Using Conditional Video Compressors for Image Restoration

Yi-Hsin Chen, Yen-Kuan Ho, Ting-Han Lin, Wen-Hsiao Peng, and Ching-Chun Huang
National Yang Ming Chiao Tung University, Taiwan

***Abstract*—To address the ill-posed nature of image restoration tasks, recent research efforts have been focused on integrating conditional generative models, such as conditional variational autoencoders (CVAE). However, how to condition the autoencoder to maximize the conditional evidence lower bound remains an open issue, particularly for the restoration tasks. Inspired by the rapid advancements in CVAE-based video compression, we make the first attempt to adapt a conditional video compressor for image restoration. In doing so, we have the low-quality image to be enhanced, which plays the same role as the reference frame for conditional video coding. Our scheme applies scalar quantization in training the autoencoder, circumventing the difficulties of training a large-size codebook as with prior works that adopt vector-quantized VAE (VQ-VAE). Moreover, it trains end-to-end a fully conditioned autoencoder, including a conditional encoder, a conditional decoder, and a conditional prior network, to maximize the conditional evidence lower bound. Extensive experiments confirm the superiority of our scheme on denoising and deblurring tasks.**

***Index Terms*—image restoration, learning-based video compression, image deblurring, image denoising**

I. Introduction

Image restoration involves repairing and/or enhancing images to restore their original quality. Deep learning-based approaches with convolutional neural networks [5]–[8] have surpassed conventional restoration methods. Emerging techniques, as discussed in [9], employ vision transformers (ViTs) to address the rather limited receptive field of CNNs. Notably, Zamir *et al.* [10] proposed Restormer, which combines a UNet structure with enhanced Transformer blocks for improved feature aggregation and transformation. These non-generative methods extract features from low-quality input images and restore their quality by decoding the refined features.

However, the restoration task is often an ill-posed problem; that is, the mapping from the original, high-quality input into its distorted version is multiple-to-one. The non-generative methods, which inherently assume that such a mapping is one-to-one, fail to formulate the problem properly. Recent research has increasingly been focused on utilizing conditional generative models, like conditional GANs, diffusion models [11], and VQ-VAEs [4], to address the restoration task.

Inspired by the recent breakthrough in conditional video coding, we present in this paper the first attempt to explore its potential for image restoration tasks. Specifically, we adopt a conditional variational auto-encoder-based (CVAE-based) compression backbone [12] for our image restoration tasks, adapting its design to suit our needs.

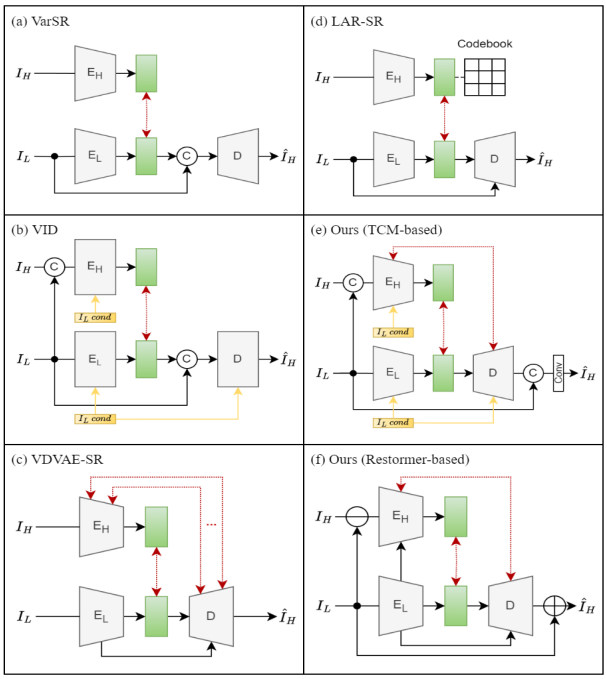

Fig. 1. Comparison of CVAE-based image restoration schemes and our proposed method: (a) VarSR [1], (b) VID [2], (c) VDVAE-SR [3], (d) LAR-SR [4], (e) Ours (TCM-based) and (f) Ours (Restormer-based). The red arrows in each sub-figure represent the evaluation of the KL divergence.

Conditional variational auto-encoders (CVAEs) have been used in various image restoration tasks [1]–[4], [13]–[16]. The training of most CVAEs is to maximize the conditional evidence lower bound:

$$\mathcal{L}(I_H, I_L, \theta, \phi) = \log E_{q_\phi(y|I_H,I_L)} \log p_\theta(I_H|y, I_L) - \mathrm{KL}(q_\phi(yI_H, I_L)||p_\theta(y|I_L)), \quad (1)$$

where I_H denotes a high-quality image, I_L represents a low-quality image. $p_\theta(y|I_L)$ indicates the prior distribution modeled by the prior encoder E_L. Conversely, $q_\phi(y|I_L, I_H)$ represents the posterior distribution modeled by the encoder E_H, which takes I_H as input and conditions the latent generation on I_L. Lastly, $p_\theta(I_H|y, I_L)$ denotes the decoding distribution implemented by the decoder D. The decoding of the latent y is likewise conditioned on I_L. At training time, y is sampled from $q_\phi(y|I_L, I_H)$, whereas at test time, it is sampled from the conditional prior $p_\theta(y|I_L)$.

The existing CVAE-based restoration models differ in how these components are implemented. For example, Hyun *et al.* [1] introduced VarSR-Net, a CVAE-based super-resolution network. As shown in Fig. 1(a), the lack of conditioning the posterior $q_\phi(y|I_H)$ on I_L impedes the minimization of the KL divergence between the posterior and the conditional prior $p_\theta(y|I_L)$. This issue is not seen in VID [2] (Fig. 1(b)), as it generates a condition signal (a density estimation map to avoid over and under deraining) from I_L, which is shared across the encoder E_H, prior encoder E_L, and decoder D for the deraining task. Furthermore, Chira *et al.* [3] proposed VDVAE-SR, a hierarchical CVAE with K layers of latents conditionally dependent on each other (Fig. 1(c)), offering a prior and a posterior of the form $p_\theta(y|I_L) = \prod_{j=1}^{K} p_\theta(y_j|y_{<j}, I_L)$ and $q_\phi(y|I_H, I_L) = \prod_{j=1}^{K} q_\phi(y_j|y_{<j}, I_H, I_L)$, respectively. Their CVAE with layers of latent is shown to improve the model expressiveness. Recently, Guo *et al.* [4] proposed a VQ-VAE-based model, known as LAR-SR (Fig. 1(d)). One issue of the VQ-VAE model is how to determine the codebook size. As shown in [17], too large a codebook can have many codewords not used at all. Another issue is that their two-stage training procedure–i.e. the reconstruction and KL divergence losses in Eq. (1) are applied sequentially and separately–is not ideal in terms of maximizing the evidence lower bound. In this paper, we explore the CVAE architecture (Fig. 1(e)) of a conditional video codec, termed DCVC-TCM [12], for our image restoration tasks. DCVC-TCM is a powerful CVAE designed as a fully conditional VAE, with conditions applied in E_H, E_L, D. We also make a slight adjustment, turning it into a hierarchical CVAE (Fig. 1(e)), with prior and posterior distributions computed in the deepest and middle layers, where the second-level distribution is conditioned on the latent representation of the preceding layer. Notably, compared to other methods where the posterior q_ϕ follows a Gaussian distribution, our q_ϕ follows a uniform distribution, consistent with the original design in video compression. Moreover, it is trained end-to-end and utilizes a hyperparameter λ to trade-off between the reconstruction and KL divergence losses in Eq. (1). The reconstruction loss determines crucially the best achievable reconstructed image quality, while the KL divergence loss is strongly connected to the image restoration quality at inference time. This approach avoids the challenges associated with training a large codebook, as seen in VQ-VAE. This work also explores an alternative CVAE architecture that utilizes Restormer [10] as the backbone (Fig. 1(f)). Our contributions are as follows:

- We make the first attempt to utilize a conditional video compression model as an image restoration framework.
- We fully condition the encoder, decoder, and the prior distribution on the low-quality image and train the entire system end-to-end to maximize the conditional evidence lower bound.
- We demonstrate that a good conditional video codec has the potential to perform comparably to or better than state-of-the-art restoration techniques in terms of both quantitative and qualitative metrics.

II. Proposed Method

Our work leverages a conditional video compression model to perform image restoration. Section II-A presents an overview of our system. Section II-B delves into the estimation of the prior and posterior distributions. Lastly, we introduce our training methodology in Section II-C.

A. System Overview

Fig. 2 illustrates our CVAE architecture. It comprises a conditional encoder E_H, a prior encoder E_L, and a conditional decoder D. These three components are conditioned on the features generated by the feature pyramid E_F. Given a pair of high- and low-quality images $I_H \in \mathbb{R}^{H\times W\times 3}$, $I_L \in \mathbb{R}^{H\times W\times 3}$, the feature pyramid E_F (Fig. 2(d)) generates the condition signals $c_1 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times N}, c_2 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times N}, c_3 \in \mathbb{R}^{H\times W\times N}$ from I_L, where N represents the number of channels. The conditional encoder E_H (Fig. 2(a)) encodes the high-quality image I_H into the content latents $y_1 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times M}$ and the kernel latents $y_2 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times N}$, based on I_L, where N and M represent the numbers of channels. The content latents are a compact set of features capturing conditionally the information about I_H according to I_L, while the kernel latents act as a correction term to compensate for the predicting error of y_1 in the decoder D at inference time (Section II-B). In addition, the prior encoder E_L encodes I_L (Fig. 2(b)) into its latent representation $\tilde{y}_1 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times M}$. During training, the conditional decoder D (Fig. 2(c)) reconstructs the high-quality image $\hat{I}_H \in \mathbb{R}^{H\times W\times 3}$ by updating $\tilde{y}_1$ and F based on the content y_1 and kernel y_2 latents, respectively. At inference time, the high-quality image I_H is not available. As such, $\tilde{y}_1$ and F are updated from samples drawn from their respective priors, respectively. The decoder incorporates the content and kernel latent blocks [18] (Figs. 2(e) and (f)) to minimize the KL divergence between the posterior and prior distributions. More details are provided in Section II-B.

To sum up, at training time, we start by encoding the input image I_H with the conditional encoder E_H and sampling features from the posterior distribution in the kernel and content latent blocks of the decoder D to reconstruct the input image $\hat{I}_H$. The quality of the reconstructed image $\hat{I}_H$ depends heavily on the distortion term in the evidence lower bound (i.e. Eq. (1)). At inference time, we start from the prior encoder E_L and sampling features from the prior in the kernel and content latent blocks to generate the restored image $\tilde{I}_H \in \mathbb{R}^{H\times W\times 3}$ from I_L. The restoration quality is controlled by the KL divergence in Eq. (1).

B. Posterior and Prior Distributions

This section details how we model the prior and posterior distributions in the content and kernel latent blocks. Minimizing the KL divergence between these two distributions is crucial to ensure high-quality restored images. In the content latent block (Fig. 2(f)), the left branch corresponds to the

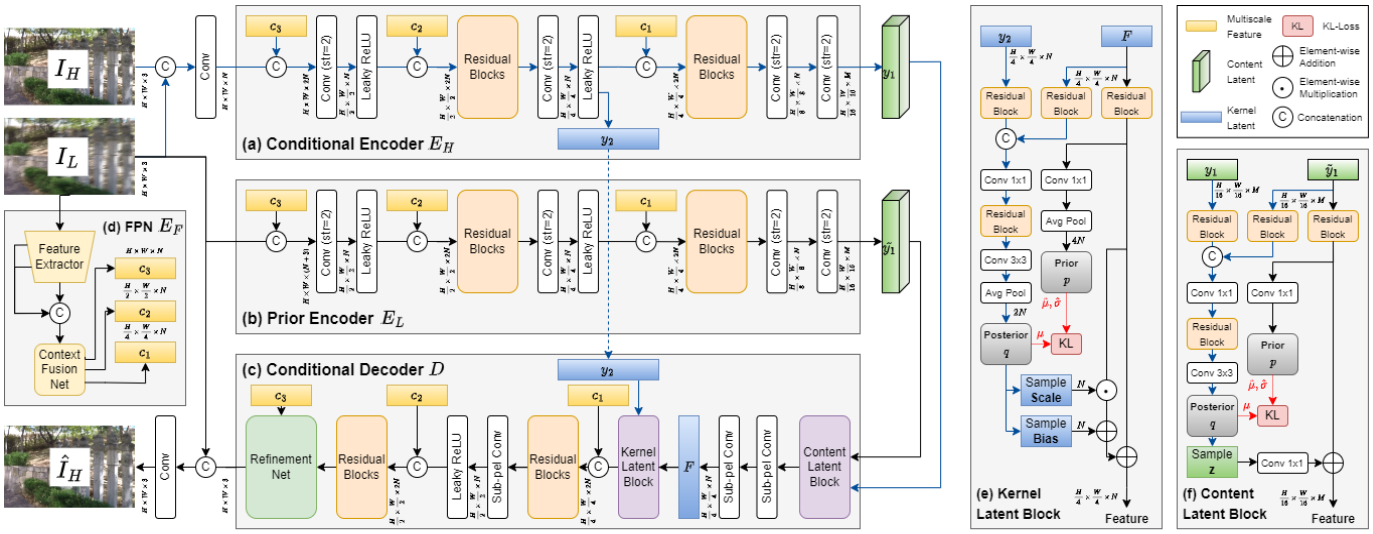

Fig. 2. Illustration of the proposed method. (a) The conditional encoder E_H converts I_H into its content latent representation y_1 and kernel latent representation y_2 given the condition I_L. (b) The prior encoder E_L converts I_L into $\tilde{y}_1$. (c) The conditional decoder D generates the reconstructed image $\hat{I}_H$ at training time. (d) The feature pyramid network E_F extracts multi-scale features from I_L as the condition signals. (e) The kernel latent block updates F based on y_2. (f) The content latent block updates $\tilde{y}_1$ based on y_1.

posterior distribution, which is assumed to follow a uniform distribution:

$$q_{\phi,1}(z) \triangleq U\left(\mu(y_1,\tilde{y}_1)-\frac{1}{2},\mu(y_1,\tilde{y}_1)+\frac{1}{2}\right), \tag{2}$$

where $\mu(y_1,\tilde{y}_1)$ denotes a function of y_1 and $\tilde{y}_1$. Note that the posterior $q_{\phi,1}$ is derived from the combination of y_1 and $\tilde{y}_1$, rather than solely from y_1. This design, inspired by Duan *et al.* [18], makes the minimization of the KL divergence relatively easy, as opposed to formulating $q_{\phi,1}$ from y_1 only. The intuition is that it depends more directly on the only variable $\tilde{y}_1$ from which the prior $p_{\theta,1}$ is constructed. The latent variable $z \in \mathbb{R}^{H\times W\times N}$ sampled from $q_{\phi,1}$ is then utilized to update $\tilde{y}_1$ in the content latent block.

The middle branch in Fig. 2(f) corresponds to the prior distribution $p_{\theta,1}$, which is assumed to follow a Gaussian distribution:

$$p_{\theta,1} \triangleq \mathcal{N}\left(\tilde{\mu}(\tilde{y}_1),\tilde{\sigma}(\tilde{y}_1)^2\right), \tag{3}$$

where $\tilde{\mu}(\tilde{y}_1)$ and $\tilde{\sigma}(\tilde{y}_1)$ represent the mean and standard deviation, respectively. Both outputs are functions of $\tilde{y}_1$.

During inference, the latent variable z is sampled from the prior distribution to decode the restored image $\tilde{I}_H$. If the prior does not approximate the posterior well, the mismatch error in the drawn sample z will propagate to the subsequent layers in the decoder, resulting in poor restoration quality.

To mitigate this issue, a kernel latent variable y_2 is introduced in the shallow layer (which is closer to the decoder output) in the decoder to rectify the mismatch error in z of the content latent block. The design of the kernel latent block is similar to that of the content latent block. One notable difference is that we perform average pooling along the spatial dimension to derive channel-wise $scale \in \mathbb{R}^N$ and $bias \in \mathbb{R}^N$ to update the main feature F in the channel dimension. The ablation study in Table VI confirms the benefit of introducing the content and kernel latent blocks.

C. Training Objective

Our training objective is to learn a CVAE by maximizing the evidence lower bound. That is, to minimize

$$\begin{aligned} L = \lambda \cdot \mathrm{D}(I_H,\hat{I}_H) + \mathrm{KL}\left(q_{\phi,1}(y_1,\tilde{y}_1))||p_{\theta,1}(\tilde{y}_1)\right) \\ +\mathrm{KL}\left(q_{\phi,2}(y_2,F)||p_{\theta,2}(F)\right), \end{aligned} \tag{4}$$

where $\hat{I}_H \in \mathbb{R}^{H\times W\times 3}$ indicates the reconstructed image. D represents the mean squared error between the ground truth I_H and $\hat{I}_H$. KL terms represent the KL divergence between the posteriors $q_{\phi,i}$ and the priors $p_{\theta,i}$ in the content and kernel latent blocks, where $i = 1, 2$. λ is a hyperparameter to balance the two losses. Table V explains how λ affects the performance.

III. Experimental Results

A. Settings

We apply our model to two restoration tasks: image deblurring and denoising. We evaluate the quality of the restored images both quantitatively and qualitatively. We adopt PSNR and SSIM to compare the image quality of our method against non-generative models. We further utilize LPIPS, DISTS, and FID to compare our method with the generative models.

B. Training Details

We implement our method using two CVAE backbones, TCM-based and Restormer-based models. We train both models end-to-end. The TCM-based model is trained from scratch with the Adam optimizer (betas=(0.9, 0.999)), with batch size

TABLE I
QUANTITATIVE RESULTS OF **SINGLE-IMAGE MOTION DEBLURRING** ON **GOPRO** DATASET

Method	PSNR↑	SSIM↑
IR-SDE [21]	30.7	0.901
MPRNet [5]	32.66	0.959
MIMO-UNet++ [22]	32.68	0.959
MAXIM-3S [23]	32.86	0.961
Restormer [10]	32.92	0.961
Ours (TCM-based)	**33.13**	**0.979**
Ours (Restormer-based)	32.90	0.960

TABLE II
QUANTITATIVE RESULTS OF **SINGLE-IMAGE MOTION DEBLURRING** ON **GOPRO** DATASET

Method	LPIPS↓	DISTS↓	FID↓
Restormer [10]	0.163	0.085	10.626
DiffIR [11]	**0.157**	0.083	9.654
Ours (TCM-based)	0.160	**0.081**	**9.417**
Ours (Restormer-based)	0.163	0.085	10.528

TABLE III
QUANTITATIVE RESULTS OF **SINGLE-IMAGE DEOISING** ON **SIDD** DATASET

Method	PSNR↑	SSIM↑
MPRNet [24]	39.71	0.958
MIRNet [7]	39.72	0.959
Uformer [25]	39.89	0.960
MAXIM-3S [23]	39.96	0.960
Restormer [10]	40.02	0.960
Ours (TCM-based)	39.52	0.961
Ours (Restormer-based)	**40.18**	**0.967**

TABLE IV
QUANTITATIVE RESULTS OF **SINGLE-IMAGE DEOISING** ON **SIDD** DATASET

Method	LPIPS↓	DISTS↓	FID↓
MIRNet [7]	0.3076	0.1513	47.71
MPRNet [24]	0.3062	0.1507	49.55
HINet [6]	0.2974	0.1491	47.38
Restormer [10]	0.2957	0.1480	47.29
Ours (TCM-based)	0.3192	**0.1471**	52.79
Ours (Restormer-based)	**0.2917**	0.1472	**46.07**

set to 4 and a learning rate of 1e-4. The value of the hyper-parameter λ is set to 0.0002. For the Restormer-based model, we start with the pre-trained weights from Restormer [10]. We also take the Adam optimizer with betas=(0.9, 0.999), weight decay set to 1e-4, a batch size of 2, and a learning rate of 5e-4. λ is set to 0.0067.

C. Image Deblurring

For the image deblurring task, we use GoPro dataset [19], which consists of high-resolution images corrupted by non-uniform blind motion blur. Our model is trained using cropped image patches of 256×256 spatial resolution from the training set and evaluated on 1280×720 images in the test set. Table I shows that our TCM-based model outperforms Restormer [10] by 0.2 dB in PSNR and 0.008 in SSIM. From Table II, our TCM-based model shows comparable LPIPS performance to DiffIR [11], which is a diffusion-based generative model. Additionally, it has a 0.237 gain in FID and a 0.002 gain in DISTS. Fig. 3 demonstrates the subjective quality of our restored images.

D. Image Denoising

For the image denoising task, we conduct experiments using Smartphone Image Denoising Dataset (SIDD) [20], which comprises approximately 30,000 high-resolution noisy images captured under various scenes and lighting conditions. These high-resolution images are cropped into 256×256 patches for both training and testing. The results are presented in Tables III and IV. Our method demonstrates superior performance to the other baseline methods. The variant with the Restormer backbone achieves a 0.16 dB gain compared to the original Restormer. Fig. 4 visualizes our restored images.

E. Ablation Study

Our ablation study explores two aspects: (a) the performance under different lambda values and (b) the benefits of using the content and kernel latent blocks. All experiments utilize a lightweight version of the TCM-based model.

Performance under different lambda values. Table V demonstrates how the choice of λ impacts the quality of $\hat{I}_H$ and $\tilde{I}_H$. A higher λ puts more emphasis on the distortion term in Eq. (4). Consequently, y_1 and y_2 contain more information about I_H, improving the quality of $\hat{I}_H$. However, this also makes it more challenging for the priors to accurately approximate the posterior, thereby causing the quality of $\tilde{I}_H$ to decrease at inference time. Conversely, the PSNR of $\tilde{I}_H$ is improved when λ is decreased.

Benefits of using the content and kernel latent blocks. Table VI demonstrates the performance of using the content and kernel latent blocks. The first row represents the experiment where only the content latent is present and the prior encoder outputs directly the Gaussian parameters. In this case, the KL divergence is evaluated by

$$\mathrm{KL}\left(U(y_1-\frac{1}{2},y_1+\frac{1}{2})\ ||\ \ \mathcal{N}(\tilde{\mu}(\tilde{y}_1),\tilde{\sigma}(\tilde{y}_1)^2)\right) \qquad (5)$$

The results in Table VI show that both the content and kernel latent blocks [18] are able to improve the restoration quality.

IV. CONCLUSION

Recognizing the limitations of existing CVAE-based image restoration methods, such as suboptimal conditional schemes in CVAE and the inability of VQ-VAE to fully optimize the codebook due to its two-stage training, we make the first attempt to adopt a conditional video codec for image restoration. Extensive experiments confirms that our approach performs comparably to or even better than the state-of-the-art restoration techniques in terms of both quantitative and qualitative quality.

Fig. 3. Subjective quality comparison on the deblurring task. Our method is able to better restore objects (e.g., cars in the scene) as compared to Restormer.

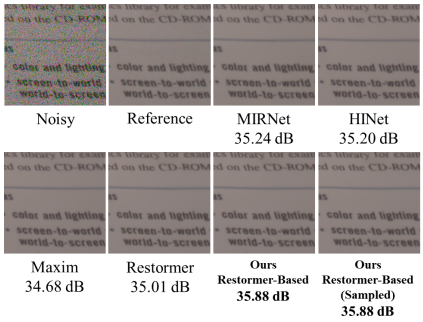

Fig. 4. Subjective quality comparison of our method and others on the denoising task.

TABLE V
PERFORMANCE UNDER DIFFERENT λ VALUES: $L = \lambda \cdot \mathrm{D} + \mathrm{KL}$

λ	KL	PSNR of $\hat{I}_H$	PSNR of $\tilde{I}_H$
0.18(high)	0.52	43.40	27.21
0.093	0.42	41.85	26.65
0.002(low)	**0.14**	**33.01**	**29.03**

TABLE VI
THE EFFECT OF THE CONTENT AND KERNEL LATENT BLOCKS

base	content block	kernel block	PSNR
v			33.3/30.8 (+0/+0)
v	v		33.5/31.1 (+0.2/+0.3)
v		v	33.6/31.3 (+0.3/+0.5)
v	v	v	**34/31.8 (+0.7/+0.1)**

REFERENCES

[1] S. Hyun and J.-P. Heo, "Varsr: Variational super-resolution network for very low resolution images," in *European Conference on Computer Vision*. Springer, 2020, pp. 431–447.

[2] Y. Du, J. Xu, Q. Qiu, X. Zhen, and L. Zhang, "Variational image deraining," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2406–2415.

[3] D. Chira, I. Haralampiev, O. Winther, A. Dittadi, and V. Liévin, "Image super-resolution with deep variational autoencoders," in *European Conference on Computer Vision*. Springer, 2022, pp. 395–411.

[4] B. Guo, X. Zhang, H. Wu, Y. Wang, Y. Zhang, and Y.-F. Wang, "Lar-sr: A local autoregressive model for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1909–1918.

[5] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 821–14 831.

[6] L. Chen, X. Lu, J. Zhang, X. Chu, and C. Chen, "Hinet: Half instance normalization network for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 182–192.

[7] S. W. Zamir, A. Arora, S. H. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *European Conference on Computer Vision*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:263784964

[8] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," *arXiv preprint arXiv:1903.10082*, 2019.

[9] F.-J. Tsai, Y.-T. Peng, Y.-Y. Lin, C.-C. Tsai, and C.-W. Lin, "Stripformer: Strip transformer for fast image deblurring," in *European Conference on Computer Vision*. Springer, 2022, pp. 146–162.

[10] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.

[11] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," *arXiv preprint arXiv:2303.09472*, 2023.

[12] X. Sheng, J. Li, B. Li, L. Li, D. Liu, and Y. Lu, "Temporal context mining for learned video compression," *IEEE Transactions on Multimedia*, 2022.

[13] I. Gatopoulos, M. Stol, and J. M. Tomczak, "Super-resolution variational auto-encoders," *arXiv preprint arXiv:2006.05218*, 2020.

[14] S. Cai, X. Liang, S. Cao, L. Yan, S. Zhong, L. Chen, and X. Zou, "Powerful lossy compression for noisy images," *arXiv preprint arXiv:2403.14135*, 2024.

[15] B. Brummer and C. De Vleeschouwer, "On the importance of denoising when learning to compress images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2440–2448.

[16] J. Peng, D. Liu, S. Xu, and H. Li, "Generating diverse structure for image inpainting with hierarchical vq-vae," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 775–10 784.

[17] F. Mentzer, D. Minnen, E. Agustsson, and M. Tschannen, "Finite scalar quantization: Vq-vae made simple," *arXiv preprint arXiv:2309.15505*, 2023.

[18] Z. Duan, M. Lu, J. Ma, Y. Huang, Z. Ma, and F. Zhu, "Qarv: Quantization-aware resnet vae for lossy image compression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[19] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3883–3891.

[20] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1692–1700.

[21] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Image restoration with mean-reverting stochastic differential equations," *arXiv preprint arXiv:2301.11699*, 2023.

[22] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4641–4650.

[23] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxim: Multi-axis mlp for image processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5769–5780.

[24] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 14 821–14 831.

[25] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 683–17 693.