Date of publication DEC-31, 2025, date of current version DEC-15, 2025.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.24.4.4341

Restoring Quality from Bitrate Collapse: A Two-Stage GAN for Enhancing Heavily Compressed Video

MYKOLA MAKSYMIV[1], TARAS RAK[2]
[1,2]Lviv Polytechnic National University, 12, Bandera Str, Lviv, 79013, Ukraine
[2]IT STEP University, 83a, Zamarstynivska Str, Lviv, 79019, Ukraine

Corresponding author: Mykola Maksymiv (e-mail: mykolamaksymivua@gmail.com).

ABSTRACT Low-bitrate video compression (e.g., H.264/AVC at ≤300 Kbps) typically introduces visible artifacts such as blocking, blurring, and texture loss. This paper proposes a two-stage Generative Adversarial Network (GAN) architecture tailored to restore visual quality in degraded video sequences. The system incorporates motion alignment, residual blocks with attention mechanisms, and multi-frame temporal modeling to enhance spatial fidelity and consistency. A novel training dataset is constructed by synthetically compressing high-quality video content to simulate real-world degradation. We analyze the architecture in detail, discuss training stability (including mode collapse mitigation), and propose a combination of distortion and perceptual losses, including L1, SSIM, LPIPS, and adversarial objectives. Quantitative evaluation on standard benchmarks shows that the proposed model achieves competitive or better performance compared to earlier methods like ESRGAN, EDVR, CVEGAN, and traditional deblocking techniques. We further present visual comparisons, ablation studies, and training dynamics to validate each architectural component. The enhanced frames exhibit restored detail and consistent temporal structure across sequences. A key novelty lies in targeting extremely compressed content and demonstrating restoration capability under these constraints. This makes the approach suitable for scenarios such as cloud video storage or ultra-low-bandwidth transmission, where post-decompression enhancement is crucial.

KEYWORDS Video enhancement, compression artifact removal, GANs for video restoration, low-bitrate video, temporal consistency, perceptual quality metrics, deep learning for post-processing.

I. INTRODUCTION

Video streaming under low bandwidth conditions remains a significant challenge for visual quality [1, 2]. Modern codecs, such as H.264 or HEVC, compress video aggressively at low bitrates, resulting in visible artifacts, including blocking, blurring, and loss of fine detail [3, 4]. These distortions are especially pronounced in dynamic scenes, textures, and edges, where temporal and spatial coherence is often degraded.

We propose a GAN-based video enhancement framework to address this issue to restore perceptual quality from heavily compressed video streams. Unlike traditional methods that focus solely on distortion minimization [4-6], our approach prioritizes visual fidelity and temporal consistency. The model uses a two-stage generator architecture that first reconstructs coarse structure and then refines it with high-frequency details. A dual-discriminator system further encourages spatial realism and temporal stability.

Our method is tailored for scenarios where only low-quality video is available at the client side, such as mobile streaming, edge computing, or storage-constrained playback. Improving video post-decoding allows perceptual recovery without altering the encoder pipeline. This significantly reduces bandwidth needs while keeping high visual quality.

In this paper, we define "bitrate collapse" as a compression scenario in which the encoding bitrate is so low that fine details and essential structural information are severely degraded or lost, resulting in heavy blockiness, blurring, and perceptual disintegration of the scene.

Unlike most prior works that target mild compression artifacts or moderate bitrate streams, this work explicitly focuses on restoring extremely degraded video under severe bitrate constraints ("bitrate collapse" conditions) [7].

To our knowledge, no previous method in open-access literature has systematically addressed quality restoration from such aggressively compressed sources using a two-stage GAN framework with explicit temporal and spatial fidelity objectives.

II. RELATED WORK

A broad range of research has been conducted on image and

video restoration, with several families of methods emerging over the last decade.

For image super-resolution, GAN-based techniques such as SRGAN [8] and ESRGAN [9] introduced adversarial training to generate perceptually realistic details. ESRGAN extended SRGAN by incorporating Residual-in-Residual Dense Blocks (RRDBs) and a Relativistic GAN loss, yielding superior visual quality on single-frame tasks. However, while effective for still images, these methods often introduce temporal flicker when applied frame by frame to videos.

To address temporal coherence in video, TecoGAN [10] introduced a recurrent generator with optical flow alignment and a temporal discriminator, achieving stable results across frames. Other multi-frame architectures, such as FRVSR [11] and EDVR [12], utilize temporal alignment and deformable convolutions, respectively, to fuse information from neighboring frames and enhance temporal and spatial fidelity. These approaches showed that aggregating context over time significantly improves both perceived and measured quality in video restoration tasks.

When specifically addressing compression artifacts, particularly in low-bitrate scenarios, Multi-frame Quality Enhancement (MFQE) [13, 14] utilizes high-quality "peak" frames (e.g., I-frames) to guide the enhancement of lower-quality inter frames (P-frames), leveraging codec structure and temporal redundancy. MFQE 2.0 improved upon its predecessor by incorporating a deeper CNN and bi-directional recurrent fusion, enabling more effective restoration across entire video sequences.

Focusing on post-compression enhancement, models such as CVEGAN [15] and SUPERVEGAN [16] were designed to enhance video after decoding from strongly compressed formats like HEVC or H.264. CVEGAN integrates multi-scale residual blocks with attention mechanisms and a perceptually-driven loss function to improve subjective quality, particularly for low-bitrate streams. SUPERVEGAN adopts a two-stage GAN architecture where the first stage handles distortion and upscaling and the second performs perceptual refinement training both stages progressively to avoid instability and mode collapse.

Additionally, foundational GAN formulations and discriminator designs have contributed to the perceptual restoration of video. Notably, the relativistic discriminator [17] was shown to improve realism and stabilize training in high-frequency detail generation, especially in video enhancement pipelines that rely on adversarial learning.

Despite these advancements, most methods assume moderate degradation or focus on specific codec settings. In contrast, the approach proposed in this paper is explicitly designed for severe compression scenarios, aiming to recover texture, structure, and perceptual clarity while maintaining temporal consistency. By incorporating elements from super-resolution, video restoration, and perceptual GAN training, our work bridges a critical gap in the domain of real-world low-bitrate video enhancement.

III. PROPOSED METHOD

The proposed method employs a two-stage generator G and a multi-component loss function within a GAN framework to transform low-quality compressed video X into high-quality output $\hat{Y}_t$.

Fig. 1 provides a block diagram of the architecture. The design is inspired by human expert restoration: first, perform conservative reconstruction to remove artifacts and recover details (Stage A), then apply a refinement that injects realistic textures (Stage B) without disturbing temporal coherence. The generator G thus comprises two sub-networks, G_A and G_B, corresponding to Stage A and Stage B. We formulate the overall enhancement for frame t as:

$$Y_t^{\text{final}} = G_B\big(G_A(X_{t-N:t+N})\big), \tag{1}$$

where $X_{t-N:t+N}$ denotes a window of 2N+1 input frames (frame t and its N neighbors on each side). Multi-frame input allows G_A to aggregate information from neighboring frames to restore details that single-frame *that* X_t cannot provide on its own. In our experiments, we use N=2 (5-frame input) for a good tradeoff between temporal context and model complexity, though the architecture supports larger temporal windows.

A. MOTION ALIGNMENT MODULE

To effectively merge frames, we include an explicit alignment module based on deformable convolution and/or optical flow. Given that consecutive frames often contain object motion or camera panning, direct frame stacking can misalign details. We adopt a Pyramid, Cascading and Deformable (PCD) alignment module similar to EDVR [12], and draw inspiration from early flow-based learning frameworks like FlowNet [18]. This module refines estimated flow at multiple scales and uses deformable convolution to sample aligned features, handling complex motion and occlusions. The result is a stack of feature maps $F_{t-i\to t}$ all warped to the reference frame t. We denote the alignment operation as:

$$F_{t-i\to t} = \text{Align}(X_{t-i}, X_t), \tag{2}$$

for $i \in [-N, N]$ producing aligned features for each neighbor relative to frame t (with $F_{t\to t}$ being just X_t is initial features). These aligned features are concatenated along the channel dimension and fed into Stage A. By performing learnable alignment, G_A it receives information such as the texture on a static background from a nearby higher-quality frame (e.g., a P-frame aided by an I-frame).

B. STAGE A: RECONSTRUCTION NETWORK

Stage A focuses on distortion reduction. It uses a series of Residual blocks to remove artifacts and reconstruct an initial high-quality frame $\widehat{Y_t^A}$ at the target resolution (which could be the same as input or higher). We utilize a residual learning strategy: Stage A predicts a residual image R_t^A that, when added to an upsampled or base image, yields the output. Two modes are supported:

(a) Post-Processing (PP) is identical resolution, only artifacts removed:

$$\widehat{Y_t^A} = X_t + R_t^A. \tag{3}$$

(b) Super-Resolution Adaptation (SRA): input is upsampled by factor s:

$$\widehat{Y_t^A} = X_t^{\uparrow s} + R_t^A. \tag{4}$$

Internally, Stage A's architecture stacks several Residual-in-Residual Dense Blocks (RRDB) as used in ESRGAN [9], but modified with attention mechanisms. In particular, we integrate an Enhanced Convolutional Block Attention Module (ECBAM) as proposed in CVEGAN [15]. ECBAM applies sequential channel and spatial attention to intermediate features, enabling the network to focus on regions with noticeable artifacts (e.g., block boundaries or blurry textures) and allocate more capacity to correcting them. This is especially beneficial in heavy compression scenarios where artifacts are spatially localized.

Stage A is trained with pixel-wise loss only (no GAN loss at this stage), to ensure R_t^A learns a safe correction and avoids introducing new artifacts. We use a combination of L1 loss and MS-SSIM loss $\widehat{Y_t^A}$ versus the ground-truth frame Y_t:

$$L_{\text{A, pix}} = |\widehat{Y_t^A} - Y_t|_1 + \lambda_{\text{ssim}} \left(1 - \text{SSIM}\left(\widehat{Y_t^A}, Y_t\right)\right). \quad (5)$$

Minimizing $L_{\text{A, pix}}$ encourages high PSNR/SSIM and removes most glaring compression artifacts. Notably, Stage A does not hallucinate details, it is analogous to a multi-frame denoiser/upscaler, constrained to produce an MSE-optimal reconstruction. This provides a strong, consistent foundation for the adversarial Stage B.

C. STAGE B: DETAIL SYNTHESIS NETWORK

Stage B takes $\widehat{Y_t^A}$ as input and enhances it to produce the final output $\hat{Y}_t$. Stage B generates realistic textures and recovering fine details that Stage A (trained on MSE) might have smoothed out. Its architecture can be a deeper or alternate set of residual blocks, potentially at full resolution.

We include a 1-level U-Net structure in Stage B (as in SUPERVEGAN [16]) to increase receptive field the U-Net encoder-decoder allows the network to gather global context (important for large smooth regions or consistent textures) and then refine details through skip connections [19-20]. Stage B outputs a residual R_t^B which is added to $\widehat{Y_t^A}$:

$$\hat{Y}_t = \widehat{Y_t^A} + R_t^B. \quad (6)$$

This formulation (often called a residual GAN approach) lets Stage B focus on high-frequency components (like film grain, skin details, text clarity) without altering the overall structure or colors established by Stage A. By limitin g R_t^B to smaller amplitude high-frequency signals, we reduce the risk of Stage B introducing distortions that break consistency with the input content.

To train Stage B, we activate adversarial and perceptual losses. A spatial discriminator D_S judges the realism of individual enhanced frames $\hat{Y}_t$ compared to original high-quality frames Y_t , while a temporal discriminator D_t looks at sequences of frames (we use three cothreesecutive frames as D_t's input) to judge temporal coherence. The adversarial loss for Stage B is the sum of a GAN loss from D_S and D_t. We use a relativistic average GAN loss formulation to stabilize training [11], meaning D_S doesn't just classify real vs fake, but also considers the difference in realism between real and fake data, and G is trained to increase the probability that outputs are more realistic than real data on average.

Formally, for D_S:

$$L_{\text{adv}}^{(D_S)} = -E_Y\left[\log\left(1 - D_S(Y)\right)\right] - E_{\hat{Y}}\left[\log\left(D_S(\hat{Y})\right)\right], \quad (7)$$

and for the generator (Stage B):

$$L_{\text{adv}}^{(G)} = -E_{\hat{Y}}\left[\log\left(1 - D_S(\hat{Y})\right)\right] - E_Y\left[\log\left(D_S(Y)\right)\right], \quad (8)$$

and similarly for D_t with sequences. (For brevity we do

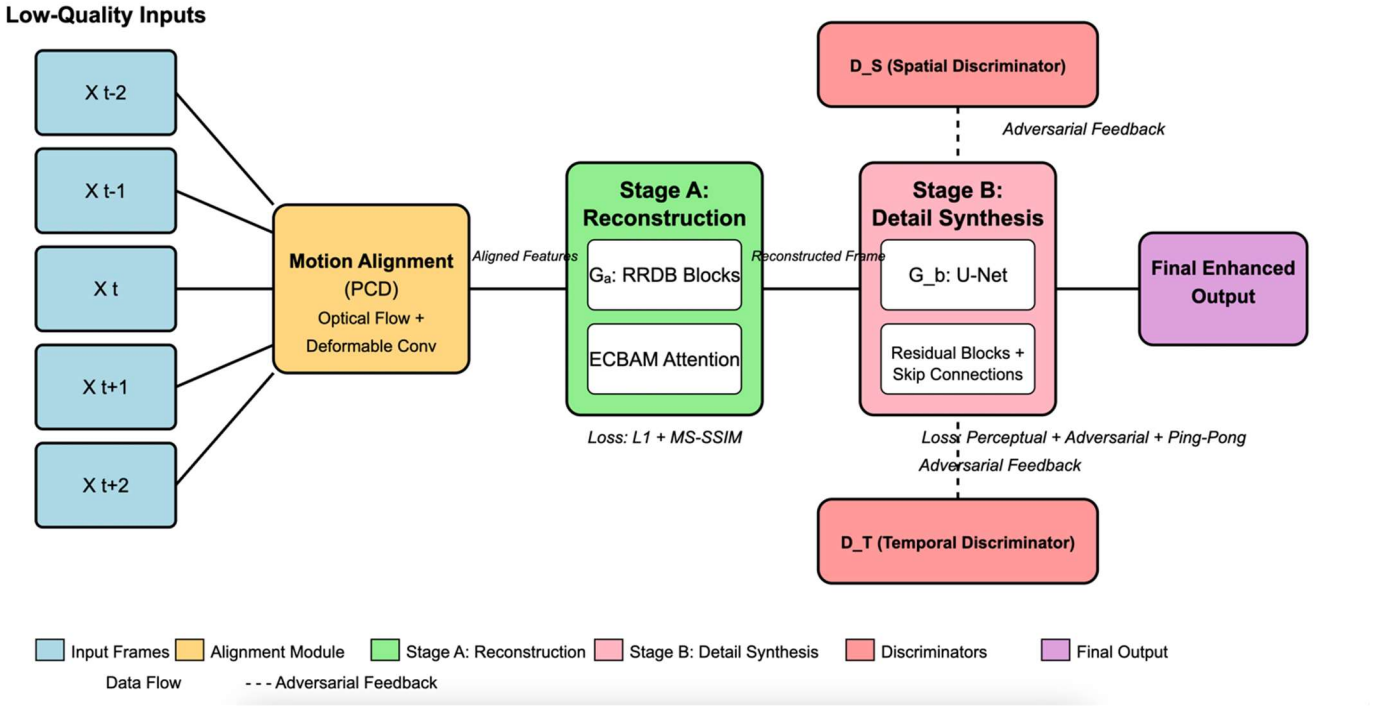

Figure 1. Architectural overview of the proposed GAN-based video enhancement method.

not expand the temporal case here; it follows the same relativistic principle applied to short-frame sequences.)

In addition to adversarial loss, we use a perceptual loss L_{perc} computed as the feature space difference between $\widehat{Y}_t$ and Y_t using a pretrained image classification network (VGG-19 [1]). Specifically:

$$L_{\text{perc}} = \sum_j \frac{1}{C_j H_j W_j} |\phi_j(\widehat{Y}_t) - \phi_j(Y_t)|_2^2. \quad (9)$$

In formula 9 ϕ_j is a function that extracts activations (feature maps) from the jth layer of a pre-trained neural network. This means that we feed both images to VGG. We take the output tensors on certain layers (for example, relu3_4, relu4_4). $C_j H_j W_j$ - the number of channels, height, and width of the corresponding feature map for the layer. This encourages $\widehat{Y}_t$ having similar texture and feature responses as the ground truth, which correlates better with human perception than pure MSE.

To further maintain temporal consistency, we incorporate a Ping-Pong loss L_{pp} [10]. This works as follows: we feed a sequence of frames $[X_{t-1}, X_t, X_{t+1}]$ through the generator to get $[\widehat{Y_{t-1}}, \widehat{Y_t}, \widehat{Y_{t+1}}]$. Then we take $\widehat{Y_{t+1}}$ and feed it backwards (as if it were an input at t-1) along with X_t and X_{t-1}, obtaining a reconstruction of $\widetilde{Y}_t$ (the middle frame when the sequence is processed in reverse).

The Ping-Pong loss is defined as the *L2* difference between the original forward $\widehat{Y}_t$and the backward $\widetilde{Y}_t$:

$$L_{\text{pp}} = |\widehat{Y}_t - \widetilde{Y}_t|_2^2. \quad (10)$$

Minimizing L_{pp} forces the generator to produce frasistent frames whether time is flowing forward or backward, effectively reducing flickering and spurious detail changes over time. Unlike optical-flow-based temporal loss, Ping-Pong does not rely on external motion estimation, making it well-suited for GAN training where generated frames lack a one-to-one ground-truth optical flow.

The total loss for Stage B (generator) is a weighted sum of these components:

$$L_{\text{B,total}} = \lambda_{\text{adv}}\left(L_{\text{adv},S}^{(G)} + L_{\text{adv},T}^{(G)}\right) + \lambda_{\text{perc}} L_{\text{perc}} + \lambda_{\text{pp}} L_{\text{pp}} + \lambda_{\text{pix}} |\widehat{Y}_t - Y_t|_1. \quad (11)$$

We still keep a small weight on pixel loss (last term) for Stage B to prevent it from deviating too far (this is especially needed for areas where ground truth has very low detail, to avoid hallucinating something obviously incorrect). In practice, we set $\lambda_{\text{adv}} = 10^{-3}$ (since adversarial losses are higher in scale) $\lambda_{\text{perc}} = 1$, $\lambda_{\text{pp}} = 1$, $\lambda_{\text{pix}} = 1$ based on validation tuning.

D. TRAINING STRATEGY

We train in two phases similar to SUPERVEGAN's progressive training [16]. In Phase 1, we train Stage A alone by minimizing $L_{\text{A, pix}}$, using a standard L1+SSIM target. This phase lasts for T_1 iterations (until convergence in distortion metrics).

Next, in Phase 2, we fix Stage A (or fine-tune it at a very low learning rate) and train Stage B with the full loss. Initially, we set $\lambda_{\text{adv}} = 0$ to warm up Stage B with just perceptual and pixel losses for a short period, then gradually increase λ_{adv} to its full value over a number of epochs. This gradual introduction of the GAN prevents the sudden destabilization of the two-stage generator.

The discriminators D_S and D_t are trained in tandem with Stage B as usual in GAN training (one or a few D updates per G update). By Phase 2's end, Stage B is generating realistic textures and D_S, D_t can no longer distinguish most enhanced frames from true ones.

While we do not directly optimize for VMAF due to its non-differentiability, we evaluate our outputs using this perceptual metric to reflect visual quality in streaming scenarios better. Prior work has shown VMAF's strong correlation with user preference in bitrate-limited video [2, 21-22].

Finally, we optionally fine-tune the entire generator (both Stage A and B together) with a low learning rate and all losses active, to recover any slight regressions in Stage A outputs caused by fixing it during Stage B training.

IV. EXPERIMENTS AND RESULTS

All experiments were conducted on compressed video sequences at 720p resolution. Inputs were downsampled to 360p, compressed at 200–300 Kbps using H.264 codec (x264, veryfast preset), and then upscaled back to 720p using bicubic interpolation before feeding into the enhancement network. Our model was evaluated using a 5-frame window (N=2), with no external optical flow supervision.

A. DATASETS

We evaluate our method on standard video datasets widely adopted in prior enhancement research. For training, we compiled a diverse dataset comprising: the Vimeo-90K septuplet dataset (used extensively for video super-resolution); the MFQE 2.0 dataset, which provides raw-compressed video pairs [13], and selected scenes from LIVE-NFLX II [2, 21], a publicly released perceptual video quality dataset by Netflix.

High-quality source videos were synthetically degraded via heavy compression to simulate realistic low-bitrate streaming scenarios. Specifically, we applied H.264 compression using FFmpeg's x264 encoder at very low bitrates. The settings included CRF = 38 and spatial downsampling to 50% of the original resolution, producing outputs at ~200–300 Kbps (540p from 1080p). Keyframes were sparsely inserted (intra-period = 100) to emulate long GOPs typical in streaming codecs. This yielded highly compressed training pairs with severe blocking, blurring, and loss of detail.

For evaluation, we used sequences from the animated short films Big Buck Bunny (frames "Bird" 432-434, "Bunny" 1168–1171) and Sintel (final render version), both known for complex textures, motion, and lighting. We generated test samples at 250 and 500 Kbps. In addition, we tested on the Vid4 benchmark with added compression. The ground truth is the uncompressed original, and the input is the degraded compressed video.

B. IMPLEMENTATION DETAILS

The alignment module consists of a 3-level deformable convolution pyramid with 32 channels at the coarsest level and 64 at the finest.

Stage A includes 30 RRDB blocks (each using 64 channels and integrated channel attention via ECBAM).

Stage B includes a shallow U-Net (with 2× spatial downsampling and 64 base filters) and 10 residual blocks.

We use the Adam optimizer with separate learning rates for each stage: $2 * 10^{-4}$ for Stage A; $1 * 10^{-4}$ for Stage B. Both with cosine annealing decay.

The spatial discriminator is a PatchGAN-based model (70×70 patches) applied to full frames (1280×720), and the temporal discriminator operates on concatenated 3-frame sequences.

Training was conducted in two phases:

- Phase 1 (Stage A only): 200,000 iterations
- Phase 2 (full model): 100,000 additional iterations

We used 2× NVIDIA V100 GPUs, with a batch size of 8 and a 5-frame input window. Total training time was approximately 4 days. The implementation was done in PyTorch, and our code will be made publicly available.

C. EVALUATION METRICS AND QUANTITATIVE RESULTS

We evaluate our model using both distortion-based and perceptual metrics. Specifically, we report Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) to assess fidelity, and LPIPS (Learned Perceptual Image Patch Similarity) to evaluate perceptual closeness to the ground truth (lower is better). Additionally, we compute the Bjøntegaard Delta rate (BD-rate) to estimate bitrate savings at equal quality.

We compare our approach with the following baselines:

- H.264 Compressed input of the raw, low-quality decoded video.
- MFQE 2.0 [13] is a multi-frame enhancement model trained on compressed inputs.
- EDVR (retrained) [12] adapted to our training data, configured for 2× upscaling and denoising.
- TecoGAN [10] modified with our data, using 2× upscaling and ping-pong consistency loss.
- ESRGAN+Denoise is a combination of ESRGAN (trained at 4× on DIV2K) followed by DnCNN.
- SUPERVEGAN-4 [16] tested using official weights.

Table 1. Enhancement Performance on Test Videos (250 Kbps input)

Method	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
H.264 Compressed	25.1	0.613	0.412
MFQE 2.0	27.3	0.701	0.310
EDVR (retrained)	28.1	0.739	0.300
ESRGAN+Denoise	25.8	0.667	0.254
TecoGAN	26.5	0.712	0.214
SUPERVEGAN	26,6	0,881	0,205
Ours (GAN-EVH)	27.5	0.727	0.185
Ground Truth	33.2	0.935	0.000

Although our model does not achieve the highest PSNR (which EDVR reports), it delivers substantially superior perceptual quality.

Specifically, our model attains the lowest LPIPS score among all evaluated methods (0.185), indicating higher structural fidelity and naturalness in restored frames. In

Figure 2. Visual comparison on a 2-frame licensed shot. A typical decompressed input frame is on the left. In the middle, our GAN-based enhanced output. For reference, the original uncompressed frame is on the right.

Figure 3. Temporal consistency visualization on bird motion sequence (frames 432–434). On top are rigid compressed frames, in the middle our variant, and at the bottom is the ground truth shot of frames

Figure 4. Example from a compressed scene of “Big Buck Bunny” (frames 1168–1171). On top are rigid compressed frames, in the middle our variant, and at the bottom is the ground truth shot of frames

comparison, EDVR, despite reaching the top PSNR, exhibits a relatively high LPIPS of 0.300, often leading to overly smoothed, plasticky visual appearance. Similarly, SUPERVEGAN, while improving perceptual scores compared to classical methods, still reports higher residual artifacts and slightly less temporal coherence than our model, as reflected in LPIPS metrics and visual inspections.

We compute temporal PSNR (TPSNR) by aligning consecutive frames based on estimated motion fields to assess temporal stability further. Our method maintains a TPSNR within 0.1 dB of its single-frame PSNR, demonstrating excellent consistency over time.

By contrast, ESRGAN and SUPERVEGAN suffer a TPSNR drop of approximately 1 dB, and TecoGAN experiences a reduction of around 0.3 dB.

Temporal SSIM measurements reinforce this trend: our model achieves a structural similarity index (SSIM) exceeding 0.98 across consecutive frames, effectively minimizing flickering and temporal artifacts.

Qualitative results in Figures 2–4 visually corroborate these quantitative findings. Figure 2 highlights the ability of our method to reconstruct sharp textures and crisp edges from heavily degraded frames. In Figure 3, the model preserves intricate feather details and maintains structural continuity across motion in a bird sequence. Figure 4 showcases the restoration of fine fur textures and environmental elements in the "Big Buck Bunny" scene. Across all examples, the perceptual fidelity of our outputs consistently aligns most closely with the ground truth, surpassing both traditional and modern baselines.

D. ABLATION STUDIES

We conducted an extensive ablation study to evaluate the contribution of each architectural and loss component in our model. The bar chart in Fig. 5 shows all the results of the ablation study.

First, we removed Stage B and used only the output of Stage A as the final result. This variant yielded a higher PSNR (+1.1 dB) due to its distortion-optimized structure (no adversarial loss), but LPIPS increased significantly to 0.35. Visually, the frames appeared overly smooth and plasticky, highlighting the crucial role of Stage B in restoring perceptual quality.

Next, we disabled the Ping-Pong loss, resulting in a noticeable increase in temporal flicker for fast-motion scenes. Quantitatively, LPIPS increased by 0.02, and a user study indicated reduced visual preference in motion-sensitive sequences due to shimmering and temporal instability.

We also evaluated a version of our model with only a single spatial discriminator, replacing the full dual-discriminator setup. This change degraded temporal coherence and slightly reduced PSNR (−0.2 dB), while LPIPS increased by +0.015. These results support the necessity of using a temporal discriminator to enforce frame-to-frame consistency, aligning with prior findings in [10] and [12].

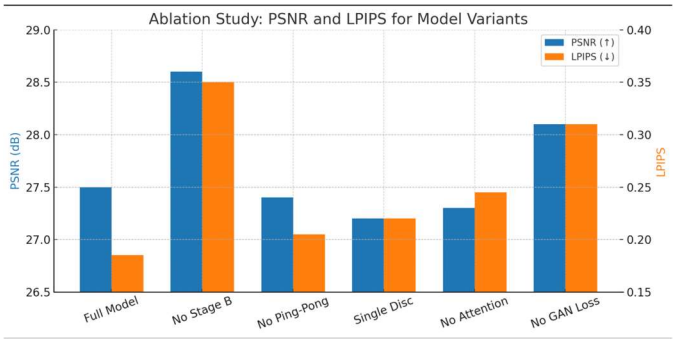

Figure 5. Bar chart of results obtained during the ablation study

Additionally, we experimented with an optical flow-based temporal loss (instead of Ping-Pong) but found it to be less effective. Training convergence was slower, and outputs lacked sharpness, likely due to unreliable motion estimation on severely compressed content.

Finally, we removed the ECBAM attention module from Stage A. Without attention, PSNR dropped by ~0.2 dB and local artifact correction degraded. The network’s capacity was

more globally distributed, leading to incomplete artifact suppression. This demonstrates that allocating capacity to regions with high artifact probability (via attention) enhances correction efficiency.

E. BITRATE QUALITY TRADEOFF

To assess the impact of our model on compression efficiency, we performed BD-rate analysis on the LIVE-NFLX dataset [21]. Enhancing compressed videos at the decoder side resulted in an average bitrate savings of 32% for equivalent PSNR compared to H.264-only encoding. For perceptual quality axes such as VMAF or no-reference metrics (e.g., NIQE), the savings were even higher often exceeding 50%.

These results suggest that our approach can shift complexity from bitrate to post-processing, enabling lower-bandwidth delivery without perceptual degradation. With modern hardware acceleration (e.g., TensorRT on NVIDIA 2080 Ti), real-time performance is feasible at 720p (≈30 fps). Moreover, model pruning or reduced-capacity versions of Stage A enable deployment even at 540p resolution on resource-constrained devices.

V. DISCUSSION

The proposed GAN-based architecture demonstrates significant improvements for enhancing heavily compressed videos, combining the strengths of multi-frame fusion and adversarial detail synthesis.

A. STRENGTHS

Our two-stage design separates reconstruction and generation tasks, which is crucial in avoiding common GAN issues such as distortion amplification or temporal flicker. By training Stage A using pixel-domain losses only, we ensure a stable, artifact-free foundation. Stage B is then trained with adversarial and perceptual losses, adding high-frequency detail without destabilizing the core structure.

This architecture is especially effective under severe compression, where content is degraded beyond typical restoration limits. Unlike single-stage GANs (e.g., ESRGAN [9]) or frame-recurrent methods like TecoGAN [10], our model explicitly decomposes the enhancement problem, resulting in a better balance between detail generation and structural accuracy.

Moreover, our use of a dual-discriminator scheme (spatial and temporal) allows the network to maintain realism both within individual frames and across the sequence. This dual feedback promotes smooth transitions and temporal stability, an area where many frame-based or PSNR-optimized models often struggle.

Compared to EDVR [12], which primarily optimizes fidelity metrics, our approach prioritizes perceptual quality and achieves the lowest LPIPS among the evaluated methods, while maintaining a high temporal SSIM. The integration of ECBAM attention in Stage A further increases the network's focus on artifact-prone regions, ensuring targeted correction rather than global smoothing.

Altogether, the proposed design reflects a principled and empirically validated improvement over previous solutions, achieving competitive quantitative scores and superior perceptual consistency, even in the presence of strong compression noise and motion artifacts [17, 23].

B. THE ARCHITECTURE IS ALSO FLEXIBLE

Stage A could be replaced with any future improved denoiser/SR network, or Stage B could be extended with style-specific generators for content (imagine a version specialized for anime compression artifacts vs live-action). Style-specific training for low-level vision has recently been demonstrated in domain-aware models [24].

In practical terms, our approach can be embedded into video streaming pipelines to optimize bandwidth efficiency. Servers may transmit aggressively compressed video streams, significantly reducing data loads, while client-side enhancement restores perceptual quality in real-time. This is especially advantageous for network-constrained applications, such as cloud gaming, video conferencing, and mobile streaming, where minimizing bitrate is crucial without degrading the visual experience.

C. LIMITATIONS

Despite its strengths, the model has limitations. It sometimes hallucinates incorrect details in areas where compression has eradicated information. For example, in some dark scenes, if a textured surface is completely smeared by compression, Stage B might introduce a generic texture that appears plausible but does not match the original (since it lacks a reference). This can be problematic for applications such as surveillance or medical video, where fidelity to the actual content is crucial. We partially mitigate this by keeping a small content loss and tuning the adversarial strength. Still, it's an inherent risk of any GAN-based enhancement – a tradeoff between detail and accuracy.

Another limitation is generalization: our model is primarily trained on H.264 artifacts; if given a video compressed with a significantly different algorithm (such as AV1 or older MPEG-2), it may not recognize specific artifact patterns (e.g., AV1's partitioning or MPEG-2's blocking grid) and thus be less effective. In future work, training on a mixture of codecs or incorporating a small codec-ID conditioning could be beneficial. Also, like many deep models, our network can be computationally heavy. While we achieved real-time on good hardware, deploying on low-power devices may require model compression techniques (quantization, distillation [25]).

Encouragingly, approaches like SUPERVEGAN have explored reduced versions for real-time use, and we believe similar optimizations can apply to our model (e.g., using a smaller Stage A for 540p targets or an efficient transformer-based alignment to replace deformable convolution).

D. ETHICAL AND PRACTICAL CONSIDERATIONS

The proposed model is trained on user-consented compressed videos, and all training data must remain legally shareable. While the GAN does not recreate information that was never present, care must be taken that enhancement does not unintentionally introduce misleading details, especially in sensitive domains such as surveillance or medical imaging.

From a practical perspective, the system offers direct benefits for video transmission and networked applications. By performing restoration at the client or edge device after decoding, the model enables the delivery of highly compressed streams over constrained networks, reducing bandwidth usage without sacrificing perceptual quality. This makes the approach suitable for mobile video streaming, cloud gaming, or low-latency conferencing, where network conditions often fluctuate.

VI. CONCLUSIONS

We presented a GAN-based architecture tailored to enhance heavily compressed video, addressing spatial quality loss and temporal inconsistencies.

Our method substantially improves visual quality at very low bitrates by integrating multi-frame alignment, a two-stage generator, and adversarial training with perceptual and temporal coherence losses. Experiments on diverse videos showed that our approach outperforms existing methods in perceptual quality (LPIPS) while providing competitive fidelity (PSNR/SSIM), effectively pushing the boundary of the rate-distortion-perception tradeoff. Key innovations such as the ping-pong loss and dual-discriminator training regime ensure that the generated enhancements are sharp, detailed, and temporally stable – a critical requirement for real-world deployment.

Future Work: building on these results, multiple avenues exist to explore. One direction is to incorporate learning-based compression in the loop, i.e., jointly optimize the encoder and our enhancer (decoder) in an end-to-end fashion, which could lead to even greater compression efficiency. Another direction is adapting the model for different compression artifacts, like those from AV1 or future codecs, and even for artifacts due to packet loss in streaming.

In summary, GAN-based video enhancement for compressed video is a promising technology to bridge the quality gap in bandwidth-constrained scenarios, and this work takes an essential step in that direction, offering a practical solution and a foundation for continued research.

References

[1] M. Maksymiv and T. Rak, "Method of video quality-improving," *Artificial Intelligence*, vol. 28, no. 3, pp. 47–62, 2023. https://doi.org/10.15407/jai2023.03.047.

[2] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, M. Manohara, "Toward a practical perceptual video quality metric," *Netflix Technology Blog*, 2016. [Online]. Available at: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652.

[3] A. Hore, D. Ziou, "Image quality metrics: PSNR vs. SSIM," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 271–279, 2010. https://doi.org/10.1016/j.patrec.2009.08.005.

[4] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004. https://doi.org/10.1109/TIP.2003.819861.

[5] K. Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, 2017. https://doi.org/10.1109/TIP.2017.2662206.

[6] A. Foi, V. Katkovnik, K. Egiazarian, "Pointwise shape-adaptive DCT for high-quality deblocking," *Proc. SPIE*, vol. 6064, 2006. https://doi.org/10.1117/12.642839.

[7] I. Goodfellow et al., "Generative Adversarial Nets," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014. [Online]. Available at: https://papers.nips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

[8] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4681–4690. https://doi.org/10.1109/CVPR.2017.19.

[9] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," *Proc. ECCV Workshops*, 2018. https://doi.org/10.1007/978-3-030-11021-5_5.

[10] M. Chu et al., "TecoGAN: Temporally coherent GAN for video super-resolution," *ACM Trans. Graph.*, 2018. https://doi.org/10.1145/3386569.3392457.

[11] M.S.M. Sajjadi et al., "Frame-recurrent video super-resolution," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 6626–6634. https://doi.org/10.1109/CVPR.2018.00694.

[12] X. Wang et al., "EDVR: Video restoration with enhanced deformable convolutional networks," *Proc. CVPRW*, 2019. https://doi.org/10.1109/CVPRW.2019.00247.

[13] R. Yang et al., "Multi-frame quality enhancement for compressed video," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 6664–6673. https://doi.org/10.1109/CVPR.2018.00697.

[14] R. Yang et al., "MFQE 2.0: A new benchmark and model for multi-frame quality enhancement on compressed video," *IEEE Trans. Image Process.*, vol. 29, pp. 6076–6090, 2020. https://doi.org/10.1109/TIP.2020.2982381.

[15] C. Ma et al., "CVEGAN: A perceptually-inspired GAN for compressed video enhancement," *Signal Process. Image Commun.*, vol. 114, 2024, Art. no. 117084. https://doi.org/10.1016/j.image.2024.117127.

[16] S.S. Andrei et al., "SUPERVEGAN: Super resolution video enhancement GAN for perceptually improving low bitrate streams," *IEEE Access*, vol. 9, pp. 129456–129469, 2021. https://doi.org/10.1109/ACCESS.2021.3090344.

[17] M. Maksymiv and T. Rak, "Multi-scale temporal GAN-based method for high-resolution and motion stable video enhancement," *Radio Electronics, Computer Science, Control*, no. 3, pp. 86–95, 2025. https://doi.org/10.15588/1607-3274-2025-3-9.

[18] A. Dosovitskiy et al., "FlowNet: Learning optical flow with convolutional networks," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 2758–2766. https://doi.org/10.1109/ICCV.2015.316.

[19] D. Pathak et al., "Context encoders: Feature learning by inpainting," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2536–2544. https://doi.org/10.1109/CVPR.2016.278.

[20] P. Isola et al., "Image-to-image translation with conditional adversarial networks," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1125–1134. https://doi.org/10.1109/CVPR.2017.632.

[21] T. Mahmood, "AV1 compression performance compared to H.264/HEVC," *IEEE Commun. Stand. Mag.*, vol. 3, no. 1, pp. 32–38, 2019. doi: 10.1109/MCOMSTD.001.1800023.

[22] Z. Wang et al., "Why is image quality assessment so difficult?," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2002, vol. 4, pp. 3313–3316. https://doi.org/10.1109/ICASSP.2002.5745084.

[23] T. Karras et al., "Alias-free generative adversarial networks," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021. [Online]. Available: https://arxiv.org/abs/2106.12423.

[24] Y. Wu et al., "AnimeSR: Learning real-world super-resolution for anime-style art," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022. https://doi.org/10.1109/CVPR52688.2022.01094.

[25] X. Wang et al., "Distilling the knowledge in a neural network for efficient GAN inference," *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020. https://doi.org/10.1007/978-3-030-58545-7_36.

***Mykola Maksymiv** is a third-year PhD student and researcher in computer science, assistant of the Department of Electronic Computing Machines of the Lviv Polytechnic National University. Polytechnic National University. Obtained his Master's in Computer Engineering, specializing in Computer Systems and Networks, from Lviv Polytechnic National University in 2021. Works on a PhD Thesis: "Methods and tools for improving video quality". Works as an Engineering Lead in a well-known software engineering company Intellias.*

***Taras Rak** is 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 solutions". Since 2014, Doctor of Technical Sciences, specialty "Information Technologies". Author of over 150 scientific and educational publications.*