All-in-One Image Compression and Restoration

Huimin Zeng[1]   Jiacheng Li[1]   Ziqiang Zheng[2]   Zhiwei Xiong[1,*]

[1]University of Science and Technology of China
[2]The Hong Kong University of Science and Technology

Abstract

Visual images corrupted by various types and levels of degradations are commonly encountered in practical image compression. However, most existing image compression methods are tailored for clean images, therefore struggling to achieve satisfying results on these images. Joint compression and restoration methods typically focus on a single type of degradation and fail to address a variety of degradations in practice. To this end, we propose a unified framework for all-in-one image compression and restoration, which incorporates the image restoration capability against various degradations into the process of image compression. The key challenges involve distinguishing authentic image content from degradations, and flexibly eliminating various degradations without prior knowledge. Specifically, the proposed framework approaches these challenges from two perspectives: i.e., content information aggregation, and degradation representation aggregation. Extensive experiments demonstrate the following merits of our model: 1) superior rate-distortion (RD) performance on various degraded inputs while preserving the performance on clean data; 2) strong generalization ability to real-world and unseen scenarios; 3) higher computing efficiency over compared methods. Our code is available at `https://github.com/ZeldaM1/All-in-one`.

1. Introduction

Image compression, which facilitates the efficient transmission and storage of image data, has served as a fundamental part of modern image data processing pipelines. Recently, deep learning-based image compression methods [23, 24, 28, 36, 42] have shown remarkable compression ratio improvement, demonstrating the superiority and flexibility over traditional standards [5, 6, 53, 56]. In the practical scenarios (*e.g.*, object detection [10, 17, 57] and autonomous driving [52, 58]) where image compression is employed, the captured images are likely to be plagued by various degradations (*e.g.*, weather-related degradations,

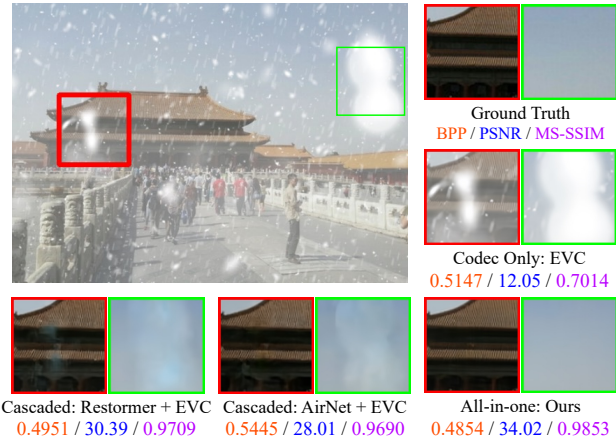*Corresponding author (zwxiong@ustc.edu.cn).

Figure 1. Results of typical solutions for degraded image compression, where BPP/PSNR/MS-SSIM are reported for each method. The image codec EVC (designed for clean images) allocates extra bits to preserve degradations. Cascaded solutions (*e.g.*, Restormer + EVC) amplify artifacts introduced in the restoration stage.

blur, and noise) due to the complex environmental conditions. However, most existing image compression methods [23, 24, 28, 36, 42] are tailored for "clean" images. For degraded images, codecs tend to spend extra bits to faithfully preserve the degradations (*e.g.*, the results of the image codec EVC [23] in Fig. 1), leading to the sub-optimal compression performance and the potential disruption for downstream tasks [62]. Given the fundamental role of image compression in the image processing pipeline, we recognize the critical need to equip the image compression model with the capability of eliminating various degradations.

Cascading independent image restoration (IR) models with compression models provides a straightforward solution for existing image compression methods to handle the degraded input images (*e.g.*, cascaded Restormer+EVC shown in Fig. 1). However, such kind of solution inevitably increases the overall complexity, resulting in inefficiency and higher requirements for the computational resources. Moreover, errors caused by the IR stage may be propagated and amplified in the subsequent compression stage, leading to error accumulation and visually unsatisfying results [15]. Therefore, there is a growing pref-

erence for joint image restoration and compression solutions. Prior works typically address degradations in terms of noise [7, 15, 21, 41, 50], blur [64], and low-light conditions [9]. A notable limitation of these specialist works is that they are designed for specific types/levels of degradations, overlooking the fact that the captured images may suffer from various degradations in practical scenarios. Consequently, they need to train separate models for each specific degradation, which limits the practicality of these methods. This limitation underlines the need for a more comprehensive solution, which is capable of addressing various degraded images encountered in practical image compression.

In this work, we aim to address the above dilemma from a novel perspective, all-in-one image compression and restoration, which requires the compression model to simultaneously recover degraded images and compress them to reduce file sizes. Additionally, it should be able to handle images corrupted by various types and levels, while maintaining the performance on clean images. The above requirements are supposed to be integrated into a unified network, using the same set of trained weights. Fulfilling these requirements presents two significant challenges for the compression model: 1) to distinguish genuine image content from degradations, ensuring the algorithm prioritizes and preserves the important image content (*e.g.*, edges and textures); and 2) to distinguish and flexibly eliminate these degradations without any degradation priors. Overcoming these challenges is essential for optimizing compression performance, as it ensures the bits are spent on genuine image content instead of encoding the degradations.

To address these challenges, we introduce a unified all-in-one image compression and restoration framework. Corresponding to the challenges above, our method performs two types of information aggregation: 1) content information aggregation, which leverages contextual information to enhance the model's understanding of the image, therefore distinguishing image content from degradations; and 2) degradation representation aggregation, which extracts the discriminative representations of degradations, enabling the model to flexibly eliminate different types of degradations and reconstruct image details. Specifically, the proposed framework consists of an encoder, a decoder, and a spatial entropy model. Both the encoder and decoder employ a hybrid-attention mechanism: the channel-wise group attention (C-GA) and the spatially decoupled attention (S-DA). The C-GA performs group-wise self-attention along the channel dimension, implicitly modeling long-range dependencies and enhancing the ability to differentiate between image content and degradations. Observing that different degradations spatially show distinctive patterns, the S-DA sequentially aggregates discriminative representations from vertical and horizontal directions, thereby distinguishing different degradations and flexibly eliminating

them. The C-GA and S-DA are integrated into the hybrid-attention transformer block (HATB), which is then incorporated into both the encoder and decoder to learn at different scales. Our contributions are summarized as follows:

- We make the first attempt to equip neural image codec with the restoration capability against various degradations, thus achieving visually satisfying results and avoiding the waste of bits on the degradations.

- We propose a unified framework for all-in-one image compression and restoration, which performs two types of information aggregation, effectively distinguishing image content from degradations and discriminating different degradations.

- Experimental results show that our method effectively addresses a wide range of degraded images without sacrificing the rate-distortion (RD) performance on clean data. It also shows strong generalization ability in real-world and unseen scenarios, while exhibiting higher computing efficiency over cascaded solutions.

2. Related Work

2.1. Neural Image Compression

Recent image compression methods [3, 4, 22, 45] have achieved tremendous improvement with auto-regressive models. To address the serial processing problem, He *et al*. [25] introduce a parallelized checkerboard context model, while David *et al*. [46] conduct channel-conditioning and latent residual prediction to reduce serial operations. EVC [23] leverages mask decay and sparsity regularization for efficiency and further improves the RD performance of the scalable encoder. DCVC-FM [36] modulates features with a learnable quantization scaler and periodically refreshing mechanism to support a wide quality range and long prediction chain. Self-attention-based methods [16, 30, 51, 68] develop various self-attention variants to capture non-local information and achieve better RD performance. Mixed architectures of transformer and CNN [42, 69] are further proposed to exploit both global and local information. Given the strong ability of generation, generative methods [1, 2] achieve visually satisfying results with extremely low bitrates. However, these methods are designed for clean data and rarely consider the practical scenario of degraded inputs, inevitably leading to the waste of bits for preserving unnecessary degradations.

2.2. All-in-one Image Restoration

Most image restoration methods [13,34,38–40,59,61,65] are designed to handle a specific type of degradation, while all-in-one image restoration methods aim to manage multiple degradations with a unified network. A majority of them [32, 37, 48, 55, 63] rely on degradation priors to guide the subsequent restoration. Li *et al*. [37] employ multi-head
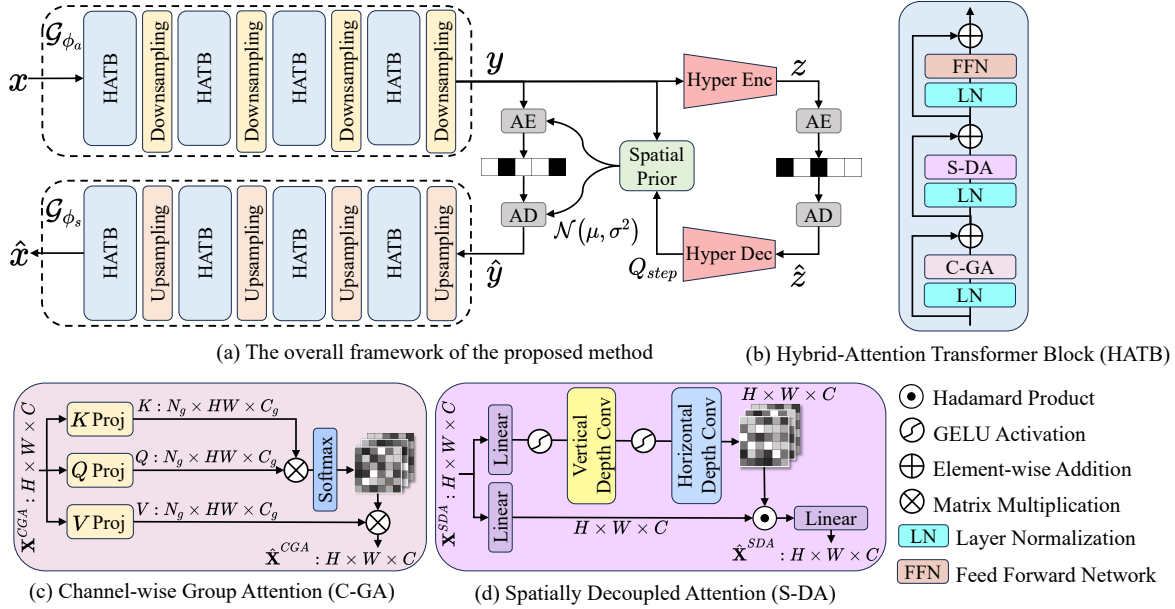
(a) The overall framework of the proposed method

(b) Hybrid-Attention Transformer Block (HATB)

(c) Channel-wise Group Attention (C-GA)

(d) Spatially Decoupled Attention (S-DA)

Figure 2. The proposed all-in-one framework, which consists of a feature encoder $\mathcal{G}\phi_a$, a feature decoder $\mathcal{G}\phi_s$ and a spatial entropy model. The HATB effectively models long-range dependencies with the C-GA, and captures discriminative representations with the S-DA.

encoders to separately embed degraded inputs. NDR [63] develops a degradation query-injection mechanism to effectively approximate and utilize the degradation representations. PromptIR [49] guides the restoration process by providing degradation-related prompts. Chen *et al*. [12] utilize independent teacher networks for different inputs, and perform knowledge distillation for a lightweight unified network. WGWSNet [67] first learns degradation-general representations and expands the parameters for specific degradations. Recent methods [32,48] adopt contrastive encoders to extract more representative degradation priors. However, extracting degradation priors involves complex encoders, posing challenges to efficiency in practical applications.

2.3. Joint Image Compression and Restoration

Nowadays, image compression methods increasingly recognize the need to incorporate the ability of restoration into the compression process. Cheng *et al*. [14] incorporate two add-on modules to equip a pre-trained image decoder with the ability of joint decoding and denoising. Cai *et al*. [8] focus on the low-light scenario and propose a signal-to-noise ratio aware branch to guide joint compression and enhancement. NARV [27] presents an end-to-end noise-adaptive ResNet VAE to handle clean and noisy input images. Nevertheless, these works consider limited degradations (*e.g*., noise and low-light), neglecting the fact that images can be affected by a wide variety of degradations.

3. Method

3.1. Problem Formulation

The proposed unified framework is fundamentally developed for image compression, however, it goes beyond pre-

vious methods [30,42,51,68,69] that are designed for high-quality clean images. Given the constraints of storage and bandwidth, our goal is to remove degradations while preserving essential image information, thereby avoiding the waste of bits on degradations and achieving visually satisfying results. Therefore, our pipeline takes the degraded, large-size image x as input and outputs the clean, compact image $\hat{x}$. This process is achieved through a unified framework (as shown in Fig. 2(a)) and the same set of trained weights. Notably, the network is trained with both degraded images and clean images as inputs, so that it still maintains the ability to compress clean input images.

3.2. Overview

As shown in Fig. 2(a), our framework consists of a feature encoder $\mathcal{G}\phi_a$, a feature decoder $\mathcal{G}\phi_s$, and a spatial entropy model. Given a degraded input image x, the encoder $\mathcal{G}\phi_a$ progressively downsamples the extracted features and obtains the latent representation y, which is then quantized to a discrete representation $\hat{y}$ and encoded into the bit-stream. During decoding, the discrete representation $\hat{y}$ is retrieved from the bit-stream and sent to the decoder $\mathcal{G}\phi_s$, which progressively upsamples the features, reconstructing the decompressed and clean output $\hat{x}$. The overall process is formulated as follows,

$$y = \mathcal{G}_{\phi_a}(x), \quad \hat{y} = \mathcal{Q}(y), \quad \hat{x} = \mathcal{G}_{\phi_s}(\hat{y}), \quad (1)$$

where $\mathcal{G}_{\phi_a}$ and $\mathcal{G}_{\phi_a}$ denote the feature encoder and decoder. $\mathcal{Q}$ indicates the operation that quantizes y using learnable quantization steps to achieve variable bit-rates with a single model. The discrete $\hat{z}$ is obtained by rounding the latent representation z. To model spatial dependencies of discrete representation $\hat{y}$ and accurately estimate the distribu-
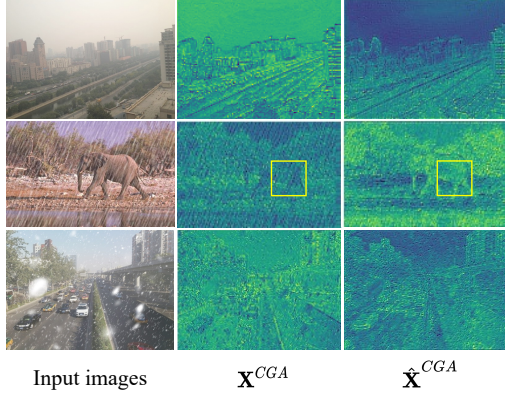
Figure 3. Visualization of the input feature $\mathbf{X}^{CGA}$ and output feature $\hat{\mathbf{X}}^{CGA}$ in C-GA. Although degradations and image signals are closely intertwined in the input features, the C-GA effectively separates degradations from the image content (*e.g.*, the elephant is distinguished from the rain streaks in the yellow box), thereby preserving image signals.

(a) Visualization of feature maps $\hat{\mathbf{X}}^{SDA}$ (b) Visualization of t-SNE
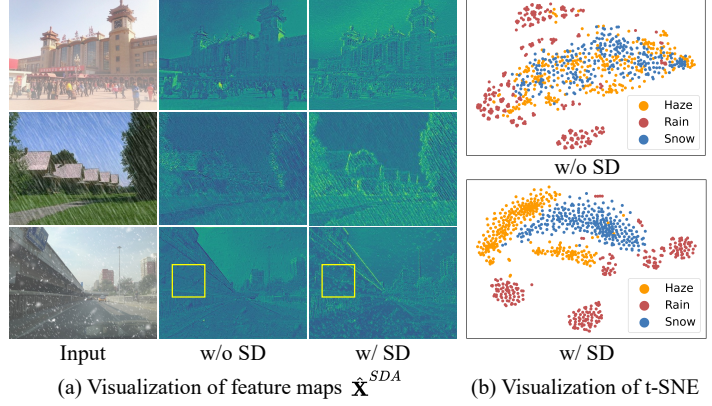
Figure 4. Visual comparisons of output feature $\hat{\mathbf{X}}^{SDA}$ in S-DA and t-SNE results, where SD indicates spatial decoupling. As can be seen, the design of spatial decoupling helps to effectively extract discriminative degradation representations (*e.g.*, the snow spots in the yellow box and distinct clusters in the t-SNE map).

tion $p_{\hat{y}|\hat{z}} \sim \mathcal{N}\left(\mu, \sigma^2\right)$, we adopt the hybrid spatial entropy model [35] to generate parameters μ and σ of the Gaussian model, and estimate the spatial prior for $\hat{y}$. The downsampling and upsampling layers are implemented with a 3×3 convolution layer followed by a pixel-shuffle layer. We elaborate on the hybrid-attention transformer block in Sec. 3.3, and describe the training scheme in Sec. 3.4.

3.3. Hybrid-Attention Transformer Block

As shown in Fig. 2(b), the hybrid attention block integrates transformer-style C-GA for contextual information with controllable complexity, and CNN-style S-DA for discriminative degradation representations with limited receptive fields. The gated-based feed-forward network [65] is adopted to transform extracted features.

Channel-wise group attention. To prioritize and preserve the image content, the unified framework needs to thoroughly understand the input images, identifying valid image signals and discardable content (*e.g.*, degradations and smooth regions). We propose employing the transformer to accomplish these objectives due to its strong ability to capture non-local information. However, a significant challenge raised by the core self-attention mechanism is that the computational complexity increases quadratically with the number of tokens ($\mathcal{O}(N^2)$ for N tokens), resulting in a computing bottleneck. Given that the channel dimension typically contains fewer tokens than the spatial dimension, we implement self-attention along the channel dimension. Such a modification implicitly provides global information for the spatial dimension with reduced complexity, thereby supporting the above image understanding process. As illustrated in Fig. 2(c), given an input feature $\mathbf{X}^{CGA} \in \mathbb{R}^{H \times W \times C}$, where H, W and C denote the height, width and number of channels, respectively, it is initially processed by a separate 1×1 convolutional layer followed

by a 3×3 convolutional layer (denoted as K Proj, Q Proj and V Proj in Fig. 2(c)) to obtain multi-group query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V} \in \mathbb{R}^{N_g \times HW \times C_g}$, where N_g denotes the number of groups and C_g represents the channels per group. For each group, channel-wise self-attention is computed using the following expression,

$$\hat{\mathbf{X}}_i^{CGA} = CGAtt(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = Softmax\left(\frac{\mathbf{Q}_i^T \mathbf{K}_i}{\sqrt{C_g}}\right)\mathbf{V}_i^T, \quad (2)$$

where i denotes the group index, $\mathbf{Q}_i$, $\mathbf{K}_i$ and $\mathbf{V}_i$ indicate the query, key and value tokens of each group. By specifying N_g and C_g, the computational complexity can be adjusted and controlled. By performing group-wise self-attention, the quadratic complexity associated with N_g is further reduced. We visualize the input and output feature maps of C-GA in Fig. 3. As can be seen, despite the input image being significantly degraded, where image features and degradations are too closely intertwined, the C-GA still demonstrates effectiveness in discerning the genuine image content (as shown in the yellow boxes of $\mathbf{X}^{CGA}$ and $\hat{\mathbf{X}}^{CGA}$), therefore preserving essential image information.

Spatially decoupled attention. Despite the effectiveness of C-GA in capturing global information, dealing with finer details requires local spatial interactions. Most importantly, such spatial interactions should contribute to distinguishing various degradations without any degradations priors. We note that different types of degradations exhibit unique spatial patterns, and their differences are accentuated when observed from different directions. For instance, the snow in Fig. 4 appears as spots, whereas rain appears as streaks, which are more anisotropic and demonstrate more shape changes in different directions. This observation motivates us to develop the S-DA, which extracts features from both horizontal and vertical directions to aggregate more distinctive degradation representations. As shown in Fig. 2(d), the

S-DA is a convolutional attention equipped with the local modeling ability to handle finer details. Furthermore, the computational complexity of S-DA grows linearly, making it more resolution-friendly than spatial self-attention mechanisms [18, 26, 44]. For an input feature $\mathbf{X}^{SDA} \in \mathbb{R}^{H \times W \times C}$, the value $\mathbf{V} \in \mathbb{R}^{H \times W \times C}$ is obtained by projecting $\mathbf{X}^{SDA}$ with a linear layer. Meanwhile, a linear layer followed by a vertical and horizontal depth-wise convolution layer is applied on $\mathbf{X}^{SDA}$ to obtain the spatial attention map $\mathbf{A} \in \mathbb{R}^{H \times W \times C}$, which evaluates the importance of each pixel. Then the S-DA is performed as follows,

$$\hat{\mathbf{X}}^{SDA} = SDAtt(\mathbf{V}, \mathbf{A})) = Linear(\mathbf{A} \odot \mathbf{V}), \quad (3)$$

where $\hat{\mathbf{X}}^{SDA}$ denotes the output feature. $Linear$ and $\odot$ denote the linear layer and Hadamard product, respectively. Note that we do not apply Sigmoid activation to the attention map $\mathbf{A}$, as we find out it declines the performance (see Sec. 4.5). The features extracted in S-DA and their t-SNE visualization are included in Fig. 4 (denoted as w/ SD). As shown in Fig. 4(a), the S-DA successfully captures degradations-related representations (*e.g.*, rain streaks and snow spots). Additionally, as illustrated by Fig. 4(b), the extracted representations are notably discriminative, leading to the distinct clusters in the t-SNE map. This characteristic significantly benefits the network to identify and flexibly remove the degradations even under the condition of without degradation priors. We include more analysis regarding the spatial decoupling design in Sec. 4.5. Depth-wise convolution is utilized to implement the vertical and horizontal layers, with kernel sizes of $1 \times K_v$ and $K_h \times 1$.

3.4. Training

Progressive training strategy. Compared with the CNN-based architecture, the attention mechanism benefits from large training patch sizes [19, 54]. To balance the performance and training time, we adopt the progressive training strategy, which involves training the network with small image patches in the earlier stage, and progressively enlarging the patch size in the later stage. Such a strategy allows the network to gradually address inputs of finer details. Furthermore, introducing varying patch sizes throughout the training process also enables the network to adaptively handle images of different sizes.

Loss function. We adopt the following rate-distortion loss as the loss function,

$$\begin{aligned} \mathcal{L} =& \lambda_d \cdot \mathcal{D}(\hat{x}, x^{gt}) + \mathcal{R}(\hat{y}) + \mathcal{R}(\hat{z}) \\ =& \lambda_d \cdot \mathbb{E}\left[\|x^{gt} - \hat{x}\|_{\mathfrak{p}}^{\mathfrak{p}}\right] \\ &- \mathbb{E}\left[\log p_{\hat{y}|\hat{z}}(\hat{y} \mid \hat{z})\right] - \mathbb{E}\left[\log p_{\hat{z}}(\hat{z})\right], \end{aligned} \quad (4)$$

where x^{gt} denotes the ground truth image, and λ_d is the hyperparameter that controls the trade-off between distortion and rate terms. $\mathcal{R}(\hat{y})$ and $\mathcal{R}(\hat{z})$ represent the bit rates of latent discrete representation $\hat{y}$ and $\hat{z}$, respectively. In

Setting	Degradation	Dataset	
		Train	Test
Weather	Haze Snow Rain	RESIDE [33] CSD [11] Rain1400 [20]	RESIDE [33] CSD [11] Rain1400 [20]
Gaussian Noise	$\sigma = 15$ $\sigma = 25$ $\sigma = 50$	Open Images [31]	Kodak [29]

Table 1. Details of dataset settings, where the specific types of degradations and adopted datasets are reported.

Method		FLOPs/G			Speed/ms		
		Sum	Restor.	Compres.	Sum	Restor.	Compres.
Cascaded	Restormer + EVC	178	141	37	804	724	80
	SwinIR + EVC	785	748		3508	3428	
	AirNet + EVC	339	302		1209	1129	
	WGWSNet + EVC	265	228		426	346	
Joint	EVC*	37			80		
Al-in-one	Ours-S	37			169		
	Ours-L	67			281		

Table 2. Computational complexity and inference speed of the compared methods and our models, where Restor. and Compres. denote restoration and compression, respectively. Cascaded solutions are denoted as *restoration+compression*. EVC* denotes converting EVC into a joint solution by training with mixed datasets.

practice, we adopt mean squared error (MSE) loss as the distortion term (*i.e.*, $\mathfrak{p} = 2$).

4. Experiments

4.1. Experimental Settings

As shown in Tab. 1, we consider two types of dataset settings (*i.e.*, the weather degradation setting and the Gaussian noise degradation setting) to evaluate the performance of the proposed method.

Weather degradation setting. This setting mainly includes weather-related degradations, *i.e.*, haze, snow and rain. We also make qualitative comparisons on REVIDE [66], Snow100K [43] and SPA+ [67], which contain realistic hazy, snowy and rainy images.

Gaussian noise degradation setting. This setting contains corruption of multiple levels of Gaussian noise. For evaluation, we compare the proposed method and cascaded solutions on the Kodak dataset [29], using the noise level included for training (*i.e.*, $\sigma = 15, 25, 50$) and unseen noise levels (*i.e.*, $\sigma = 35, 45, 55$). Since the proposed pipeline is inherently an image compression method, during training, we randomly select clean images as input with the probability of 0.2 for two settings.

Compared methods. As shown in Tab. 2, we compare our approach with both cascaded and joint solutions. The cascaded solutions are composed of independent image restoration and compression models. For the weather degradation setting, we consider two types of IR models for a comprehensive comparison: 1) AirNet [32] and WGWS-Net [67], which are developed for all-in-one image restoration, and 2) Restormer [65] and SwinIR [40], which are de-
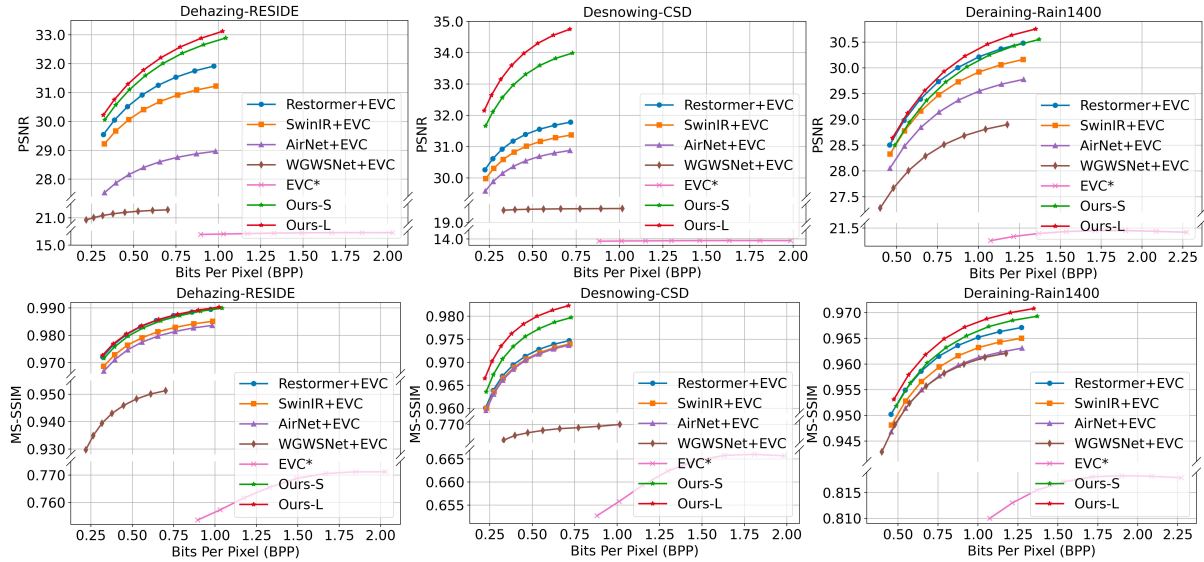
Figure 5. RD performance evaluation on the RESIDE [33], CSD [11] and Rain1400 [20] dataset, where we evaluate the results with both PSNR and MS-SSIM.

signed for specific restoration tasks but serve as strong baselines for image restoration. For the Gaussian noise degradation setting, we select Restormer [65] and AirNet [32] as representative IR methods. We convert these IR models into all-in-one restoration methods by training with mixed datasets. For the compression model, we retrain the Large variant of EVC [23] on the clean datasets of each setting. During evaluation, the degraded inputs are first restored by IR models, and then devoted to EVC [23] for image compression. Since there are rare joint solutions for this all-in-one task, we provide a joint solution (denoted as EVC*) by training EVC [23] with mixed datasets. Notably, directly training EVC with mixed datasets leads to instability and frequent collapse, we therefore reduce the learning rate and train multiple times to obtain the reported results.

Model series. We propose two variants of different complexity, namely Ours-S ($C_g = 32$) and Ours-L ($C_g = 48$). Comparisons of computational complexity and inference speed between the compared methods and our models are provided in Tab. 2, and elaborated in Sec. 4.3.

Evaluation. To quantitatively evaluate the RD performance, we adopt PSNR and MS-SSIM to measure the distortion, and adopt BPP to assess the bitrates.

4.2. Rate-Distortion Performance

Weather degradation setting. RD performance on degraded images is shown in Fig. 5, where the cascaded solutions are referred to as *restoration+compression* (as outlined in Sec. 4.1). As shown by the red curves in Fig. 5, the proposed Ours-L shows superior performance across three benchmarks in comparison with other methods. Compared with the well-performing Restormer+EVC (the blue curves), Ours-L achieves a BD-PSNR for 0.85 dB, 2.49 dB and 0.11 dB on the RESIDE [33], CSD [11] and Rain1400 [20] datasets, respectively. Ours-S (the green

curves) surpasses EVC* by a large margin and outperforms almost all cascaded methods with much fewer FLOPs and higher speed, which is further elaborated in Sec. 4.3.

Gaussian noise degradation setting. We report the RD performance evaluated with PSNR versus BPP in Fig. 6, and include the results evaluated with MS-SSIM in the supplementary materials. As can be seen, Ours-L (the red curves) shows superior performance over compared methods across all noise levels. At the noise level $\sigma = 35$, Ours-L achieves a BD-PSNR of 0.13 dB and 0.51 dB over cascaded AirNet+EVC (the purple curve) and joint EVC* (the pink curve), respectively. The efficient Ours-S (the green curves) demonstrates competitive performance at lower noise levels (*i.e.*, 15, 25 and 35), and much better performance over AirNet+EVC at higher noise levels (*i.e.*, 45, 50 and 55).

Since the proposed framework is intrinsically an image codec, we include the RD performance evaluated on clean Kodak dataset [29] in Fig. 7. As can be seen, Ours-L (the red curves) demonstrates comparable performance with the clean-specific image codec EVC [23] (the blue curves). When evaluated with PSNR, Ours-L exceeds EVC [23] at lower bitrates, and shows only a slight performance drop at higher bitrates, achieving an overall BD-rate improvement of -0.15%. Similarly, Ours-S (the green curves) also shows competitive performance in comparison to EVC [23]. It is worth noting that both degraded and clean images are processed with a single model and the same trained weights. This underlines the superiority of our method to not only effectively address various degraded images, but also to maintain robust RD performance on clean images.

4.3. Efficiency Analysis

To analyze the computing efficiency of the compared methods and our models, we report the FLOPs and inference speed in Tab. 2, which are evaluated with an in-
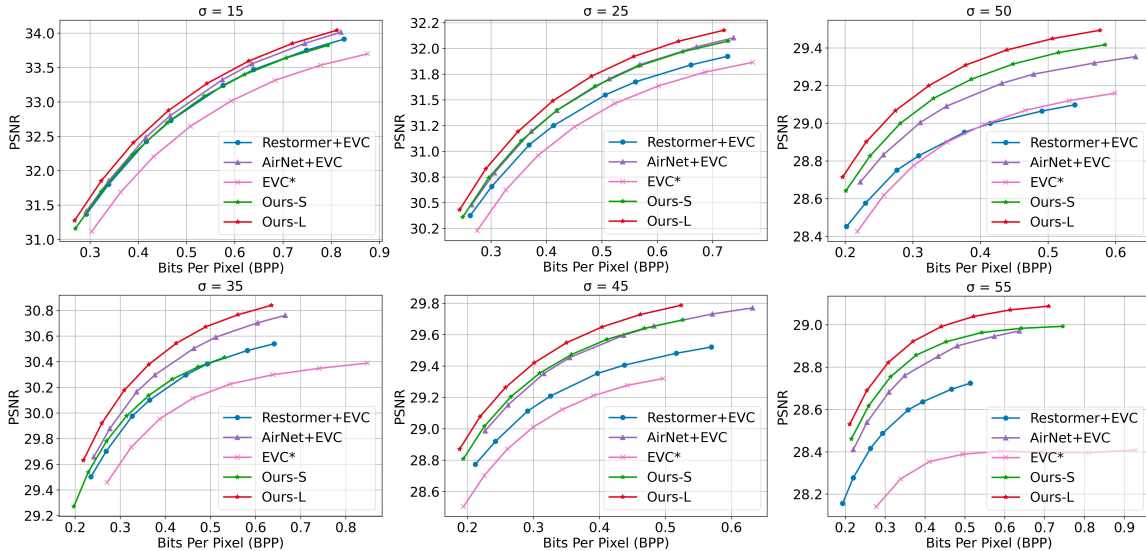
Figure 6. RD performance evaluation on the Kodak dataset [29], where inputs are corrupted by known levels (*i.e.*, 15, 25 and 50) and unknown levels (*i.e.*, 35, 45 and 55) of Gaussian noise. We evaluate the results with PSNR.
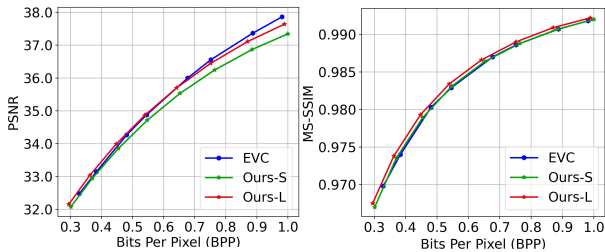

Figure 7. RD performance evaluation on clean Kodak dataset [29], where the results are summarized with both PSNR and MS-SSIM.

put size of 256×256 and 768×512, respectively. Given significantly better performance than the cascaded methods, our models yield much fewer FLOPs and show higher speed. For instance, compared with the strong baseline Restormer+EVC, Ours-L takes up 37.64% of the FLOPs and achieves a $2.86\times$ speedup. Ours-S delivers comparable performance with only 20.79% of FLOPs and achieves a $4.76\times$ speedup. In comparison with AirNet+EVC, Ours-S takes up only 10.91% of FLOPs and achieves a $7.15\times$ speedup, while providing superior RD performance. Despite equipping the same FLOPs with the joint EVC*, Ours-S provides much better RD performance (as shown in Fig. 5 and Fig. 6) and stability during training.

4.4. Qualitative Results

For the weather degradation setting, we provide qualitative comparisons on realistic degraded images in Fig. 8. Qualitative results of synthetic degraded images and Gaussian noise setting are included in the supplementary materials. As shown in Fig. 8, cascaded and joint methods are not effective in removing the degradations in realistic scenarios, and even introduce additional distortion. For instance, the rainy images of the cascaded methods contain unremoved rain streaks. The cascaded SwinIR+EVC introduces arti-

facts for the hazy image. The joint EVC* additionally introduces visually unpleasant noise, which may result from the inherent conflict between compression (preserving image content) and restoration (eliminating degradations). Due to the lack of degradation-specific designs, EVC* struggles to distinguish degradations from valid content, leading to retained degradations and artificial textures/noise in an attempt to enhance "details". In contrast, our method demonstrates superior generalization ability for realistic scenarios.

4.5. Ablation Studies

We start with a baseline model constructed by C-GA, with the configuration of $N_g = 4$ (see supplementary materials). Then, we integrate S-DA into the baseline model to assess the effectiveness of S-DA and spatial decoupling design. We further compare our HATB with two existing popular attention variants to explore its potential. All ablation studies are conducted with Ours-S on the weather degradation setting, and evaluated on the RESIDE dataset [33].

Effectiveness of S-DA. Based on the baseline model with C-GA, we integrate S-DA into the transformer blocks to construct a complete network (denoted as Baseline + S-DA). As depicted by the green curve in Fig. 9, the integration of S-DA leads to a significant improvement of RD performance for the baseline (blue curve), demonstrating the effectiveness of S-DA.

Effectiveness of spatial decoupling design. To demonstrate the benefits of the spatial decoupling design in S-DA, we replace the vertical and horizontal layers with simple depth-wise convolution layers (marked as w/o SD). As depicted in Fig. 9, compared to the complete structure (green curve), discarding the spatial decoupling (orange curve) results in a significant drop in performance. We further provide the visualizations of the features extracted under the condition of with and without spatial decoupling, and their
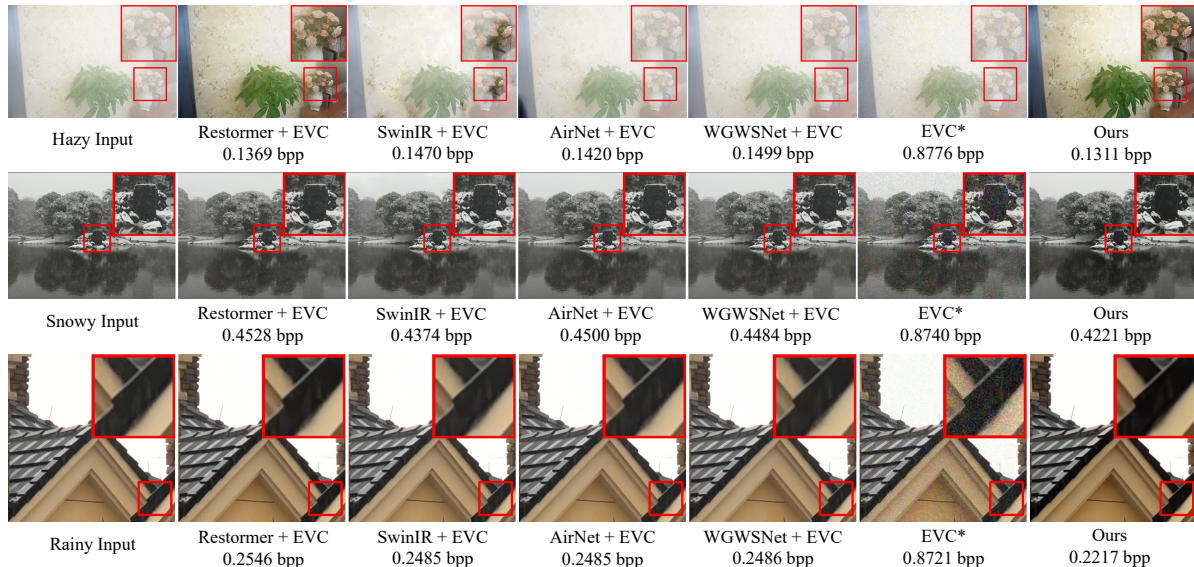
Figure 8. Qualitative comparisons on *realistic* hazy, snowy and rainy images, where cascaded solutions are denoted as *restoration+compression*, and Ours denotes the results of Ours-L. We include BPP for each image.
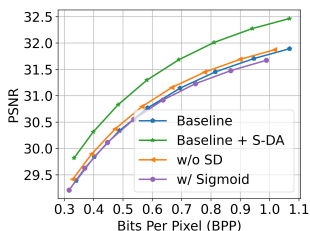

Figure 9. Ablation study on the effectiveness and specific designs of S-DA.
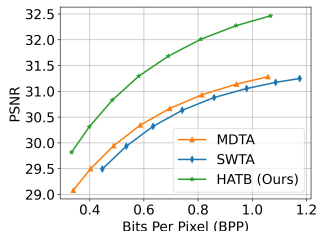

Figure 10. RD performance comparison of different attention variants.

Method	Parameters/M	FLOPs/G	Speed/ms
MDTA	36.39	32	164
SWTA	56.28	48	239
HATB (Ours)	38.39	37	169

Table 3. Complexity of models constructed by different attention variants, where the computational complexity is evaluated with an input size of 256×256. The inference speed is measured on the NVIDIA V100 Tensor Core GPU with an input size of 768×512.

t-SNE results in Fig. 4. As shown in Fig. 4(a), disposing of spatial decoupling hinders the extraction of degradation-related features (*e.g.*, rain streaks and snow), resulting in the indistinguishable t-SNE clusters in Fig. 4(b).

Effectiveness of Sigmoid activation. Considering that applying Sigmoid activation for attention maps is a standard design in spatial attention [47, 60], we additionally apply Sigmoid activation on the spatial attention map (denoted as w/ Sigmoid) to assess its effectiveness. As illustrated by the purple curve in Fig. 9, applying the Sigmoid activation damages the performance compared with the original implementation (depicted by the green curve).

Discussions of attention variants. We attribute the superior RD performance to the hybrid attention mechanism, which integrates C-GA and S-DA to effectively model global dependencies and capture discriminative degradation representations. We further compare it with two popular attention mechanisms: 1) multi-head depth-wise transform attention (MDTA) [65], and 2) swin-transformer attention (SWTA) [44]. We conduct comparisons by replacing the HATB with the aforementioned attention blocks. The comparisons of RD performance and model complexity are provided in Fig. 10 and Tab. 3, respectively.

As illustrated in Fig. 10, with similar model complexity, our HATB-based model (green curve) demonstrates superior RD performance than the MDTA-based model (orange curve). Compared to the SWTA-based model (blue curve), our HATB-based model provides significantly better performance with only 68.21% of the parameters and 77.08% of the FLOPs, achieving a speedup of $1.41\times$. This underlines the potential of our HATB to serve as a versatile block to boost existing frameworks.

5. Conclusion

We propose a unified framework for all-in-one image compression and restoration, which equips neural image codec with the restoration ability against various degradations with the same set of trained weights. We leverage a hybrid attention mechanism to effectively distinguish genuine image information from degradations, and differentiate different types of degradations. Extensive experiments are conducted to demonstrate the superior RD performance of our method in handling degraded inputs without sacrificing the performance on clean data. The ablation studies further verify the rationality and effectiveness of our design.

References

[1] Eirikur Agustsson, David Minnen, George Toderici, and Fabian Mentzer. Multi-realism image compression with a conditional generator. In *CVPR*, pages 22324–22333, 2023. 2

[2] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *ICCV*, pages 221–231, 2019. 2

[3] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *ICLR*, 2017. 2

[4] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *ICLR*, 2018. 2

[5] Fabrice Bellard. Bpg image format. *URL https://bellard. org/bpg*, 1(2):1, 2015. 1

[6] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *TCSVT*, 31(10):3736–3764, 2021. 1

[7] Benoit Brummer and Christophe De Vleeschouwer. On the importance of denoising when learning to compress images. In *WACV*, pages 2440–2448, 2023. 2

[8] Shilv Cai, Liqun Chen, Sheng Zhong, Luxin Yan, Jiahuan Zhou, and Xu Zou. Make lossy compression meaningful for low-light images. In *AAAI*, volume 38, pages 8236–8245, 2024. 3

[9] Shilv Cai, Xu Zou, Liqun Chen, Luxin Yan, and Sheng Zhong. Jointly optimizing image compression with low-light image enhancement. *arXiv preprint arXiv:2305.15030*, 2023. 2

[10] Yuxuan Cai, Hongjia Li, Geng Yuan, Wei Niu, Yanyu Li, Xulong Tang, Bin Ren, and Yanzhi Wang. Yolobile: Real-time object detection on mobile devices via compression-compilation co-design. In *AAAI*, volume 35, pages 955–963, 2021. 1

[11] Wei-Ting Chen, Hao-Yu Fang, Cheng-Lin Hsieh, Cheng-Che Tsai, I Chen, Jian-Jiun Ding, Sy-Yen Kuo, et al. All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In *ICCV*, pages 4196–4205, 2021. 5, 6

[12] Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen Kuo. Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In *CVPR*, pages 17653–17662, 2022. 3

[13] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *CVPR*, pages 5896–5905, 2023. 2

[14] Yi-Hsin Chen, Kuan-Wei Ho, Shiau-Rung Tsai, Guan-Hsun Lin, Alessandro Gnutti, Wen-Hsiao Peng, and Riccardo Leonardi. Transformer-based learned image compression for joint decoding and denoising. *arXiv preprint arXiv:2402.12888*, 2024. 3

[15] Ka Leong Cheng, Yueqi Xie, and Qifeng Chen. Optimizing image compression via joint learning with denoising. In *ECCV*, pages 56–73, 2022. 1, 2

[16] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *CVPR*, pages 7939–7948, 2020. 2

[17] Benjamin Deguerre, Clément Chatelain, and Gilles Gasso. Fast object detection in compressed jpeg images. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 333–338, 2019. 1

[18] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *CVPR*, pages 12124–12134, June 2022. 5

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[20] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *CVPR*, pages 3855–3863, 2017. 5, 6

[21] Mario Gonzalez, Javier Preciozzi, Pablo Muse, and Andres Almansa. Joint denoising and decompression using cnn regularization. In *CVPRW*, pages 2598–2601, 2018. 2

[22] Zongyu Guo, Zhizheng Zhang, Runsen Feng, and Zhibo Chen. Causal contextual prediction for learned image compression. *TCSVT*, 32(4):2329–2341, 2021. 2

[23] Wang Guo-Hua, Jiahao Li, Bin Li, and Yan Lu. Evc: Towards real-time neural image compression with mask decay. In *ICLR*, 2023. 1, 2, 6

[24] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *CVPR*, pages 5718–5727, 2022. 1

[25] Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *CVPR*, pages 14771–14780, 2021. 2

[26] Qibin Hou, Cheng-Ze Lu, Ming-Ming Cheng, and Jiashi Feng. Conv2former: A simple transformer-style convnet for visual recognition. *arXiv preprint arXiv:2211.11943*, 2022. 5

[27] Yuning Huang, Zhihao Duan, and Fengqing Zhu. Narv: An efficient noise-adaptive resnet vae for joint image compression and denoising. In *ICMEW*, pages 188–193. IEEE, 2023. 3

[28] Seungmin Jeon, Kwang Pyo Choi, Youngo Park, and Chang-Su Kim. Context-based trit-plane coding for progressive image compression. In *CVPR*, pages 14348–14357, 2023. 1

[29] Eastman Kodak. Kodak lossless true color image suite (photocd pcd0992). *URL http://r0k. us/graphics/kodak*, 6, 1993. 5, 6, 7

[30] A Burakhan Koyuncu, Han Gao, Atanas Boev, Georgii Gaikov, Elena Alshina, and Eckehard Steinbach. Contextformer: A transformer with spatio-channel attention for context modeling in learned image compression. In *ECCV*, pages 447–463, 2022. 2, 3

[31] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3):18, 2017. 5

[32] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *CVPR*, pages 17452–17462, 2022. 2, 3, 5, 6

[33] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *TIP*, 28(1):492–505, 2018. 5, 6, 7

[34] Jiacheng Li, Chang Chen, Zhen Cheng, and Zhiwei Xiong. Toward dnn of luts: Learning efficient image restoration with multiple look-up tables. *TPAMI*, 2024. 2

[35] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *MM*, pages 1503–1511, 2022. 4

[36] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with feature modulation. In *CVPR*, pages 26099–26108, 2024. 1, 2

[37] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *CVPR*, pages 3175–3185, 2020. 2

[38] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *CVPR*, pages 18278–18289, 2023. 2

[39] Yinglong Li, Jiacheng Li, and Zhiwei Xiong. Look-up table compression for efficient image restoration. In *CVPR*, pages 26016–26025, June 2024. 2

[40] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, pages 1833–1844, 2021. 2, 5

[41] Huan Liu, George Zhang, Jun Chen, and Ashish J Khisti. Lossy compression with distribution shift as entropy constrained optimal transport. In *ICLR*, 2021. 2

[42] Jinming Liu, Heming Sun, and Jiro Katto. Learned image compression with mixed transformer-cnn architectures. In *CVPR*, pages 14388–14397, 2023. 1, 2, 3

[43] Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *TIP*, 27(6):3064–3073, 2018. 5

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 5, 8

[45] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *NeurIPS*, 31, 2018. 2

[46] David Minnen and Saurabh Singh. Channel-wise autoregressive entropy models for learned image compression. In *ICIP*, pages 3339–3343. IEEE, 2020. 2

[47] Diganta Misra, Trikay Nalamada, Ajay Uppili Arasanipalai, and Qibin Hou. Rotate to attend: Convolutional triplet attention module. In *WACV*, pages 3139–3148, January 2021. 8

[48] Dongwon Park, Byung Hyun Lee, and Se Young Chun. All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations. In *CVPR*, pages 5815–5824. IEEE, 2023. 2, 3

[49] Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one blind image restoration. *arXiv preprint arXiv:2306.13090*, 2023. 3

[50] Javier Preciozzi, Mario González, Andrés Almansa, and Pablo Musé. Joint denoising and decompression: A patch-based bayesian approach. In *ICIP*, pages 1252–1256, 2017. 2

[51] Yichen Qian, Xiuyu Sun, Ming Lin, Zhiyu Tan, and Rong Jin. Entroformer: A transformer-based entropy model for learned image compression. In *ICLR*, 2022. 2, 3

[52] Xuebin Sun, Sukai Wang, Miaohui Wang, Shing Shin Cheng, and Ming Liu. An advanced lidar point cloud sequence coding scheme for autonomous driving. In *MM*, pages 2793–2801, 2020. 1

[53] David S Taubman, Michael W Marcellin, and Majid Rabbani. Jpeg2000: Image compression fundamentals, standards and practice. *Journal of Electronic Imaging*, 11(2):286–287, 2002. 1

[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357, 2021. 5

[55] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *CVPR*, pages 2353–2363, 2022. 2

[56] Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991. 1

[57] Shiyao Wang, Hongchao Lu, and Zhidong Deng. Fast object detection in compressed video. In *ICCV*, pages 7104–7113, 2019. 1

[58] Yiting Wang, Pak Hung Chan, and Valentina Donzella. Semantic-aware video compression for automotive cameras. *IEEE Transactions on Intelligent Vehicles*, 8(6):3712–3722, 2023. 1

[59] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17683–17693, 2022. 2

[60] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 8

[61] Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. Robust web image/video super-resolution. *TIP*, 19(8):2017–2028, 2010. 2

[62] Zizheng Yang, Jie Huang, Jiahao Chang, Man Zhou, Hu Yu, Jinghao Zhang, and Feng Zhao. Visual recognition-driven image restoration for multiple degradation with intrinsic semantics recovery. In *CVPR*, pages 14059–14070, 2023. 1

[63] Mingde Yao, Ruikang Xu, Yuanshen Guan, Jie Huang, and Zhiwei Xiong. Neural degradation representation learning for all-in-one image restoration. *TIP*, 2024. 2, 3

[64] Juncheol Ye, Hyunho Yeo, Jinwoo Park, and Dongsu Han. Accelir: Task-aware image compression for accelerating neural restoration. In *CVPR*, pages 18216–18226, 2023. 2

[65] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022. 2, 4, 5, 6, 8

[66] Xinyi Zhang, Hang Dong, Jinshan Pan, Chao Zhu, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Fei Wang. Learning to restore hazy video: A new real-world dataset and a new method. In *CVPR*, pages 9239–9248, 2021. 5

[67] Yurui Zhu, Tianyu Wang, Xueyang Fu, Xuanyu Yang, Xin Guo, Jifeng Dai, Yu Qiao, and Xiaowei Hu. Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions. In *CVPR*, pages 21747–21758, 2023. 3, 5

[68] Yinhao Zhu, Yang Yang, and Taco Cohen. Transformer-based transform coding. In *ICLR*, 2021. 2, 3

[69] Renjie Zou, Chunfeng Song, and Zhaoxiang Zhang. The devil is in the details: Window-based attention for image compression. In *CVPR*, pages 17492–17501, 2022. 2, 3