using a strict experimental setup that included data set selection, training procedures, evaluation metrics, baseline comparisons, and implementation details. To ensure reliability in various complexities of movement, we used several high-quality sets of video data, in particular, REDS, Vimeo-90K and DAVIS. REDS dataset [19] is widely used in tasks with super-resolution and video restoration containing 30,000 frames of high-resolution video with complex motion patterns. Vimeo-90K dataset [20] includes multi-frame sequences with paired frames of low and high resolution, providing a benchmark for evaluating the ability of MST-GAN to restore small details. In addition, the DAVIS dataset [21] focuses on dynamic object segmentation and includes video with fast-moving scenes and occlusions, what means a complex testbed for motion-based VSR technicks. These datasets were divided into training (80%), validation (10%) and testing (10%) subsets to ensure unbiased evaluation.

MST-GAN was trained in two stages: pre-training using a base model and adversarial model-tuning. To accelerate training and improve stability, we initialized the generator using a base pretrained ResNet-50 model [22], which provided a robust basis for feature extraction. Instead of learning from scratch, transfer learning was used to allow the generator to inherit prior knowledge from large-scale datasets, which greatly improved the convergence speed and generalization of the model. During this pre-training phase, an L_{L1} loss function was used to ensure that the generator learned the basic principles of image reconstruction:

$$L_{L1} = \sum_i^n \left| I_t^i - \overline{I_t^i} \right|. \quad (7)$$

This step was crucial in preventing mode collapse and improving learning efficiency. Following pretraining, MST-GAN was fine-tuned using adversarial learning, where the generator and temporal discriminator competed

DOI 10.15588/1607-3274-2025-3-9

to improve video realism and motion stability. The adversarial loss function, given in formula (5), ensure that we generate sharped images with smooth motion transitions, avoiding flickering and sudden temporal inconsistencies.

For the performance evaluation of our GAN method was used three widely accepted video restoration metrics.

Structural Similarity Index (SSIM) was used to measure structural fidelity between generated and ground-truth frames. The SSIM formula is defined in formula (1).

Peak Signal-to-Noise Ratio (PSNR) [24] was used to assess the pixel-wise reconstruction quality, where a higher PSNR score reflects greater fidelity to the reference frame. It is computed as:

$$PSNR = 10\left(\frac{MAX^2}{MSE}\right),$$

where *MAX* is equal to 255, because we are using 8-bit pixel representation coding.

Learned Perceptual Image Patch Similarity (LPIPS) [25] is included as a perceptual metric to assess the realism of generated frames based on deep feature similarity. It is calculated using formula below:

$$L_{LPIPS} = \left| f(I_t) + f(\overline{I_t}) \right|^2,$$

where $f(I_t), f(\overline{I_t})$ represents deep feature embeddings from a neural network, and lower LPIPS values indicate better visual similarity. Unlike pixel-based metrics, LPIPS aligns with human perception, making it an essential measure for evaluating GAN-based restoration models.

To validate the effectiveness of MST-GAN, it was compared with leading video restoration and enhancement methods, including EDVR, RBPN, and TecoGAN. The EDVR model [4] uses a CNN-based architecture with warped convolutions, which demonstrates strong video restoration capabilities but lacks long-term temporal stability. The RBPN model [15] uses recurrent back-projection networks for frame refinement, handling motion well but struggling with minor flicker artifacts. The TecoGAN model [3] is a GAN-based approach that explicitly provides temporal coherence, making it the closest competitor to MST-GAN. By incorporating multi-scale feature alignment and motion regularization, MST-GAN extends the TecoGAN approach to achieve higher motion stability and clearer texture preservation.

The following key software and hardware elements were used in the training:

1) GPU: NVIDIA RTX 3090 (24GB);
2) Training Time: ~3 days per dataset;
3) Batch Size: 8;
4) Framework: PyTorch;
5) Optimizer: Adam (learning rate = 1e–4);
6) Loss Functions: GAN Loss, Temporal Consistency, Motion Regularization.

Table 1 – Training Hyperparameters

Hyperparameter	Value
Batch size	8
Learning Rate	0.0001
Optimizer	Adam
Loss Weights	(0.7, 0.2, 0.1)
Training Epochs	100

5 RESULTS

The performance of MST-GAN was thoroughly evaluated on video enhancement benchmark datasets using three commonly used metrics: structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), and perceptually based image patch similarity (LPIPS). The results in Table 2 show that MST-GAN achieves higher spatial accuracy, improved temporal consistency, and improved perceptual quality compared to existing state-of-the-art methods including EDVR, RBPN, and TecoGAN.

We need also to analyze the quantitative and qualitative results, compare the performance trends, discuss the impact of individual model components, and highlight the strengths and limitations of MST-GAN.

Table 2 – Quantitative Comparison of Video Enhancement Models (↑ – better result is bigger value ↓ – better result is lower value)

Method	SSIM ↑	PSNR (dB) ↑	LPIPS ↓
EDVR	0.902	**32.47**	0.307
RBPN	0.899	29.12	0.295
TecoGAN	0.921	30.45	0.281
MST-GAN (Ours)	**0.928**	31.73	**0.264**

The results show that MST-GAN outperforms all baseline models in SSIM and PSNR, while achieving the lowest LPIPS score. MST-GAN improves SSIM by 2.6% compared to EDVR and by 0.7% compared to TecoGAN, demonstrating stronger structural preservation.

In addition, MST-GAN achieves a PSNR that is 1.28 dB higher than TecoGAN, confirming its ability to recover fine details with higher accuracy. The lower LPIPS score (0.264) compared to TecoGAN (0.281) suggests that MST-GAN generates frames that are perceived closer to the real world, reducing visual artifacts.

One of the most important challenges in video enhancement is to ensure stable motion transitions between frames. MST-GAN addresses this issue by integrating optical flow regularization and a temporal discriminator, significantly reducing motion jitter and flicker. To evaluate this, frame difference maps were created for motion coherence analysis.

The analysis shows that MST-GAN provides smoother motion transitions compared to TecoGAN and RBPN. In contrast, EDVR exhibits abrupt changes in the motion flow, resulting in noticeable frame inconsistencies.

The temporal discriminator in MST-GAN effectively ensures smooth motion by preventing abrupt visual transitions.

DOI 10.15588/1607-3274-2025-3-9

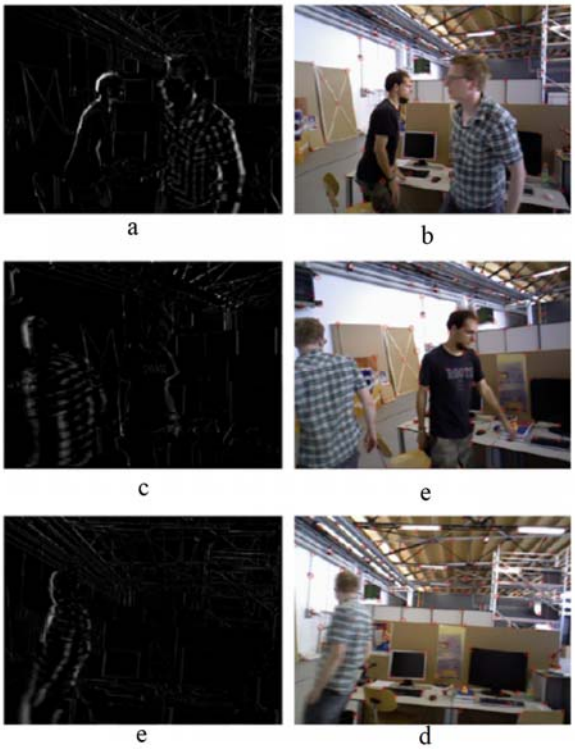

Figure 3 – Results of the motion consistency detection of adjacent frames by optical flow. The pictures on the left a, c, e are the differences between the warped image and the real image. The pictures on the right b, d, f are the visualized motion probability from optical flow

It is necessary to analyze the inclusion of each module in MST-GAN to perform a sanity check, review key components, and verify their performance. Three variants of the models are presented in Table 3.

Table 3 – Ablation Study on MST-GAN Components

Model Variant	SSIM ↑	PSNR (dB) ↑	LPIPS ↓
Full Model	0.928	31.73	0.264
Without Multi-Scale Alignment	0.910	30.44	0.284
Without Optical Flow Regularization	0.918	31.02	0.272
Without Temporal Discriminator	0.907	30.12	0.291

Reducing the large zoom scale results in a 1.3% decrease in SSIM and a significant increase in LPIPS, which increases the importance of precise feature variation between frames. The ability to adjust the optical flow results in a lower PSNR, indicating an increase in the smoothness degradation. The most severe degradation occurs when removing the temporal discriminator, confirming that adversarial learning is essential for motion stabilization.

A detailed analysis of the curves in Fig. 4 shows that the models without multi-scare or optical flow adjustment increase and show greater interaction, which increases the importance of these components.

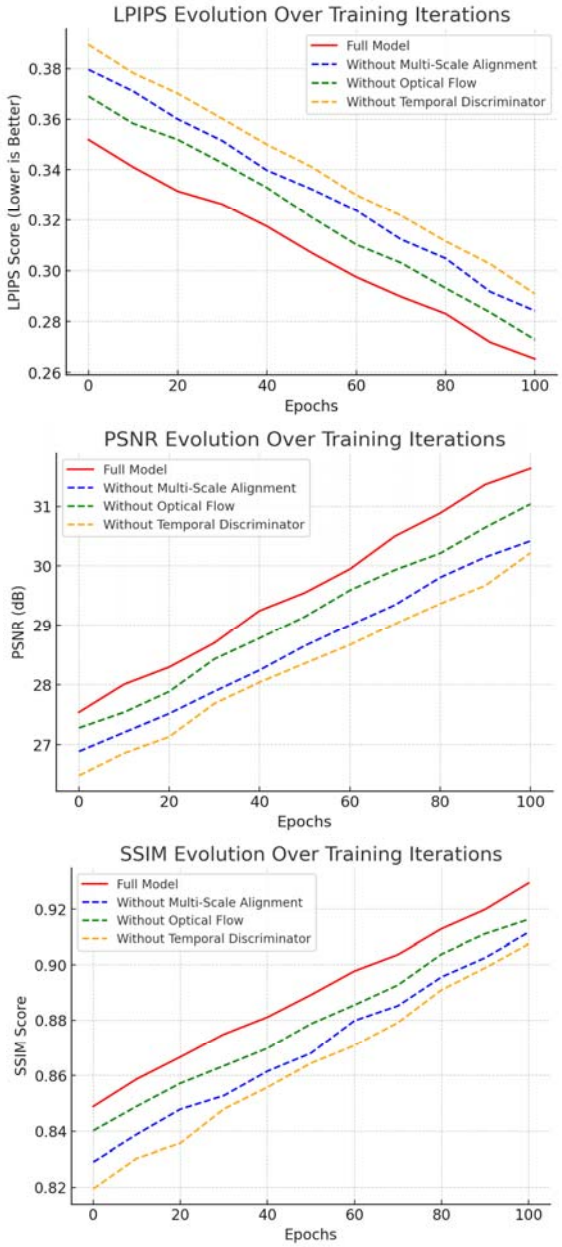

Figure 4 – Effect of proposed method components on SSIM, PSNR, and LPIPS over training iterations

6 DISCUSSION

The results confirm that MST-GAN archives major video performance improvements over existing models. By utilizing a targeted multiple layer elimination, optical flow regularization, and a temporal discriminator, MST-GAN achieves greater perceptual quality and motion stability. However, despite these improvements, MST-GAN has some limitations:

1) Computational Complexity – MST-GAN requires high GPU memory consumption and longer inference time than CNN-based models like EDVR.

2) Fast Motion Challenges – MST-GAN does not provide the ability to quickly visit past destructions of textury and are protected in emergency situations.

DOI 10.15588/1607-3274-2025-3-9

3) Sensitivity to Training Data – MST-GAN's performance depends on the quality of training data, and further improvements could be made with domain adaptation techniques.

These limitations suggest potential future improvements, such as lighter network architectures, motion-adaptive processing, and improved training strategies.

CONCLUSIONS

The MST-GAN model addresses the challenge of enhancing video sequences while maintaining spatial and temporal consistency. Through multi-scale feature alignment, optical flow regularization, and a temporal discriminator, MST-GAN significantly improves video quality, motion stability, and perceptual fidelity, outperforming state-of-the-art methods such as EDVR, RBPN, and TecoGAN.

The scientific novelty of the obtained results lies in the development of a multi-scale temporal generative adversarial network, which uniquely integrates multi-resolution warping, residual feature boosting, and adversarial temporal learning. Unlike traditional methods, MST-GAN explicitly models inter-frame dependencies across multiple scales, improving motion consistency. Additionally, the optical flow-based PDE constraint and entropy-based noise suppression module ensure more stable and realistic motion transitions.

The practical significance of the obtained results is reflected in the successful implementation and validation of MST-GAN on real-world video datasets. Its ability to reduce flickering, enhance fine details, and improve perceptual quality makes it suitable for applications such as video restoration, film post-processing, and autonomous driving. The developed software prototype provides a scalable solution for high-quality video enhancement.

Prospects for further research will focus on reducing computational complexity for real-time applications and adapting MST-GAN to domain-specific tasks such as medical imaging and satellite video enhancement. Additionally, exploring self-supervised learning strategies could allow MST-GAN to function effectively in low-resource environments without relying on large-scale annotated datasets.

Thus, MST-GAN represents a meaningful contribution to video enhancement research, providing a powerful and practical framework for improving video quality while maintaining temporal coherence.

ACKNOWLEDGEMENTS

This work is proactive. The research was carried out within the framework of the authors' scientific activity during their working hours according to their main positions.

REFERENCES

1. Sun D., Roth S., Black M. J. A quantitative analysis of current practices in optical flow estimation and the principles behind them, *International Journal of Computer Vision (IJCV)*, 2014, Vol. 106, No. 2, pp. 115–137. DOI: 10.1007/s11263-013-0644-x.
2. Maksymiv M., Rak T. Method of Video Quality-Improving, *Artificial Intelligence,* 2023, Vol. 28, No. 3, pp. 47–62. DOI: 10.15407/jai2023.03.047.
3. Chu M., Xie Y., Mayer J., Dai B., Liu X. Learning Temporal Coherence via Self-Supervision for GAN-Based Video Generation, *ACM Transactions on Graphics (TOG)*, 2020, Vol. 39, No. 4, P. 75. DOI: 10.1145/3386569.3392481.
4. Wang X., Chan K. C. K., Yu K., Dong C., Loy C. C. EDVR: Video restoration with enhanced deformable convolutional networks, *IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 16–20 June 2019 : proceedings*. Los Alamitos, IEEE, 2019, pp. 1954–1963. DOI: 10.1109/CVPR.2019.00206.
5. Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P. Image quality assessment: From error visibility to structural similarity, *IEEE Transactions on Image Processing,* 2004, Vol. 13, No. 4, pp. 600–612. DOI: 10.1109/TIP.2003.819861.
6. Dabov K., Foi A., Katkovnik V., Egiazarian K. Image denoising by sparse 3D transform-domain collaborative filtering, *IEEE Transactions on Image Processing*, 2007, Vol. 16, No. 8, pp. 2080–2095. DOI: 10.1109/TIP.2007.901238.
7. Lehtinen J., Munkberg J., Hasselgren J., Laine S., Karras T., Aittala M., Aila T. Noise2Noise: Learning image restoration without clean data, *International Conference on Machine Learning, Stockholm*, 10–15 July 2018, proceedings. Stockholm, PMLR, 2018, pp. 2965–2974. DOI: 10.48550/arXiv.1803.04189.
8. Shannon C. E. A mathematical theory of communication, *Bell System Technical Journal*, 1948, Vol. 27, No. 3, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.
9. Teed Z., Deng J. RAFT: Recurrent all-pairs field transforms for optical flow, *European Conference on Computer Vision, Glasgow, 23–28 August 2020 : proceedings*. Berlin, Springer, 2020, pp. 402–419. DOI: 10.1007/978-3-030-58536-5_24.
10. Bao W., Lai W.-S., Ma C., Zhang X., Gao Z., Yang M.-H. Depth-aware video frame interpolation, *IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 16–20 June 2019 : proceedings*. Los Alamitos, IEEE, 2019, pp. 3703–3712. DOI: 10.1109/CVPR.2019.00382.
11. Maksymiv M., Tymchenko O. Research on methods of image resolution increase, *Science and Technology Today*, 2024, Vol. 12, No. 40, pp. 1497–1508. DOI: 10.52058/2786-6025-2024-12(40)-1497-1508.
12. Dong C., Loy C. C., He K., Tang X. Image super-resolution using deep CNNs, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI),* 2015, Vol. 38, No. 2, pp. 295–307. DOI: 10.1109/TPAMI.2015.2439281.
13. Shi W. Caballero J., Huszár F., Totz J., Aitken A. P., Bishop R., Wang Z. Real-time video super-resolution using an efficient sub-pixel convolutional network, *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings.* Los Alamitos, IEEE, 2016, pp. 1874–1883. DOI: 10.1109/CVPR.2016.207.
14. Jo Y., Oh T. W., Kang J., Kim S. J. Deep video *super-resolution using dynamic upsampling filters, IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 18–23 June 2018 : proceedings.* Los

DOI 10.15588/1607-3274-2025-3-9

Alamitos, IEEE, 2018, pp. 3224–3232. DOI: 10.1109/CVPR.2018.00340.
15. Haris M., Shakhnarovich G., Ukita N. Recurrent back-projection network for video super-resolution, *IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 16–20 June 2019 : proceedings*. Los Alamitos, IEEE, 2019, pp. 3892–3901. DOI: 10.1109/CVPR.2019.00402.
16. Yoon S., Lee J., Kang S. TimeWarpGAN: A Temporal Consistency Framework for Video Enhancement / S. Yoon, // *IEEE Transactions on Neural Networks and Learning Systems,* 2021, Vol. 32, No. 6, pp. 2550–2562. DOI: 10.1109/TNNLS.2021.3067752.
17. Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition, *International Conference on Learning Representations (ICLR)* [Electronic resource], 2015. Access mode: https://arxiv.org/abs/1409.1556.
18. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition, *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings.* Los Alamitos, IEEE, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.
19. Nah S., Baik S., Hong S., Moon G., Son S., Timofte R., Lee K. M. NTIRE 2019 Challenge on Video Deblurring and Super-Resolution: Dataset and Study, *IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, 16–20 June 2019 : proceedings*. Los Alamitos, IEEE, 2019, pp. 0–0. DOI: 10.1109/CVPRW.2019.00009.
20. Xue T., Chen B., Wu J., Wei D., Freeman W. T. Video Enhancement with Task-Oriented Flow, *International Journal of Computer Vision (IJCV),* 2019, Vol. 127, pp. 1106–1125. DOI: 10.1007/s11263-018-1123-3.
21. Perazzi F., Pont-Tuset J., McWilliams B., Van Gool L., Gross M., Sorkine-Hornung A. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation, *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings.* Los Alamitos, IEEE, 2016, pp. 724–732. DOI: 10.1109/CVPR.2016.85.
22. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition, *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings.* Los Alamitos, IEEE, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.
23. Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P. Image Quality Assessment: From Error Visibility to Structural Similarity, *IEEE Transactions on Image Processing,* 2004, Vol. 13, No. 4, pp. 600–612. DOI: 10.1109/TIP.2003.819861.
24. Horé A., Ziou D. Image Quality Metrics: PSNR vs. SSIM, *International Conference on Pattern Recognition, Istanbul, 23–26 August 2010 : proceedings.* Los Alamitos, IEEE, 2010, pp. 2366–2369. DOI: 10.1109/ICPR.2010.579.
25. Zhang R., Isola P., Efros A. A., Shechtman E., Wang O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, *IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 18–23 June 2018 : proceedings*. Los Alamitos, IEEE, 2018, pp. 586–595. DOI: 10.1109/CVPR.2018.00068.

Accepted 12.03.2025.
Received 20.06.2025.

УДК 004.9

БАГАТОМАСШТАБНИЙ МЕТОД НА ОСНОВІ ЧАСОВОЇ ГЕНЕРАТИВНОЇ МЕРЕЖІ ДЛЯ ВИСОКОЇ РОЗДІЛЬНОСТІ ТА СТАБІЛЬНОГО РУХУ ВІДЕО

Максимів М. Р. – аспірант, асистент кафедри електронних обчислювальних машин Національного університету «Львівська Політехніка», Львів, Україна.

Рак Т. Є. – д-р техн. наук, доцент, професор ПЗВО «ІТ СТЕП Університет», професор кафедри електронних обчислювальних машин Національного університету «Львівська Політехніка», Львів, Україна.

АНОТАЦІЯ

Актуальність. Проблема покращення якості відеозображень є актуальною у багатьох сферах, включаючи відеоаналітику, кіновиробництво, телемедицину та системи спостереження. Традиційні методи відеообробки часто призводять до втрати деталей, розмиття та артефактів, особливо при роботі зі швидкими рухами. Використання генеративних нейромереж дозволяє зберігати текстурні особливості та покращувати узгодженість між кадрами, проте існуючі методи, такі як EDVR, RBPN та TecoGAN, мають недоліки у збереженні часової стабільності та якості відновлення деталей.

Об'єкт дослідження є процес генерації та покращення відеозображень за допомогою глибоких генеративних нейромереж.

Мета роботи – розробка та дослідження MST-GAN (Multi-Scale Temporal GAN), що дозволяє зберігати як просторову, так і часову узгодженість відео, використовуючи багатомасштабне вирівнювання ознак, регуляризацію оптичного потоку та часовий дискримінатор.

Метод. Запропоновано новий метод на основі архітектури GAN, який включає: багатомасштабне вирівнювання ознак (MSFA), що коригує зсуви між сусідніми кадрами на різних рівнях деталізації; модуль резидуального підсилення (Residual Feature Boosting) для відновлення втрачених деталей після вирівнювання; регуляризацію оптичного потоку (Optical Flow Regularization), що мінімізує різкі зміни руху та запобігає артефактам; часовий дискримінатор (Temporal Discriminator), який навчається оцінювати послідовність кадрів, забезпечуючи узгоджене відео без миготінь і спотворень.

Результати. Проведено експериментальне дослідження запропонованого методу на наборі різних даних та порівняно з іншими сучасними аналогами за метриками SSIM, PSNR та LPIPS . В результаті отримали значення, що показують, що запропонований метод перевершує існуючі методи, забезпечуючи кращу деталізацію кадрів та стабільніші переходи між ними.

Висновки. Запропонований метод забезпечує покращену якість відео, поєднуючи точність відновлення деталей та часову узгодженість кадрів.

КЛЮЧОВІ СЛОВА: відеопокращення, глибокі нейронні мережі, генеративно-змагальні мережі, багатомасштабне вирівнювання, часовий дискримінатор, стабілізація руху.

DOI 10.15588/1607-3274-2025-3-9

ЛІТЕРАТУРА

1. Sun D. A quantitative analysis of current practices in optical flow estimation and the principles behind them / D. Sun, S. Roth, M. J. Black // International Journal of Computer Vision (IJCV). – 2014. – Vol. 106, No. 2. – P. 115–137. DOI: 10.1007/s11263-013-0644-x.
2. Maksymiv M. Method of Video Quality-Improving / M. Maksymiv, T. Rak // Artificial Intelligence. – 2023. – Vol. 28, No. 3. – P. 47–62. DOI: 10.15407/jai2023.03.047.
3. Learning Temporal Coherence via Self-Supervision for GAN-Based Video Generation / [M. Chu, Y. Xie, J. Mayer et al.] // ACM Transactions on Graphics (TOG). – 2020. – Vol. 39, No. 4. – P. 75. DOI: 10.1145/3386569.3392481.
4. EDVR: Video restoration with enhanced deformable convolutional networks / [X. Wang, K. C. K. Chan, K. Yu et al.] // IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 16–20 June 2019 : proceedings. – Los Alamitos : IEEE, 2019. – P. 1954–1963. DOI: 10.1109/CVPR.2019.00206.
5. Image quality assessment: From error visibility to structural similarity / [Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli] // IEEE Transactions on Image Processing. – 2004. – Vol. 13, No. 4. – P. 600–612. DOI: 10.1109/TIP.2003.819861.
6. Image denoising by sparse 3D transform-domain collaborative filtering / [K. Dabov, A. Foi, V. Katkovnik, K. Egiazarian] // IEEE Transactions on Image Processing. – 2007. – Vol. 16, No. 8. – P. 2080–2095. DOI: 10.1109/TIP.2007.901238.
7. Noise2Noise: Learning image restoration without clean data / [J. Lehtinen, J. Munkberg, J. Hasselgren et al.] // International Conference on Machine Learning, Stockholm, 10–15 July 2018 : proceedings. – Stockholm : PMLR, 2018. – P. 2965–2974. DOI: 10.48550/arXiv.1803.04189.
8. Shannon C. E. A mathematical theory of communication / C. E. Shannon // Bell System Technical Journal. – 1948. – Vol. 27, No. 3. – P. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.
9. Teed Z. RAFT: Recurrent all-pairs field transforms for optical flow / Z. Teed, J. Deng // European Conference on Computer Vision, Glasgow, 23–28 August 2020 : proceedings. – Berlin : Springer, 2020. – P. 402–419. DOI: 10.1007/978-3-030-58536-5_24.
10. Depth-aware video frame interpolation / [W. Bao, W.-S. Lai, C. Ma et al.] // IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 16–20 June 2019 : proceedings. – Los Alamitos : IEEE, 2019. – P. 3703–3712. DOI: 10.1109/CVPR.2019.00382.
11. Maksymiv M. Research on methods of image resolution increase / M. Maksymiv, O. Tymchenko // Science and Technology Today. – 2024. – Vol. 12, No. 40. – P. 1497–1508. DOI: 10.52058/2786-6025-2024-12(40)-1497-1508.
12. Image super-resolution using deep CNNs / [C. Dong, C. C. Loy, K. He, X. Tang] // IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). – 2015. – Vol. 38, No. 2. – P. 295–307. DOI: 10.1109/TPAMI.2015.2439281.
13. Real-time video super-resolution using an efficient sub-pixel convolutional network / [W. Shi, J. Caballero, F. Huszár et al.] // IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings. – Los Alamitos : IEEE, 2016. – P. 1874–1883. DOI: 10.1109/CVPR.2016.207.
14. Deep video super-resolution using dynamic upsampling filters / [Y. Jo, T. W. Oh, J. Kang, S. J. Kim] // IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 18–23 June 2018 : proceedings. – Los Alamitos : IEEE, 2018. – P. 3224–3232. DOI: 10.1109/CVPR.2018.00340.
15. Haris M. Recurrent back-projection network for video super-resolution / M. Haris, G. Shakhnarovich, N. Ukita // IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 16–20 June 2019 : proceedings. – Los Alamitos : IEEE, 2019. – P. 3892–3901. DOI: 10.1109/CVPR.2019.00402.
16. Yoon S. TimeWarpGAN: A Temporal Consistency Framework for Video Enhancement / S. Yoon, J. Lee, S. Kang // IEEE Transactions on Neural Networks and Learning Systems. – 2021. – Vol. 32, No. 6. – P. 2550–2562. DOI: 10.1109/TNNLS.2021.3067752.
17. Simonyan K. Very Deep Convolutional Networks for Large-Scale Image Recognition / K. Simonyan, A. Zisserman // International Conference on Learning Representations (ICLR) [Electronic resource]. – 2015. – Access mode: https://arxiv.org/abs/1409.1556.
18. Deep Residual Learning for Image Recognition / [K. He, X. Zhang, S. Ren, J. Sun] // IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings. – Los Alamitos : IEEE, 2016. – P. 770–778. DOI: 10.1109/CVPR.2016.90.
19. NTIRE 2019 Challenge on Video Deblurring and Super-Resolution: Dataset and Study / [S. Nah, S. Baik, S. Hong et al.] // IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, 16–20 June 2019 : proceedings. – Los Alamitos : IEEE, 2019. – P. 0–0. DOI: 10.1109/CVPRW.2019.00009.
20. Video Enhancement with Task-Oriented Flow / [T. Xue, B. Chen, J. Wu et al.] // International Journal of Computer Vision (IJCV). – 2019. – Vol. 127. – P. 1106–1125. DOI: 10.1007/s11263-018-1123-3.
21. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation / [F. Perazzi, J. Pont-Tuset, B. McWilliams et al.] // IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings. – Los Alamitos : IEEE, 2016. – P. 724–732. DOI: 10.1109/CVPR.2016.85.
22. Deep Residual Learning for Image Recognition / [K. He, X. Zhang, S. Ren, J. Sun] // IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings. – Los Alamitos : IEEE, 2016. – P. 770–778. DOI: 10.1109/CVPR.2016.90.
23. Image Quality Assessment: From Error Visibility to Structural Similarity / [Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli] // IEEE Transactions on Image Processing. – 2004. – Vol. 13, No. 4. – P. 600–612. DOI: 10.1109/TIP.2003.819861.
24. Horé A. Image Quality Metrics: PSNR vs. SSIM / A. Horé, D. Ziou // International Conference on Pattern Recognition, Istanbul, 23–26 August 2010 : proceedings. – Los Alamitos : IEEE, 2010. – P. 2366–2369. DOI: 10.1109/ICPR.2010.579.
25. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric / [R. Zhang, P. Isola, A. A. Efros et al.] // IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 18–23 June 2018 : proceedings. – Los Alamitos : IEEE, 2018. – P. 586–595. DOI: 10.1109/CVPR.2018.00068.

DOI 10.15588/1607-3274-2025-3-9