


Article

PixRevive: Latent Feature Diffusion Model for Compressed Video Quality Enhancement

Weiran Wang¹, Minge Jing^{1,*}, Yibo Fan¹ and Wei Weng² 

¹ School of Microelectronics, Fudan University, Shanghai 200433, China; wrwang0922@gmail.com (W.W.); fanyibo@fudan.edu.cn (Y.F.)

² Department of Liberal Arts and Science, Kanazawa University, Ishikawa 920-1192, Japan; weng@staff.kanazawa-u.ac.jp

* Correspondence: mejing@fudan.edu.cn

Abstract: In recent years, the rapid prevalence of high-definition video in Internet of Things (IoT) systems has been directly facilitated by advances in imaging sensor technology. To adapt to limited uplink bandwidth, most media platforms opt to compress videos to bitrate streams for transmission. However, this compression often leads to significant texture loss and artifacts, which severely degrade the Quality of Experience (QoE). We propose a latent feature diffusion model (LFDM) for compressed video quality enhancement, which comprises a compact edge latent feature prior network (ELPN) and a conditional noise prediction network (CNPN). Specifically, we first pre-train ELPNet to construct a latent feature space that captures rich detail information for representing sharpness latent variables. Second, we incorporate these latent variables into the prediction network to iteratively guide the generation direction, thus resolving the problem that the direct application of diffusion models to temporal prediction disrupts inter-frame dependencies, thereby completing the modeling of temporal correlations. Lastly, we innovatively develop a Grouped Domain Fusion module that effectively addresses the challenges of diffusion distortion caused by naive cross-domain information fusion. Comparative experiments on the MFQEv2 benchmark validate our algorithm's superior performance in terms of both objective and subjective metrics. By integrating with codecs and image sensors, our method can provide higher video quality.

Keywords: compressed video restoration; diffusion model; rich detail information; group-wise domain fusion



Citation: Wang, W.; Jing, M.; Fan, Y.; Weng, W. PixRevive: Latent Feature Diffusion Model for Compressed Video Quality Enhancement. *Sensors* **2024**, *24*, 1907. <https://doi.org/10.3390/s24061907>

Academic Editor: Yun Zhang

Received: 2 February 2024

Revised: 4 March 2024

Accepted: 13 March 2024

Published: 16 March 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The thriving development of Internet of Things (IoT) technologies has led to an explosion in video data traffic. However, the massive costs of data storage and limited upload bandwidth pose obstacles to the continuous transmission of high-quality videos. To tackle this challenge, classic video coding standards have emerged, including H.263 [1], H.264/AVC [2], and H.265/HEVC [3]. These schemes leverage the spatial and temporal redundancies in video content to enable efficient transmission and storage through lossy compression. Meanwhile, breakthroughs in image sensor technologies have steadily improved video resolution, dynamic range, and denoising capabilities. This provides superior initial conditions for compression coding and richer quality clues for subsequent video refinement and restoration algorithms. However, inevitable bitrate reduction introduces multifaceted data loss and compression artifacts like blurring, blockiness, and edge fluctuations [4,5]. Such distortions fail to satisfy the requisite user experience quality (QoE) [6,7]. Additionally, disruption and damage introduced in frame coding adversely affect downstream computer vision tasks reliant on video content like scene analysis and object tracking, thus compromising visual fidelity. Therefore, developing powerful compressed video restoration algorithms to rectify compression-induced reductions in image quality is imperative.

Earlier traditional methods typically optimize transform coefficients based on specific compression standards [8,9]. Such codec-dependent approaches struggle to generalize across standards. In contrast, CNN-based methods, like QE-CNN [10], MFQEv2 [11], STDF [5], and RFDA [12], demonstrate superior performance on video enhancement tasks. With the advent of neural network architectures like Vision Transformers [13,14], learning-based video restoration techniques have also made significant strides. State-of-the-art designs such as STCF [15] and TVQE [16] prove effective for restoration. Beyond task-specific solutions, recent research has also established unified frameworks, like BasicVSR [17] and BasicVSR++ [18], to address compression artifacts. However, the limitations imposed by these methods impede their performance, making it challenging to effectively address highly uncertain issues, such as images that are severely damaged or have significant information loss. It is difficult to accurately infer the Possibility distribution of missing parts from the remaining valid pixels. Therefore, tighter integration of sensor technologies and video codecs to generate outstanding high-quality video remains key for advancing compressed video perceptual quality enhancement algorithms.

To address the aforementioned challenges, we intend to utilize cutting-edge conditional generation modeling (diffusion probability model) [19,20] as the foundation. By leveraging advanced sensor imaging systems and robust generative architectures with strong representation and generalization capabilities, we aim to progressively enhance the quality of data during the reconstruction process, thereby generating more intricate and lifelike images. This approach allows the model to focus on detail recovery incrementally, rather than attempting to solve the entire complex problem at once, ultimately improving the efficiency of video restoration. However, without careful guidance, directly applying the diffusion model to video damage repair may disrupt inter-frame dependencies and inevitably lead to detail distortion. To surmount these limitations, we propose a novel synergistic framework between denoising diffusion and CNNs to ameliorate compression video impairments. Our model first extracts edge information from video frames through the ELPNet based on discrete wavelet transform, enabling more targeted and higher-quality reconstruction of high-frequency components. This constructs a pseudo ground-truth feature space guiding the diffusion model's denoising process. Finally, the outputs are fused together. Through this collaborative framework, highly correlated information complements each other to effectively restore low-quality video, achieving state-of-the-art restoration quality on the MFQEv2 dataset compared to previous approaches.

Our contributions are summarized as follows:

1. We propose the first diffusion-model-based video compression restoration network, surpassing the performance limitations of previous neural network methods.
2. We design a frequency-domain filling block (FFB), the core idea of which is leveraging the multi-resolution frequency-domain features provided by wavelet transforms to guide detail restoration. It provides more high-frequency knowledge to reconstruct sharp texture details.
3. Theoretical analysis reveals domain discrepancies between diffusion models and deep convolutional networks. Direct latent feature fusion may exacerbate these gaps, inducing distortions. To mitigate this, we design a simple yet effective group-wise domain fusion module.
4. Extensive experiments and ablation studies validate the superior performance of our proposed technique.

2. Related Work

2.1. Compressed Image/Video Restoration

Inspired by the success of deep learning, a multitude of recent works [21–28] have demonstrated that convolutional neural networks (CNNs) exhibit superior performance in enhancing image and video compression quality. The ARCNN designed by Dong et al. [22] pioneers the leverage of CNNs to mitigate artifacts introduced by JPEG encoding. Owing to its robustness, DnCNN [23] is frequently employed as the benchmark for image

restoration, including denoising and artifact reduction. QE-CNN [10] utilizes two models to reduce distortions for I frames and P/B frames. MFQEv2 [11] utilizes motion compensation between two adjacent peak quality frames extracted by optical flow estimation to enhance low-quality frames. Additionally, to effectively process motion relations, STDF [5] proposes a spatiotemporal deformable fusion scheme to aggregate temporal information to eliminate unpleasant distortions. RFDA [12] further refines STDF through recursive fusion and deformable spatiotemporal attention modules to simulate long-range motion compensation. To enhance perceptual quality, a new generative adversarial network named MW-GAN+ [29] leverages multi-level wavelet packet transform (WPT) to recover high-frequency details and fine-grained textures. Recently, researchers have introduced Transformer-based frameworks into the field of video compression restoration and achieved promising results. Zhang et al. [15] designed a parallel structure combining Swin Transformer and CNN, which integrates motion compensation and global context information. Another work, TVQE [16], designed novel modules that are capable of not only learning local and global features for correlational modeling but also aggregating inter-frame information. These methods can effectively restore the artifacts caused by video compression. However, these methods falter in reconstructing high-frequency details, especially along image edges. Additionally, over-reliance on intrinsic learning patterns during training hampers texture expressiveness, yielding blurry, smoothed outputs, thus rendering the restoration work unsatisfactory.

2.2. Diffusion Models

Diffusion-based [30] generative models have recently regained widespread attention. This class of models sequentially perturbs data samples by introducing additive noise to simplify them into elementary distributions (e.g., Gaussian), then reverses the process, and learns to recover the latent variables in the simple distribution back to data in the complex distribution by optimizing a variational lower bound of the likelihood function, using parameterized Markov chains. Subsequently, these models gradually denoise samples from the noisy distribution via Langevin dynamics [31], yielding target samples from the data distribution.

Recently, DDPM [32] has shown state-of-the-art performance across various tasks, including image super-resolution [33,34], restoration [35,36], and translation tasks (restoration, colorization, etc.) [37,38]. Additionally, the learned feature representations from diffusion models also prove very useful for discriminative tasks, including image classification [39], segmentation [40,41], and object detection [42]. Diffusion models have been extensively used for sample generation owing to the high quality and diversity of their generated samples. With the continuous advancement of diffusion models across domains, they have surpassed the long-standing dominance of GANs in image generation. However, intrinsic defects persist for utilizing diffusion models in video restoration. Specifically, we have empirically shown, through experiments, that merely applying diffusion models fails at temporal modeling, contrarily deteriorating performance. Hence, our work ingeniously overcomes the innate deficiencies of diffusion models in inter-frame modeling through innovative architectural designs.

2.3. Neural Network Combined with Diffusion Model

To better enhance the image restoration capability of diffusion models, existing works [43,44] incorporate latent features from conditional neural networks into training diffusion models. Specifically, the method extracts integrated features from low-resolution images through a neural network for conditioning to guide image generation. Then, the neural network features are simply linearly combined with the probability distribution features from the diffusion model; while moderately improving restoration on specific domains, there are some limitations: (1) the weak detail restoration capabilities; (2) disregarding domain discrepancies and simply conducting linear fusion lead to unsatisfactory detail effects or even distortions; and (3) the fusion mainly aims to improve restoration on

specialized domains rather than generalizing to common visible light images. In contrast, our method has three main advantages: (1) Our guiding network leverages discrete wavelet transforms to obtain richer texture details, abstracted into the latent space for enhancing detail restoration and generation capacity. We then integrate this wavelet-enhanced network with the diffusion model for targeted performance gains. (2) We devise a simple yet effective patch-wise domain matching module to bridge domain gaps for seamless fusion, alongside an efficient fusion mechanism. (3) We have extended our model to common visible light domains and achieved state-of-the-art results.

3. Preliminaries: Diffusion

In this paper, we adopt diffusion models to generate accurate restorations for compressed damaged video frames. This is achieved by learning Markov chains that progressively convert the Gaussian noise distribution to the trained model's data distribution. The process comprises two key phases: forward diffusion and reverse diffusion. As illustrated in Figure 1, given the true data distribution $x_0 \sim p(x)$, the forward diffusion process injects Gaussian noise over T timesteps to incrementally corrupt the distribution. This yields a series of noisy samples, parametrized by the variance schedule $(\beta_1, \beta_2, \dots, \beta_t)$. Noise samples denote latent variables sharing the original data dimension. Each iteration of the forward process, transforming x_0 into $x_T \sim \mathcal{N}(0, I)$, can be described as:

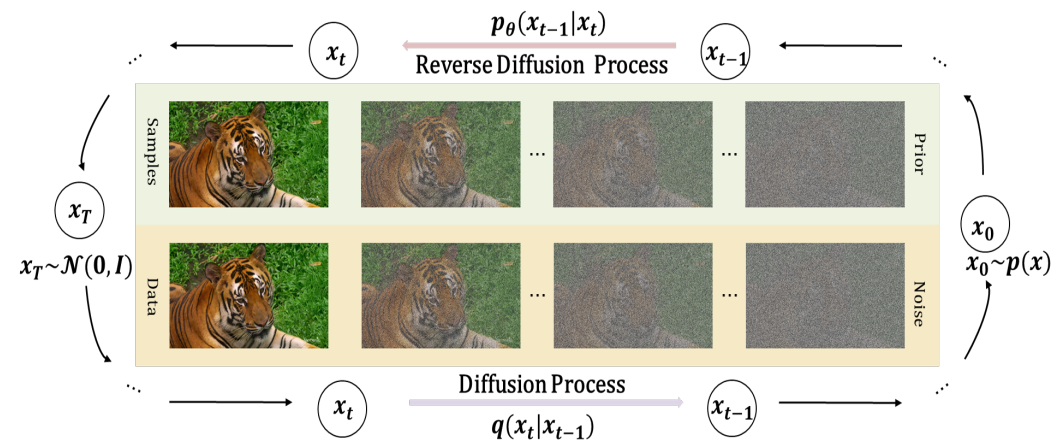


Figure 1. The diffusion process and inverse diffusion process of diffusion models for compressed video frame restoration.

$$p(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I). \quad (1)$$

For ease of calculation and formula representation, let $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=0}^t \alpha_i$; Equation (1) can be further reduced to:

$$p(x_t | x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \sqrt{\bar{\alpha}_t})I). \quad (2)$$

This suggests that the data distribution $p(x_t|x_0)$ can be computed directly from Equation (2) for any moment t without iteration. As t increases, the fraction of the introduced noise escalates, while that of the original data x_0 diminishes. When Gaussian noise dominates, the distribution of $p(x_t|x_0)$ converges to the Gaussian distribution $\mathcal{N}(0, I)$, indicating the completion of the forward diffusion phase where structural information corrodes.

The learning of diffusion models is achieved by reversing the forward process defined in Equation (1) to construct a reverse Markov chain. Specifically, define a joint distribution $p_\theta(x_0, \dots, x_T)$ controlled by θ , and then construct a reverse process based on this joint distribution, that is, starting from the standard normal distribution $p(x_T) = \mathcal{N}(x_T; 0, I)$,

perform Gaussian denoising step by step until approximating the true data distribution. The formulas are as follows:

$$q_{\theta}(x_0, \dots, x_{T-1} | x_T) := \prod_{t=1}^T q_{\theta}(x_{t-1} | x_t), \quad (3)$$

$$q_{\theta}(x_{t-1} | x_t) := \mathcal{N}(x_{t-1}; \mu_{\theta}(x_t, t), \sigma_{\theta}(x_t, t)^2 I). \quad (4)$$

The parameters involved in the backward process, such as $\mu_{\theta}, \sigma_{\theta}$, represent the mean and variance of the Gaussian distribution, respectively, which are estimated by a neural network. In addition, the sequence of variances β_t mentioned in the forward process can participate in joint model learning or remain unchanged.

In the training phase, we construct an upper bound on the negative log-likelihood by adding a non-negative KL dispersion term to the negative log-likelihood function $-\log p_{\theta}(x_0)$ of the target data distribution $p_{\theta}(x_0)$, denoted as Equation (5), and the specific expansion can be expanded into [32].

$$\begin{aligned} -\log p_{\theta}(x) &\leq -\log p_{\theta}(x_0) + D_{\text{KL}}[q(x_{1:T} | x_0) \| p_{\theta}(x_{1:T} | x_0)] \\ \mathbb{E}_{q(x_0)}[-\log p_{\theta}(x_0)] &\leq \mathbb{E}_q \left[\underbrace{D_{\text{KL}}(q(x_T | x_0) \| p(x_T))}_{L_T} - \underbrace{\log p_0(x_0 | x_1)}_{L_0} \right] \\ &\quad + \sum_{t>1} \underbrace{D_{\text{KL}}(q(x_{t-1} | x_t, x_0) \| p_{\theta}(x_{t-1} | x_t))}_{L_{T-1}}, \end{aligned} \quad (5)$$

In the L_{T-1} term in the above formula, the KL divergence of the two Gaussian distributions $p_{\theta}(x_{t-1} | x_t)$ and $q(x_{t-1} | x_t, x_0)$ is calculated; the latter is based on the original data X_0 . The posterior distribution of the true unknown generation process is inferred from the global perspective of the entire diffusion model. The specific expression is as follows:

$$q(x_{t-1} | x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I), \quad (6)$$

where mean $\tilde{\mu}_t(x_t, x_0) = \frac{1}{\sqrt{\tilde{\alpha}_t}} \left(x_t - \epsilon \frac{1 - \tilde{\alpha}_t}{\sqrt{1 - \tilde{\alpha}_t}} \right)$, variance $\tilde{\beta}_t = \frac{1 - \tilde{\alpha}_{t-1}}{1 - \tilde{\alpha}_t} \beta_t$, and ϵ represents the noise in x_t , which is the only uncertain variable in the reverse process. The diffusion model uses a denoising network $\epsilon_{\theta}(x_t, t)$ to estimate ϵ . Finally, based on the description in [32], we perform the parameter optimization of the network by means of Equation (7).

$$\nabla_{\theta} \left\| \epsilon - \epsilon_{\theta} \left(\sqrt{\tilde{\alpha}_t} x_0 + \epsilon \sqrt{1 - \tilde{\alpha}_t}, t \right) \right\|_2^2. \quad (7)$$

4. Approach

Given a compressed low-quality video sequence, $V_{lq} = \{X_k \in \mathbb{R}^{C \times H \times W}\}$ with K frames, where $k \in (1, K)$. C , H , and W denote the channel, height, and width of each frame, respectively. As shown in Figure 2, we demonstrate the overall pipeline of the Latent Feature Diffusion Model (LFDM). In our methodology, we feed the current frame into ELPN and additionally introduce adjacent frames to enhance the richness of the original input information, which enables the network to construct a more coherent spatiotemporal representation, thus preserving inter-frame dependencies. The corresponding reference frame input is $X_f = \{X_{k-1}, X_k, X_{k+1}\}$. When enhancing it into a high-quality frame, we extract and store the mapped features as a pseudo ground-truth feature bank to provide more accurate conditional features for reverse diffusion. This allows the diffusion model to probe a solution space akin yet not identical to the conditional features, chasing improved outcomes while retaining correlation with the multi-frame data. We use Equation (2) to convert X_k into $P(X_t | X_k)$ as the input for the diffusion model. Finally, fusing its output

with the repository features produces the optimal result. Overall, the enhanced frame \hat{Y}_t of the compressed frame X_k is generated as:

$$\begin{aligned} F_k &= F_{\text{con}}(X_f), \quad k \in \{0, 1, 2, 3\}, \\ D_i^P &= P(X_t|X_k), \\ D_o^P &= \text{Diff}(D_i^P, F_k), \quad k \in \{0, 1, 2, 3\}, \\ \hat{Y}_t &= \text{Fusion}(D_o^P, F_k), \end{aligned} \quad (8)$$

where $F_{\text{con}}(\cdot)$ denotes the decoder of the ELPNet, $\{F_k \mid k \in 0, 1, 2, 3\}$ represents decoding-end features of varying sizes extracted from the ELPNet, $\text{Diff}(\cdot)$ refers to the diffusion model's conditional denoising network, and fusion signifies the final module fusing information across domains. This effectively mitigates deficiencies induced by directly fusing cross-domain features, thereby unleashing the potential of heterogeneous information to better achieve the target task. The details of ELPNet, diffusion, and fusion will be elaborated in Sections 4.1, 4.2, and 4.3, respectively.

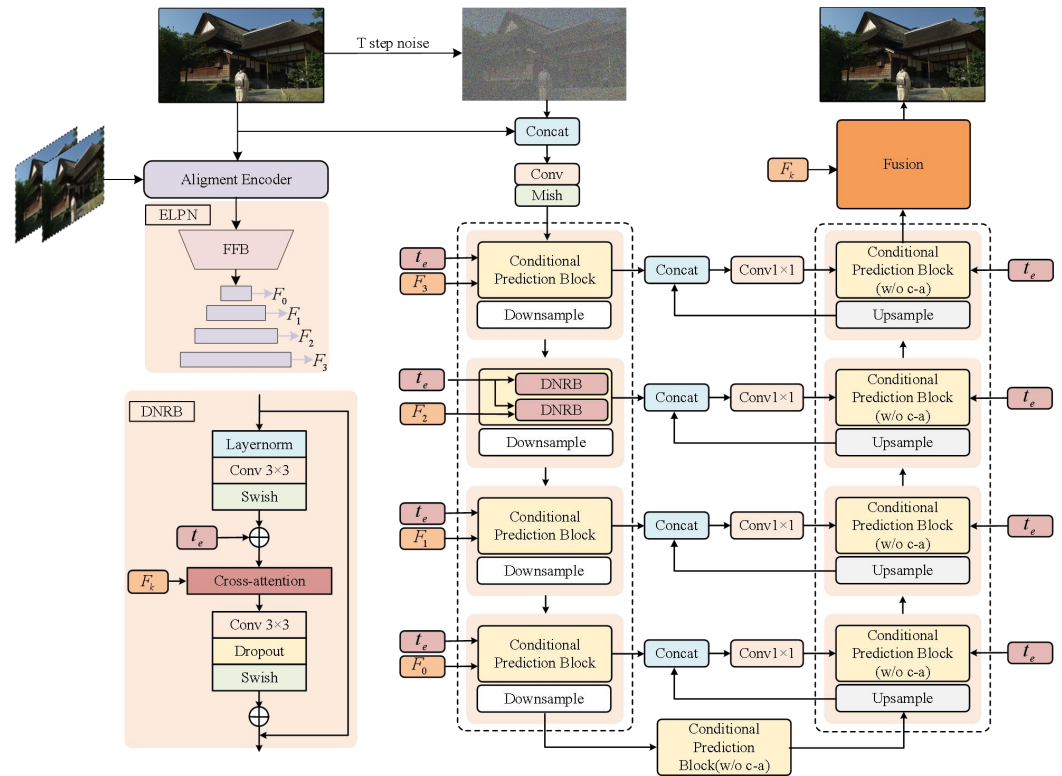


Figure 2. The overall architecture of the proposed LFDN. First, the current frame and neighboring frames are fed into the ELPNet for pre-training. Second, the ELPNet extracts the prior latent features and feeds them to the CNPN to direct its generation process. The details of the CNPN are illustrated in the figure. Finally, feature information from different domains is consolidated via “fusion”, comprehensively elaborated on in Section 4.3. Here, $t \sim \text{Uniform}\{1, \dots, T\}$ and transformed into t_e through an MLP; FFB represents the frequency-domain filling block; (w/o c-a) denotes without cross-attention.

4.1. ELPNet

Before introducing the ELPNet, we first present a spatiotemporal alignment module [45] that harnesses optical flow estimation (OFE) to compute forward and backward flows between adjacent frames. These optical flows then warp the input frames temporally, which is vital to leverage useful information from neighboring frames for restoring the target.

Our CNN branch, namely, ELPNet (Figure 3), aims to directly learn the mapping from damaged to pristine images. Its encoded integrated features serve as conditioning to guide diffusion model generation. To achieve this, we adopt the same architecture as the diffusion model's denoising network for constructing the ELPNet. By conducting feature extraction through ELP-Resblock (structure in Figure 3, left), which blends frequency-domain information using Discrete Wavelet Transforms (DWTs), we can retain more texture details during restoration while forcing the network to learn both high and low frequencies.

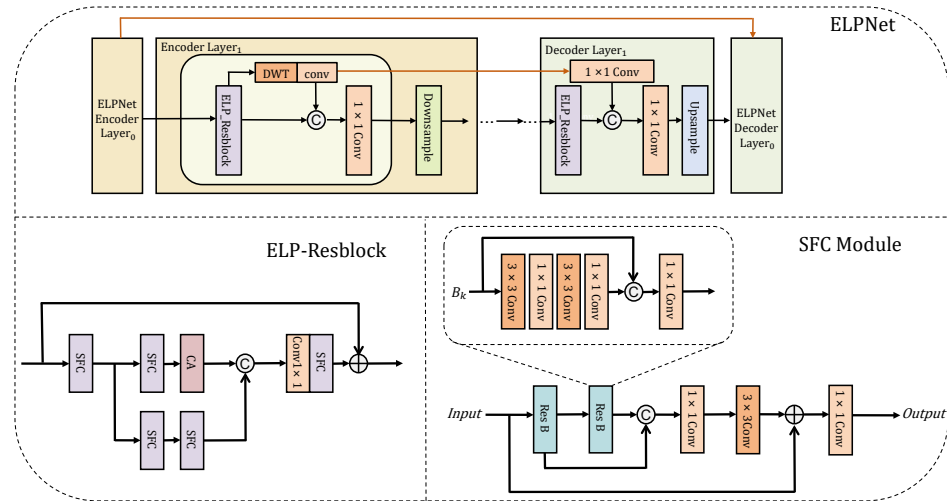


Figure 3. The overall structure of our proposed ELPNet. DWT refers to the Discrete Wavelet Transform and CA denotes the Channel Attention mechanism.

Specifically, a fixed-parameter low-pass filter (L_{FF}) and high-pass filters (H_{FF1} , H_{FF2} , and H_{FF3}) perform stride 2 convolution calculations to decompose images or feature maps into four sub-bands (X_{LF} , X_{HF1} , X_{HF2} , X_{HF3}). We denote X_{LF} as $(L_{FF} \otimes X) \downarrow_2$, which represents the convolutional computation, where \downarrow_2 indicates a 2x scaling factor. We embed the Haar DWT [46] into our proposed network, $L_{FF} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $H_{FF1} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}$, $H_{FF2} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}$, $H_{FF3} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$. Then, the value at the (i, j) -th position of X_{LF} after 2D Haar wavelet transformation can be calculated by Equation (9):

$$X_{LF}(i, j) = X(2i - 1, 2j - 1) + X(2i - 1, 2j) + X(2i, 2j - 1) + X(2i, 2j). \quad (9)$$

The expressions for the high-frequency sub-bands are similar to the expression for X_{LF} . The integration of low-frequency components as encoding side features with downsampled features provides powerful semantic information and a relatively coarse spatial layout. Furthermore, high-frequency components are integrated into the decoding side region through a multitude of skip connections, guaranteeing the preservation and enhancement of fine image details during the image reconstruction phase. This approach enables our network to not only amalgamate rich information from spatial and frequency domains during the learning process but also enhances its capability to capture high-frequency features like image textures and contours. The experimental results show that the embedding of DWT indeed greatly improves the restoration capability of the network (see Section 5.3 for details).

To ensure the retention of ample texture information in the final restoration results, thereby assisting the diffusion model in recovering intricate and clear details, we apply the following loss function to ELPNet for training, which can be represented as:

$$L = L_{Char} + \alpha L_{MS} + \beta L_{Per}, \quad (10)$$

where L_{Char} refers to Charbonnier loss [47], L_{MS} refers to MS-SSIM loss [48], and L_{Per} is perceptual loss [49]. After experimentation, $\alpha = 0.2$ and $\beta = 0.001$ were finally determined as the hyperparameter weights for each loss function part.

4.2. Noise Prediction with Modified Conditional Feature

At this stage, we aim to harness diffusion models' powerful data generation capability for restoring video frames. Initially, ELPNet's pretrained decoder produces dimension-aligned decoding features as conditioning to guide restoration. This establishes meaningful associations between the target view and the rectified feature view, enhancing the diffusion model's holistic image understanding to improve detail generation fidelity. An auto-alignment strategy is adopted throughout to ensure alignment between decoded features and corresponding generation content. By effectively utilizing decoded features' contextual information, this adjusted alignment strategically guides the generation process.

Specifically, the predictor's main network adopts a U-Net [50] architecture comprising encoder, middle, and decoder steps. The input D_i^P first undergoes 2D convolutions and Mish activations to extract suitable features. Next, within the Resblocks, cross-attention fuses the pseudo ground-truth features with the denoiser's intermediates, guiding the network to produce accurate predictions. This is formulated as:

$$\begin{aligned} p(x_{t-1} | x_t, F_k), &= N(x_{t-1}; \mu(x_t, F_k), \sigma_t^2 I), \\ D_{en0}^P &= D_{enc0}(F_k, D_i^P), k = 4, \\ D_{en1}^P &= D_{enc1}(F_k, D_{en0}^P), k = 3, \\ D_{en2}^P &= D_{enc2}(F_k, D_{en1}^P), k = 2, \\ D_{en3}^P &= D_{enc3}(F_k, D_{en2}^P), k = 1. \end{aligned} \quad (11)$$

Formula (11) demonstrates our latent image features guiding diffusion model generation toward high detail retention. Multi-resolution image features ensure the model obtains adequate guidance under varying receptive fields for improved representations. Moreover, our guidance derives from the designed prior frequency-domain blocks, enriching textures and sharpening salient patterns. Consequently, the architecture's detail restoration and generation capabilities significantly improve. Specifically, time t is sinusoidally position-encoded as t_e and embedded via multilayer perceptrons (MLPs) [51]. Every encoder step has two conditional prediction blocks (CPBs) and a downsampling block where 2D convolutions with a stride of 2 are employed to halve the size of the feature map. Each decoder step contains two CPBs without cross-attention and an upsampling block, doubling the size via transposed convolutions. Applying two-dimensional convolution on decoder outputs reconstructs the predicted noise value $\delta\epsilon$ to recover x_{t-1} over T iterations, generating the restored frame.

4.3. Multi-Scale Group-Wise Information Fusion

Since the output of the conditional neural network belongs to the latent image feature distribution, and the output of the deep diffusion model belongs to the conditional probability distribution, there is a large domain discrepancy between them. If they are directly linearly or nonlinearly combined, the desired performance results cannot be obtained. The existing methods, such as those in [43,52], that fuse convolutional neural networks with diffusion models directly fuse features from the two domains with gaps, which will inevitably lead to image distortions and detail losses. Therefore, how to organically and concisely achieve the fusion of the two has become a universally recognized challenge. This method proposes a simple and innovative solution.

According to the difference between the two domains, we have designed two different fusion paths and finally set up a reasonable network module to fuse them, which ensures effective alignment of their features. As shown in Figure 4, we use the diffusion denoising network to extract multi-scale features and fuse them with ELPNet's features. For the

denoising backbone of this paper, its extended part contains four convolutional layers, with the output feature size ranging from $(8C, H/8, W/8)$ to (C, H, W) . We use a multi-scale feature fusion module to fuse the feature information of the four stages. Eventually, these features are summed and sequentially fed into the fusion head, producing the final result $\hat{Y}_t \in \mathbb{R}^{3 \times H \times W}$. Specifically, three dilated convolutions with different dilation rates ($r = 1, 2, 4$) are applied to map the high-dimensional combination of the two branches to a 3-channel output. Each pixel is acquired by convolutions with $3 \times 3, 5 \times 5$, and 7×7 receptive fields, using Leaky ReLU as the activation function.

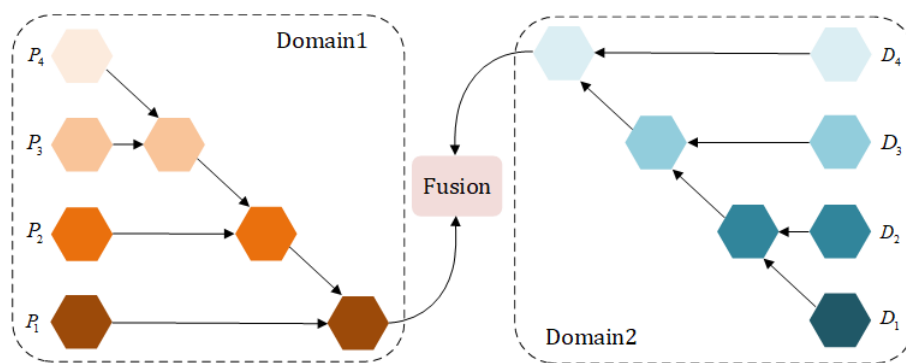


Figure 4. The structure of the fusion module. The left half is the image domain features obtained from the neural network, the right half is the probability distribution features obtained from the diffusion model, and the center represents the fusion of the heterogeneous information.

Using a simple linear weighting method may not result in more enriched semantic representations. The features extracted by different models may overlap and contain redundancies, and directly combining them could exacerbate this issue, ultimately causing a decline in model efficiency. This implicit cross-domain fusion circumvents direct feature interaction across domains; specifically, by introducing an implicit layer, it ensures that the aggregation of information does not hinder the flow of information between different domains. This allows the final features to interact and fuse in a carefully designed common space, enabling information from different domains to complement each other while maintaining their independence, greatly mitigating the negative impacts of mismatch. This strategy helps prevent anomalous uncertain restorations in the outputs. Essentially, this succinct and controllable fusion technique yields more continuous, coherent, and logical restored details. By better achieving our targeted task, it generates more realistic and naturalistic results.

5. Experiments

5.1. Dataset

We chose to utilize the widely acknowledged MFQEv2 [11] standard dataset within the realm of image and video compression for training our pre-trained models, the ELPNet and the conditional noise prediction network (CNP). Subsequently, we conducted evaluations to assess the effectiveness of our approach. This dataset encompasses 126 video sequences sourced from Xiph, VQEG, and JCT-VC [53], spanning diverse content and resolutions, establishing it as a robust benchmark for evaluating algorithmic robustness. Adhering to prevailing evaluation standards in this domain, we adhered to a training set–test set ratio of 6:1 for data partitioning. All video sequences underwent compression processing at three different compression rates (QP values of 27, 32, 37, and 42) using HM16.20 and HEVC LowDelay-P (LDP) configurations. Elevated compression rates correspond to more pronounced compression distortions. The utilization of various compression rates enables a comprehensive evaluation of the method’s recovery and generalization capabilities across different levels of compression distortion. In our algorithmic comparative experiments, we conducted an impartial assessment, taking into account the impact of content complexity, resolution, and compression rate on image and video quality.

5.2. Experiment Settings

In our research, we developed a model consisting of two key networks: ELPNet, responsible for extracting information from compressed videos to recover corrupted frames, and a conditional noise predictor, a diffusion model network based on the U-Net architecture, for performing the final video frame restoration. Both networks are designed to receive 64 input channels ($C = 64$). During the training phase of the model, we randomly crop small blocks of 128×128 pixels from compressed videos, which serve as training samples to simulate the original data. To enhance the model's robustness in handling video jitter, we applied a series of data augmentation operations to the training dataset, including random rotation and flipping. We used the Adam optimization algorithm to update the parameters of the conditional noise predictor, where the learning-rate-related hyperparameters δ_1 and δ_2 were set to 0.9 and 0.999. In the training process of the diffusion model, we empirically set the forward and backward diffusion steps to 1000 steps. Additionally, the selection of noise sequences β_1, \dots, β_T followed the recommendations in the literature. At the beginning of training, the learning rate was set to 1×10^{-4} and decreased to one-tenth after completing 70% of the iteration cycles. All experiments were conducted on a high-performance server equipped with an Intel Core i9-13900K CPU, 64 GB of memory, and two NVIDIA® GeForce RTX 4090 GPUs (NVIDIA, Santa Clara, CA, USA), using PyTorch 2.0.0, Python 3.9, CUDA 11.8, and CuDNN 8.6.0. Building upon the method put forth in [54], this paper implements a repair approach with unconstrained dimensions. In the evaluation process, we used two main performance metrics to quantify the improvement in video quality: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). These standardized metrics allow us to accurately measure and compare the effect of the proposed model on enhancing the quality of video frames.

5.3. Comparisons with Previous Algorithms

We presented optimal results on the MFQEv2 dataset, including ARCNN [23], DnCNN [24], MFQEv2 [11], STDF-R3L [5], and RFDA [12]. The results of several of these methods are cited from the relevant literature, and the relevant parameters are strictly configured according to the authors' recommendations in their publications. Recently the BasicVSR++ [17] method has demonstrated state-of-the-art performance on several video restoration tasks [55]. Considering that the official version of BasicVSR++ was pre-trained and fine-tuned on other datasets, for a fair comparison, we re-trained BasicVSR++ on the MFQEv2 datasets (QP32 and QP37), keeping the same experimental setup as other Baseline methods.

5.3.1. Qualitative Visual Effect Comparison

Our method yields visually satisfying results, as depicted in Figure 5, highlighting its exceptional capability to restore intricate details and textures within enhanced frames. In comparison to alternative methods, our restoration outcomes closely align with the ground truth, devoid of issues like excessive smoothing and detail loss. This robustly affirms the effectiveness of our method in rectifying details and texture information in damaged images. Notably, in the BasketballPass sequence, it is evident that contours and object boundary details lost during the compression process are effectively reinstated in our results. The Racehorses sequence similarly showcases this effect, illustrating the preservation of details and textures. The robust capability of our method for detail reconstruction is attributed to the innovative design of the model architecture. The incorporation of detail/texture-sensitive components in the loss function and the integration of a multi-scale sub-network empower the network to adeptly learn how to reconstruct rich and realistic details from contextual information within damaged regions. This presents a robust and effective solution for enhancing the quality of detail and texture restoration in image recovery tasks.

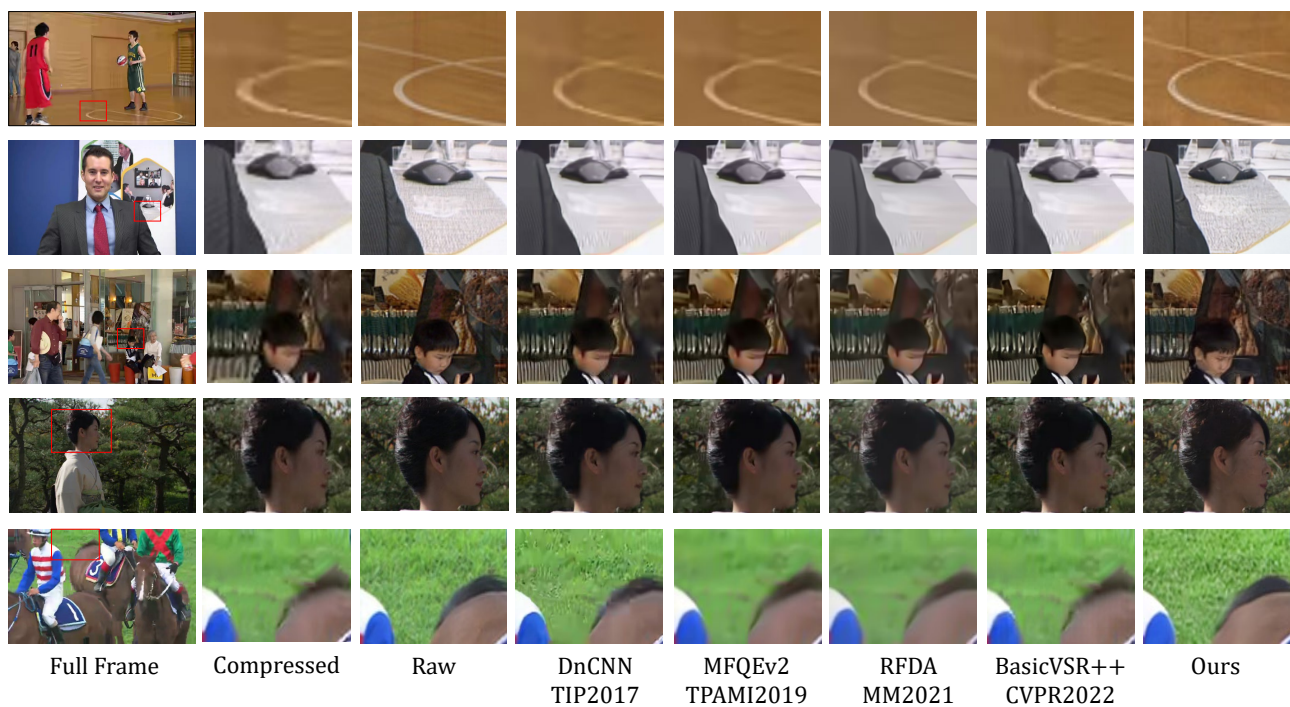


Figure 5. Subjective comparison results between state-of-the-art methods and our proposed method in five video sequences at QP = 37. Test video names (from top to bottom): BasketballPass, Johanny, BQMall, Kimono, and Racehorses. The zoom-in of red box area is shown.

5.3.2. Quality Fluctuation

Fluctuations in video quality serve as critical evaluation metrics [11]. Random variations in quality can result in significant temporal inconsistencies and a diminished user experience. We utilize Standard Deviation (SD) and Peak–Valley Difference (PVD) [56] to quantify the quality fluctuations for each test sequence. Table 1 presents the average PVD and SD values for different methods across all test sequences. The results indicate that our proposed method exhibits the smallest average PVD and SD. This suggests that, in comparison to other baseline methods, our approach demonstrates smaller quality fluctuations, contributing to a more stable enhancement effect. Furthermore, Figure 6 depicts four PSNR curves for various test sequence groups, representing the original HEVC compressed sequence, RFDA, BasicVSR++, and our method’s processed sequences. It is evident that, when compared to alternative methods, our approach achieves significantly improved performance on compressed frames, demonstrating the lowest fluctuation amplitude.

Table 1. Averaged PVD/SD of test sequences for PSNR at QP = 27, 32, and 37.

Method	QP27	QP32	QP37
HEVC	1.07/0.83	1.38/0.82	1.42/0.79
ARCNN	1.07/0.83	1.38/0.82	1.44/0.80
DnCNN	1.06/0.83	1.40/0.83	1.44/0.80
MFQEv2	0.77/0.74	0.98/0.70	0.96/0.67
RFDA	0.63/0.61	0.70/0.63	0.69/0.61
BasicVSR++	—	0.73/0.67	0.71/0.66
STCF	0.57/0.58	0.62/0.59	0.61/0.61
Ours	0.59/0.58	0.57/0.55	0.54/0.53

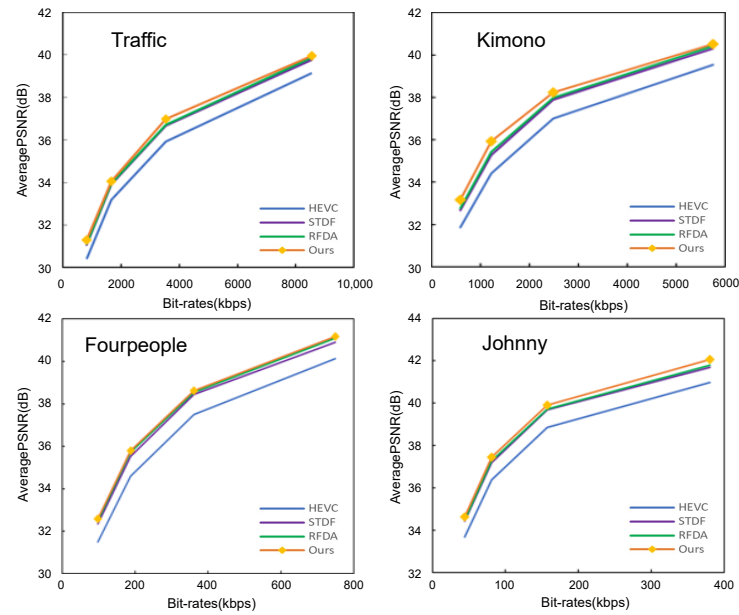


Figure 6. Rate–distortion curves of four test sequences.

5.3.3. Rate–Distortion Performance

In comparison to other methods, we conducted a comprehensive evaluation of the rate–distortion performance of our proposed approach. Figure 7 illustrates the rate–distortion curves for our method and other state-of-the-art methods on four selected sequences. The observation reveals that, at similar bit rates, our method consistently attains a higher PSNR compared to other methods, indicating its superior rate–distortion performance.

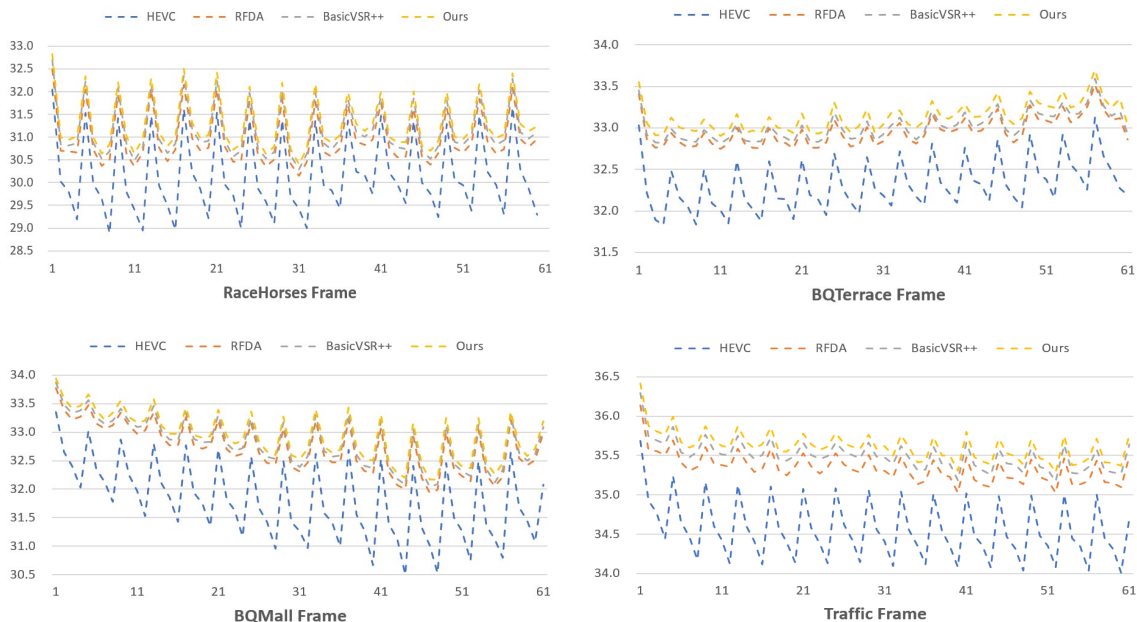


Figure 7. PSNR curves of HEVC, RFDA, BasicVSR++, and ours on four test sequence Cactus at QP = 37.

5.3.4. Overall Performance

Table 2 illustrates the overall improvement of our method in terms of PSNR and SSIM metrics. The results indicate that, regardless of the QP value, our method surpasses other state-of-the-art methods in terms of average metric improvement. For instance, compared to BasicVSR++, we achieve an improvement of 0.13–0.20 dB in PSNR. When contrasted

with STCF, our method exhibits a PSNR improvement ranging from 0.02 to 0.06 dB, with a more pronounced enhancement in SSIM. Unlike BasicVSR++ with a bidirectional motion compensation mechanism and STCF's 7-video-frame restoration approach, our method enhances the target frame by exploring richer texture details and global contextual information through adjacent frame fusion. This is attributed to the targeted design of our diffusion model based on prior latent feature modulation and the group-wise domain fusion module. The extensive experimental results validate the overall superiority of our method in the task of compressed video restoration.

Table 2. Overall performance comparison in terms of Δ PSNR(dB)/ Δ SSIM($\times 10^{-2}$) over the test sequences at four QPs. Video resolutions: Class A (2560 \times 1600), Class B (1920 \times 1080), Class C (832 \times 480), Class D (480 \times 240), Class E (1280 \times 720). Bold indicates best performance.

QP	Sequences	ARCNN	DNCNN	MFQEv2.0	STDF-R3L	RFDA	BasicVSR++	TVQE	STCF	Ours
		[23]	[24]	[11]	[5]	[12]	[17]	[16]	[28]	LFDM
A	Traffic	0.24/0.47	0.24/0.57	0.59/1.02	0.73/1.15	0.80/1.28	0.94/1.52	0.88/1.44	0.91/1.44	1.04/1.64
	PeopleonStreet	0.35/0.75	0.41/0.82	0.92/1.57	1.25/1.96	1.44/2.22	1.37/2.23	1.49/2.33	1.62/2.43	1.58/2.37
B	Kimono	0.22/0.65	0.24/0.75	0.55/1.18	0.85/1.61	1.02/1.86	1.41/2.18	0.99/1.82	1.21/1.94	1.52/2.26
	ParkScene	0.14/0.38	0.14/0.50	0.46/1.23	0.59/1.47	0.64/1.58	0.86/2.25	0.66/1.76	0.74/1.79	0.95/2.30
	Cactus	0.19/0.38	0.20/0.48	0.50/1.00	0.77/1.38	0.83/1.49	0.62/1.51	0.85/1.57	0.93/1.61	0.82/1.61
C	BQTerrace	0.20/0.28	0.20/0.38	0.40/0.67	0.63/1.06	0.65/1.06	0.71/1.25	0.74/1.34	0.75/1.25	0.82/1.38
	BasketballDrive	0.23/0.55	0.25/0.58	0.47/0.83	0.75/1.23	0.87/1.40	1.02/1.53	0.85/1.46	1.09/1.59	1.06/1.74
	RaceHorses	0.22/0.43	0.25/0.65	0.39/0.80	0.55/1.35	0.48/1.23	0.76/1.84	0.61/1.59	0.69/1.59	0.86/1.84
D	BQMall	0.28/0.68	0.28/0.68	0.62/1.20	0.99/1.80	1.09/1.97	1.17/2.24	1.06/2.02	1.25/2.21	1.24/2.32
	PartyScene	0.11/0.38	0.13/0.48	0.36/1.18	0.68/1.94	0.66/1.88	0.44/1.71	0.80/2.27	0.73/2.28	0.78/2.36
	BasketballDril	0.25/0.58	0.33/0.68	0.58/1.20	0.79/1.49	0.88/1.67	0.87/1.67	0.98/2.01	0.96/1.76	0.89/1.88
E	RaceHorses	0.27/0.55	0.31/0.73	0.59/1.43	0.83/2.08	0.85/2.11	1.02/2.74	0.86/2.30	1.02/2.47	1.17/2.90
	BQSquare	0.08/0.08	0.13/0.18	0.34/0.65	0.94/1.25	1.05/1.39	0.61/0.93	1.25/1.74	1.06/1.48	1.02/1.57
	BlowingBubbles	0.16/0.35	0.18/0.58	0.53/1.70	0.74/2.26	0.78/2.40	0.69/2.65	0.83/2.60	0.80/2.53	0.85/2.62
E	BasketballRass	0.26/0.58	0.31/0.75	0.73/1.55	1.08/2.12	1.13/2.24	1.22/2.66	1.12/2.41	1.32/2.63	1.30/2.73
	FourPeople	0.37/0.50	0.39/0.60	0.73/0.95	0.94/1.17	1.13/1.36	1.13/1.38	1.16/1.42	1.11/1.33	1.20/1.42
	Johnny	0.25/0.10	0.32/0.40	0.60/0.68	0.81/0.88	0.90/0.94	0.99/0.97	1.12/1.33	1.00/1.13	1.06/1.25
E	KristenAndSara	0.41/0.50	0.42/0.60	0.75/0.85	0.97/0.96	1.19/1.15	1.20/1.13	1.27/1.23	1.12/1.11	1.15/1.21
	Average	0.23/0.45	0.26/0.58	0.56/1.09	0.83/1.51	0.91/1.62	0.95/1.80	0.98/1.82	1.02/1.81	1.08/1.93
	42	Average	0.29/0.96	0.22/0.77	0.59/1.65	0.76/2.04	0.82/2.20	— / —	0.99/2.64	0.88/2.34
32	Average	0.18/0.19	0.26/0.35	0.52/0.68	0.86/1.04	0.87/1.07	0.89/1.25	0.93/1.24	1.07/1.32	1.09/1.55
27	Average	0.18/0.14	0.27/0.24	0.49/0.42	0.72/0.57	0.82/0.68	— / —	0.87/0.80	1.05/0.88	1.03/1.17

5.4. Ablation Study

5.4.1. The effect of ELPNet and fusion in participation

In this section, the results of ablation experiments convincingly demonstrate a significant improvement in the performance of the restoration network when the features from the ELPNet are integrated, as compared to using either the diffusion model alone or the ELPNet in isolation. As shown in Table 3, when the features extracted by the ELPNet are not included, the PSNR and SSIM indices of the diffusion model are noticeably lower. Similarly, when only the ELPNet is utilized for restoration, there is a significant decrease in performance due to the inability to leverage the diffusion model to generate missing image structures. Ultimately, the complete network, after integrating ELPNet features, achieves the optimal improvement in PSNR and SSIM (1.08/1.93). This underscores that the prior latent features extracted by the ELPNet provide crucial guidance for the diffusion

model, resulting in the generation of higher-fidelity restoration results through fusion. The synergy between the two components mutually enhances the final image quality. Therefore, incorporating the ELPNet structure in the restoration network is deemed essential, playing an indispensable role in improving restoration effectiveness. The experimental results validate that a single model struggles to achieve a balance between preserving fine details and maintaining overall structural coherence. In this context, feature fusion provides a valuable avenue for complementary enhancement.

Table 3. The impact of ELPNet’s involvement on PSNR and SSIM within the test sequences.

Method	Fusion Scheme	Δ PSNR	Δ SSIM
Diffusion-only	—	0.78	1.40
ELPNet-only	—	0.51	1.32
Diffusion and ELPNet	Cross-attention	0.96	1.67
Diffusion and ELPNet	Cross-attention and fusion	1.08	1.93

Figure 8 depicts subjective comparison images of our method with and without ELPNet involvement in the diffusion model. It is evident that, upon introducing the prior latent features extracted by the ELPNet, the generated results progressively align with the real image, showcasing an enhancement in texture details.

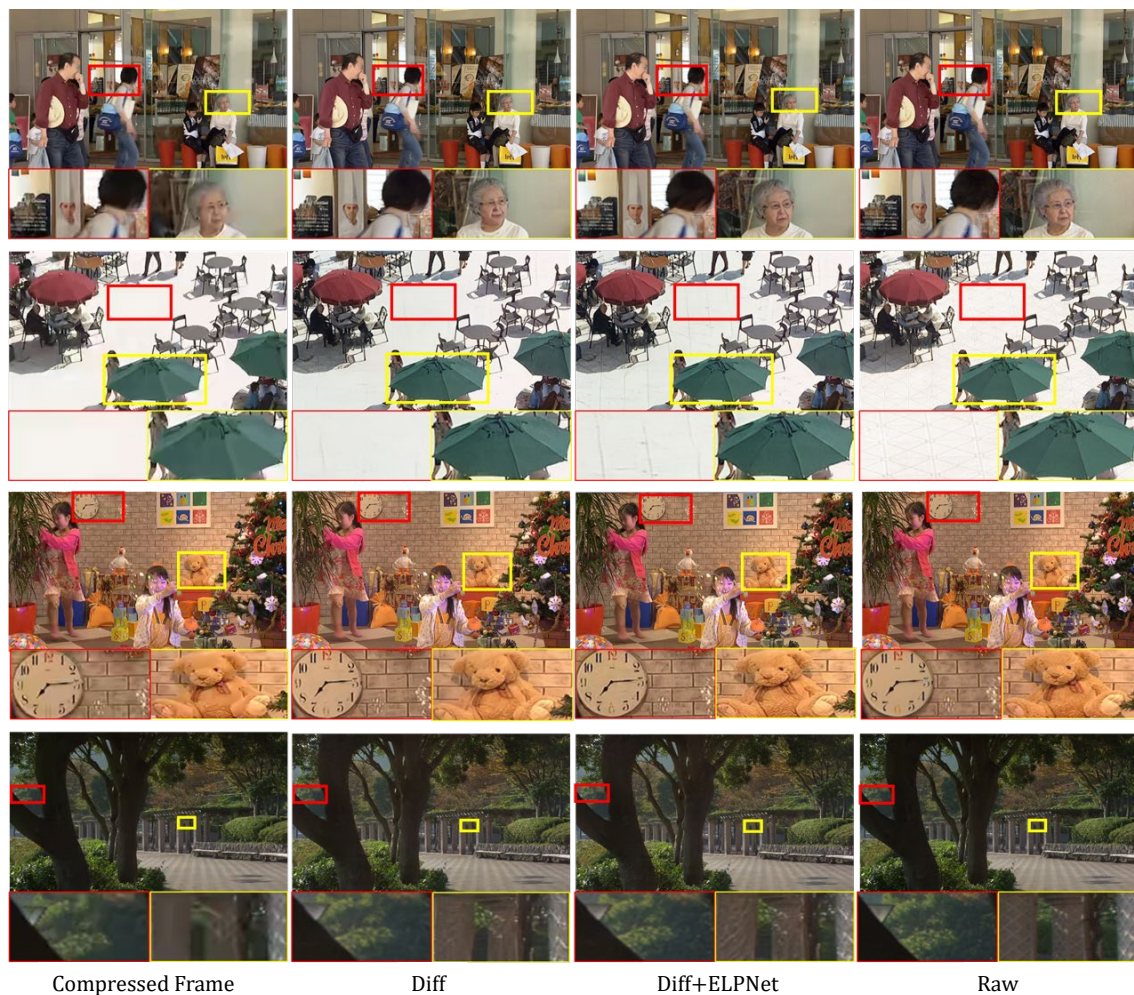


Figure 8. Subjective comparison images depicting the restoration with and without ELPNet intervention. The zoom-in of red box and yellow box area is shown.

5.4.2. The significance of DWT

To thoroughly substantiate the pivotal role of Discrete Wavelet Transform (DWT) in augmenting image restoration quality, we conducted a comparative analysis of the network's performance before and after the integration of the DWT module. As depicted in Table 4, the experimental findings distinctly showcase a significant enhancement in various evaluation metrics for ELPNet with DWT, compared to the standard network lacking the DWT module. Notably, the PSNR metric exhibited an increase of 0.08, while the SSIM metric witnessed a noteworthy improvement of 0.27. The significant improvement lies in the substantial increase in SSIM metrics, particularly noticeable in the reconstruction of texture details, highlighting the critical role of frequency-domain information in reinstating high-frequency content in damaged images. This robustly affirms that the DWT module empowers the network to assimilate frequency domain prior knowledge, thereby producing visually richer and more realistic texture effects. Consequently, it can be conclusively stated that the introduction of wavelet transforms is pivotal for elevating the quality of image restoration. The DWT module devised in this study assumes an indispensable role in the reconstruction of details and texture information.

Table 4. The effects of DWT and various loss functions on PSNR and SSIM for test sequences. ✓ indicates that the feature or component was enabled, while × signifies that it was not enabled.

Method	L_{cha}	L_{ssim}	L_{per}	Δ PSNR	Δ SSIM
w/o DWT ELPNet	✓	×	×	0.40	0.97
ELPNet	✓	×	×	0.48	1.24
ELPNet	✓	✓	×	0.49	1.29
ELPNet	✓	✓	✓	0.51	1.32

5.4.3. Addition of loss function

In addition, we underscore the importance of the employed loss functions in this study. Upon scrutiny of rows two to four in Table 4, it becomes evident that each loss plays an effective and pivotal role in enhancing both PSNR and SSIM. The Charbonnier loss offers pixel-level supervision, while the perceptual loss guarantees that the output consistently aligns with the ground truth within the deep feature space. Through the comprehensive integration of all losses during the training phase, our model attains optimal performance.

6. Conclusions

We propose a novel LFDN approach, completing compressed video damage restoration by designing a neural network combined with sensors and codecs to generate detail-preserving latent features. These judiciously guide the diffusion model to recover fine-grained image information. Specifically, we modulate the diffusion probability distribution by enhancing neural network detail perception using Discrete Wavelet Transforms. Cross-attention is particularly effective for guiding the model's probability distribution features. Additionally, considering the domain discrepancy between neural networks and diffusion models, our simple yet effective group-wise domain fusion module integrates both to mitigate detail losses and distortions. This substantially boosts model performance. Systematic experiments on public datasets verify our model's superiority over other state-of-the-art models. Moving forward, this method can be integrated with the High-Efficiency Video Coding (HEVC) standard to restore compression-induced quality degradation during the post-processing stage. This would provide the industry with a practical video restoration solution to significantly improve the visual quality of compressed images.

Author Contributions: Conceptualization, W.W. (Weiran Wang) and M.J.; methodology, W.W. (Weiran Wang); software, W.W. (Weiran Wang); validation, W.W. (Weiran Wang) and M.J.; formal analysis, M.J.; investigation, Y.F.; resources, Y.F.; data curation, M.J.; writing—original draft preparation, W.W. (Weiran Wang); writing—review and editing, W.W. (Weiran Wang), M.J. and W.W. (Wei Weng); visualization, W.W. (Wei Weng); supervision, M.J.; project administration, M.J.; funding acquisition, M.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rijkse, K. H.263: Video coding for low-bit-rate communication. *IEEE Commun. Mag.* **1996**, *34*, 42–45. [CrossRef]
2. Telecom, I. T. U. Advanced Video Coding for Generic Audiovisual Services. ITU-T Recommendation H. 264. 2003. Available online: <https://www.itu.int/rec/T-REC-H.264> (accessed on 15 January 2024).
3. Sullivan, G.J.; Ohm, J.R.; Han, W.J.; Wiegand, T. Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *22*, 1649–1668. [CrossRef]
4. Zeng, K.; Zhao, T.; Rehman, A.; Wang, Z. Characterizing perceptual artifacts in compressed video streams. In Proceedings of the SPIE Proceedings, Human Vision and Electronic Imaging XIX, San Francisco, CA, USA, 2–6 February 2014. [CrossRef]
5. Deng, J.; Wang, L.; Pu, S.; Zhuo, C. Spatio-Temporal Deformable Convolution for Compressed Video Quality Enhancement. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 10696–10703. [CrossRef]
6. Yeo, H.; Jung, Y.; Kim, J.; Shin, J.; Han, D. Neural adaptive content-aware internet video delivery. In *Operating Systems Design and Implementation, Operating Systems Design and Implementation*; OmniScriptum S.R.L.: London, UK, 2018.
7. Yin, X.; Jindal, A.; Sekar, V.; Sinopoli, B. A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP. In Proceedings the 2015 ACM Conference on Special Interest Group on Data Communication, London, UK, 17–21 August 2015. [CrossRef]
8. Foi, A.; Katkovnik, V.; Egiazarian, K. Pointwise Shape-Adaptive DCT for High-Quality Denoising and Deblocking of Grayscale and Color Images. *IEEE Trans. Image Process.* **2007**, *6*, 1395–1411. [CrossRef]
9. Zhang, X.; Xiong, R.; Fan, X.; Ma, S.; Gao, W. Compression Artifact Reduction by Overlapped-Block Transform Coefficient Estimation With Block Similarity. *IEEE Trans. Image Process.* **2013**, *22*, 4613–4626. [CrossRef]
10. Yang, R.; Xu, M.; Liu, T.; Wang, Z.; Guan, Z. Enhancing quality for HEVC compressed videos. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 2039–2054. [CrossRef]
11. Guan, Z.; Xing, Q.; Xu, M.; Yang, R.; Liu, T.; Wang, Z. MFQE 2.0: A new approach for multi-frame quality enhancement on compressed video. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 949–963. [CrossRef] [PubMed]
12. Zhao, M.; Xu, Y.; Zhou, S. Recursive fusion and deformable spatiotemporal attention for video compression artifact reduction. In Proceedings of the 29th ACM international conference on multimedia, Virtual Event, China, 20–24 October 2021; pp. 5646–5654.
13. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. SwinIR: Image Restoration Using Swin Transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Virtual Conference, 11–17 October 2021. [CrossRef]
14. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient Transformer for High-Resolution Image Restoration. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022. [CrossRef]
15. Zhang, X.; Yang, S.; Luo, W.; Gao, L.; Zhang, W. Video compression artifact reduction by fusing motion compensation and global context in a swin-CNN based parallel architecture. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 3489–3497.
16. Yu, L.; Chang, W.; Wu, S.; Gabbouj, M. End-to-End Transformer for Compressed Video Quality Enhancement. In *IEEE Transactions on Broadcasting*; IEEE: New York, NY, USA, 2023.
17. Chan, K.C.; Wang, X.; Yu, K.; Dong, C.; Loy, C.C. Basicvsr: The search for essential components in video super-resolution and beyond. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4947–4956.
18. Chan, K.C.; Zhou, S.; Xu, X.; Loy, C.C. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5972–5981.

19. Dhariwal, P.; Nichol, A. Diffusion Models Beat GANs on Image Synthesis. *Neural Inf. Process. Syst. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
20. Ho, J.; Saharia, C.; Chan, W.; Fleet, D.J.; Norouzi, M.; Salimans, T. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.* **2022**, *23*, 2249–2281.
21. Liu, D.; Li, Y.; Lin, J.; Li, H.; Wu, F. Deep learning-based video coding: A review and a case study. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 1–35. [[CrossRef](#)]
22. Dong, C.; Deng, Y.; Loy, C.; Tang, X. Compression Artifacts Reduction by a Deep Convolutional Network. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
23. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155. [[CrossRef](#)] [[PubMed](#)]
24. Dai, Y.; Liu, D.; Wu, F. A Convolutional Neural Network Approach for Post-Processing in HEVC Intra Coding. In *MultiMedia Modeling, Lecture Notes in Computer Science*; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 28–39. [[CrossRef](#)]
25. He, X.; Hu, Q.; Zhang, X.; Zhang, C.; Lin, W.; Han, X. Enhancing HEVC compressed videos with a partition-masked convolutional neural network. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; IEEE: New York, NY, USA, 2018; pp. 216–220.
26. Ding, D.; Kong, L.; Chen, G.; Liu, Z.; Fang, Y. A Switchable Deep Learning Approach for In-loop Filtering in Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*; IEEE: New York, NY, USA, 2020; p. 1. [[CrossRef](#)]
27. Xue, T.; Chen, B.; Wu, J.; Wei, D.; Freeman, W.T. Video Enhancement with Task-Oriented Flow. *Int. J. Comput. Vis.* **2019**, *127*, 1106–1125. [[CrossRef](#)]
28. Yang, R.; Sun, X.; Xu, M.; Zeng, W. Quality-gated convolutional LSTM for enhancing compressed video. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; IEEE: New York, NY, USA, 2019; pp. 532–537.
29. Wang, J.; Xu, M.; Deng, X.; Shen, L.; Song, Y. MW-GAN+ for Perceptual Quality Enhancement on Compressed Video. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 4224–4237. [[CrossRef](#)]
30. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015.
31. Welling, M.; Teh, Y. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In Proceedings of the International Conference on Machine Learning, Honolulu, HA, USA, 18–21 December 2011.
32. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *Neural Inf. Process. Syst. Inf. Process. Syst.* **2020**, *33*, 6840–6851.
33. Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D.J.; Norouzi, M. Image Super-Resolution Via Iterative Refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 64713–4726. [[CrossRef](#)]
34. Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; Chen, Y. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing* **2022**, *479*, 47–59. [[CrossRef](#)]
35. Whang, J.; Delbracio, M.; Talebi, H.; Saharia, C.; Dimakis, A.; Milanfar, P. Deblurring via Stochastic Refinement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
36. Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; Van Gool, L. Diffir: Efficient diffusion model for image restoration. *arXiv* **2023**, arXiv:2303.09472.
37. Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; Van Gool, L. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022. [[CrossRef](#)]
38. Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; Yoon, S. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021. [[CrossRef](#)]
39. Han, X.; Zheng, H.; Zhou, M. Card: Classification and regression diffusion models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 18100–18115.
40. Wollleb, J.; Sandkühler, R.; Bieder, F.; Valmaggia, P.; Cattin, P.C. Diffusion models for implicit image segmentation ensembles. In Proceedings of the International Conference on Medical Imaging with Deep Learning, PMLR, Zurich, Switzerland, 6–8 July 2022; pp. 1336–1348.
41. Wu, J.; Fu, R.; Fang, H.; Zhang, Y.; Yang, Y.; Xiong, H.; Liu, H.; Xu, Y. Medsegdiff: Medical image segmentation with diffusion probabilistic model. *arXiv* **2022**, arXiv:2211.00611.
42. Chen, S.; Sun, P.; Song, Y.; Luo, P. Diffusiondet: Diffusion model for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 19830–19843.
43. Han, L.; Zhao, Y.; Lv, H.; Zhang, Y.; Liu, H.; Bi, G.; Han, Q. Enhancing remote sensing image super-resolution with efficient hybrid conditional diffusion model. *Remote. Sens.* **2023**, *15*, 3452. [[CrossRef](#)]
44. Wu, C.; Wang, D.; Bai, Y.; Mao, H.; Li, Y.; Shen, Q. HSR-Diff: Hyperspectral image super-resolution via conditional diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 7083–7093.

45. Xu, L.; He, G.; Zhou, J.; Lei, J.; Xie, W.; Li, Y.; Tai, Y.W. Transcoded Video Restoration by Temporal Spatial Auxiliary Network. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 2875–2883. [[CrossRef](#)]
46. Mallat, S. A theory for multiresolution signal decomposition: the wavelet representation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*; IEEE: New York, NY, USA, 1989; pp. 674–693. [[CrossRef](#)]
47. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–16 July 2017; pp. 624–632.
48. Snell, J.; Ridgeway, K.; Liao, R.; Roads, B.D.; Mozer, M.C.; Zemel, R.S. Learning to generate images with perceptual similarity metrics. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: New York, NY, USA, 2017; pp. 4277–4281.
49. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 694–711.
50. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
51. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
52. Baranchuk, D.; Rubachev, I.; Voynov, A.; Khrukov, V.; Babenko, A. Label-Efficient Semantic Segmentation with Diffusion Models. *arXiv* **2021**, arXiv:2112.03126.
53. Lainema, J.; Bossen, F.; Han, W.J.; Min, J.; Ugur, K. Intra coding of the HEVC standard. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *22*, 1792–1801. [[CrossRef](#)]
54. Wang, Y.; Yu, J.; Zhang, J. Zero-shot image restoration using denoising diffusion null-space model. *arXiv* **2022**, arXiv:2212.00490.
55. Yang, R. NTIRE 2021 challenge on quality enhancement of compressed video: Methods and results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 647–666.
56. Xu, Y.; Gao, L.; Tian, K.; Zhou, S.; Sun, H. Non-Local ConvLSTM for Video Compression Artifact Reduction. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27–28 October 2019. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.