Multimedia Systems (2024) 30:75
https://doi.org/10.1007/s00530-024-01272-5

REGULAR PAPER

Indirect: invertible and discrete noisy image rescaling with enhancement from case-dependent textures

Huu-Phu Do[1] · Yan-An Chen[1] · Nhat-Tuong Do-Tran[1] · Kai-Lung Hua[2] · Wen-Hsiao Peng[1] · Ching-Chun Huang[1]

Received: 11 April 2023 / Accepted: 18 January 2024 / Published online: 7 March 2024
© The Author(s) 2024

Abstract
Rescaling digital images for display on various devices, while simultaneously removing noise, has increasingly become a focus of attention. However, limited research has been done on a unified framework that can efficiently perform both tasks. In response, we propose INDIRECT (INvertible and Discrete noisy Image Rescaling with Enhancement from Case-dependent Textures), a novel method designed to address image denoising and rescaling jointly. INDIRECT leverages a jointly optimized framework to produce clean and visually appealing images using a lightweight model. It employs a discrete invertible network, DDR-Net, to perform rescaling and denoising through its reversible operations, efficiently mitigating the quantization errors typically encountered during downscaling. Subsequently, the Case-dependent Texture Module (CTM) is introduced to estimate missing high-frequency information, thereby recovering a clean and high-resolution image. Experimental results demonstrate that our method achieves competitive performance across three tasks: noisy image rescaling, image rescaling, and denoising, all while maintaining a relatively small model size.

Keywords Noisy image rescaling · Discrete invertible neural network · Case-dependent texture

1 Introduction

Over the past few decades, portable devices have become ubiquitous due to their convenience and ease of use. These devices are often equipped with advanced image sensors, simplifying the capture and sharing of high-resolution (HR) images. Given the varying resolutions of these devices and the need to reduce storage capacity and transmission bandwidth, there is a growing demand for efficient image-rescaling technology [1–5]. Image rescaling typically involves downscaling an HR image to a low-resolution (LR) format for storage/transmission, and subsequently upscaling this LR image back to the HR domain. Notably, images captured by portable devices tend to contain noise, attributable primarily to the compact design of their sensors. Therefore, the challenge of rescaling images while simultaneously removing noise to reconstruct visually pleasing and noiseless HR images, a process known as noisy image rescaling, is an emerging and significant topic in this field.

Though image denoising and image rescaling have been studied extensively as separate tasks in the literature [1, 2, 6–13], there is little research on noisy image rescaling, which combines denoising and rescaling in a single task. One straightforward solution to noisy image rescaling is to

Communicated by C. Yan.

✉ Ching-Chun Huang
chingchun@nycu.edu.tw

Huu-Phu Do
dohuuphu25.ee11@nycu.edu.tw

Yan-An Chen
my91015.cs08g@nctu.edu.tw

Nhat-Tuong Do-Tran
tuongdotn.cs11@nycu.edu.tw

Kai-Lung Hua
hua@mail.ntust.edu.tw

Wen-Hsiao Peng
wpeng@cs.nctu.edu.tw

[1] Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu City 300093, Taiwan

[2] Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei City 106335, Taiwan

perform denoising and rescaling in sequential steps as two separate processes. As far as image rescaling is concerned, the existing works can be divided into two main categories. One category views the downscaling and upscaling as two independent operations, while the other category regards them as a joint task. The former, which is also referred to as super-resolution (SR) [14–17] in the literature, can perform upscaling whether the downscaling kernel (e.g., bicubic) is known or not. However, focusing only on image upscaling, SR is an ill-posed problem. The recent work IRN [5], which belongs to the second category, alleviates the problem by simultaneously modeling downscaling and upscaling through an invertible network. It decomposes an HR image into a visually-pleasing LR image and high-frequency information (HFInfo) via a series of coupling layers [18–20]. Particularly, HFInfo is assumed to follow a case-independent (or input-independent) Gaussian distribution. As such, the HR image can be reconstructed by feeding the LR image and a random Gaussian sample to the coupling layers configured in a reverse way. Although IRN shows very promising rescaling results, it still has room for improvement. First, HFInfo is in fact case-specific (or input-dependent). The case-independent Gaussian sample can hardly be a good estimate of the missing high-frequency information. Second, the LR image must be quantized for transmission, compression, or display, which may cause quality degradation in the rescaling process. This issue is not fully addressed in IRN.

As regards image denoising, deep learning-based methods [6, 7] have made breakthrough progress on datasets with synthetic noise over the conventional image denoising methods [21, 22]. Meanwhile, some researchers started to address the real-world image denoising task, for which the input data are more complicated and challenging than datasets with synthetic noise for denoising. However, most of their designs adopt large models that are impractical for portable devices. Recently, InvDN [8] proposed an invertible image denoising network that effectively reduces the model size while achieving good denoising performance. Built upon IRN [5], InvDN [8] first disentangles the noisy input image into two parts: the clean LR image and the HFInfo component modeling the noise embedded in the input image. Denoising is subsequently achieved by discarding the HFInfo in the reverse operation of the network. Although InvDN [8] is designed initially for image denoising, it has the potential for addressing the noisy image rescaling task due to its intrinsic IRN-like [5] architecture. However, like IRN [5], it suffers from the same quantization issue of the LR image when applied to the noisy image rescaling task.

To this end, we propose a novel framework, INvertible and Discrete Noisy Image Rescaling with Enhancement from Case-dependent Textures (INDIRECT), to tackle noisy image rescaling in a joint optimization manner. Inspired by IRN [5] and InvDN [8], INDIRECT is built upon invertible networks and involves two sub-modules: (a) the discrete invertible denoising and rescaling networks (DDR-Net) and (b) the case-dependent texture module (CTM). Similar to IRN [5], DDR-Net first disentangles the noisy HR image into an informative LR image (embedding possibly some useful high-frequency information) and the noisy HFInfo part. DDR-Net is a discretized and invertible network with the aim of mitigating the quantization effects of the LR image during downscaling. Next, the noisy HFInfo is discarded for denoising during downscaling. For upscaling, instead of simply drawing a case-independent random Gaussian sample for the missing HFInfo, our CTM effectively exploits the informative LR image to predict a more representative and clean HFInfo. This design is justified by the fact that the coupling architecture of DDR-Net allows partial HFInfo to be embedded into the LR image. Therefore, CTM can estimate a case-specific HFInfo based on the LR image for better reconstruction quality. Besides, we also present a perceptual version of INDIRECT to enhance the perceptual quality of the reconstructed image. To sum up, our contributions are as follows:

- To the best of our knowledge, our work is the first attempt to tackle the noisy image rescaling task in a unified framework that simultaneously addresses and optimizes image denoising and rescaling.
- Our CTM module effectively leverages the embedded information in the downscaled LR image to generate the case-specific HFInfo and reconstruct a better clean HR image.
- We present DDR-Net, a novel and efficient way to tackle the quantization errors resulting from the downscaling process by modeling the invertible network in a discrete manner.

Furthermore, we note that the proposed method (i.e., INDIRECT) is an extension of our previous conference work (i.e., DIRECT) [23], which focuses only on the image rescaling task. The differences between INDIRECT and DIRECT are threefold. First, INDIRECT is a unified framework that simultaneously addresses image denoising and rescaling; in contrast, DIRECT focuses only on image rescaling. The difference in problem setting makes them fundamentally distinct. Second, we provide more comprehensive experiments and comparisons for image denoising, image rescaling, and noisy image rescaling, which are not covered in [23]. Third, we detail the implementation, datasets, and ablation study in depth, which are missing in [23].

2 Related works

Due to the advances in portable devices, a tremendous amount of HR images is captured every day. It is challenging to reduce the storage requirements or transmission bandwidth of HR images while having to remove the noise introduced by the device's image sensor. The following summarizes the related works on image rescaling and image denoising.

2.1 Image rescaling

Image rescaling can be divided into two sequential steps: image downscaling and image upscaling. Conventionally, image super-resolution is applied to perform the image rescaling task. Depending on whether the downscaling kernel is known or not, it can upscale the input LR image in a non-blind or blind manner. However, image super-resolution tends to regard upscaling and downscaling as uncorrelated processes and is by nature an ill-posed problem. Some recent works jointly optimize the downscaling and upscaling processes as a new task called image rescaling [1–4], by leveraging an encoder-decoder framework. Specifically, the encoder learns the downscaling process in the hopes of acquiring a more informative and visually pleasing LR image; meanwhile, a better HR image is recovered through the jointly optimized decoder. However, this framework implicitly removes and recovers the HFInfo in the downscaling and upscaling process instead of explicitly modeling the HFInfo.

Recently, IRN [5] proposed a novel invertible network that views upscaling and downscaling as reverse operations of each other. This network is bijective in the sense that it can carry out upscaling and downscaling in an invertible way. Architecture-wise, it has two sub-modules: 2D Haar wavelet transform and an invertible neural network [18–20]. First, 2D Haar wavelet transform is applied to decompose the input HR image x_{HR} into a low-frequency component x_{LHaar} and a high-frequency one x_{HHaar}. Next, x_{LHaar} and x_{HHaar} are processed via the invertible neural network. Through the invertible network, x_{LHaar} will be converted into a visually pleasing clipping LR image $\tilde{x}_{LR}$, and x_{HHaar} into x_{HF}, which is trained to follow an isotropic Gaussian distribution. Since x_{HF} is unknown at inference time, IRN [5] samples from the Gaussian distribution to get z for upscaling. Finally, the HR image is recovered by inversely passing z and the LR image $\tilde{x}_{LR}$ through the network.

It is to be noted that the HFInfo sampled from an isotropic Gaussian is case-agnostic and may not be able to represent diverse scenes dynamically and precisely. Moreover, without considering the quantization effect of the LR image, IRN [5] may suffer from poor HR reconstruction results. In contrast, our proposed method exploits the embedded information in the clipping LR image $\tilde{x}_{LR}$ to predict a case-dependent HFInfo ($\tilde{x}_{HF}$), and alleviates the quantization effect at the same time by means of a discretized invertible network. We will detail these insights of our method in the Sect. 3.

2.2 Image denoising

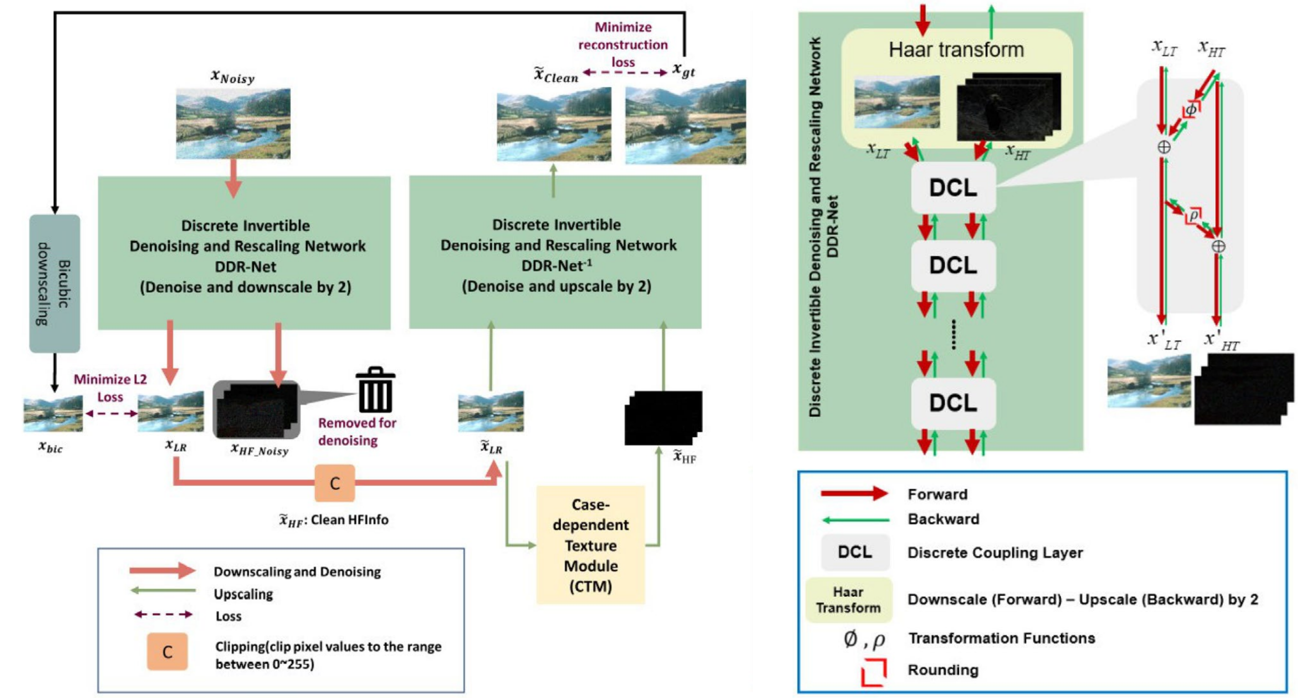
In recent years, deep learning-based denoising methods have achieved great success on datasets with synthetic noise, as compared with conventional methods such as NLM [21] and BM3D [22]. There are also works addressing the real-world image denoising task, which is even more challenging than denoising on synthetic noise data. DANet [6] proposes a unified framework to achieve noise removal and noise generation simultaneously. Its GAN-based architecture contains two parallel branches: one for noise removal and the other for noise generation. In this way, the denoising process can be performed by fooling the discriminator with three kinds of paired data, which are (clean, noisy), (clean, generated noisy), and (denoised, noisy). Besides, MIRNet [7] introduces a multi-scale residual architecture to deal with image restoration. It utilizes the multi-scale feature aggregation along with the spatial and channel attention mechanism to fully exploit the information exchanged between multi-scale features, achieving significant performance improvement. However, both DANet [6] and MIRNet [7] adopt a strategy of using larger and deeper models such as U-Net [24] structure. The heavy computational costs lead to limited practicality in real-world applications.

In contrast, InvDN [8] applies the invertible network in resolving the image denoising problem. Its lightweight architecture and invertibility make InvDN [8] suitable for portable devices like smartphones. InvDN [8] first disentangles the noisy image x into a clean low-resolution image x_{LR} and a noisy high-frequency component x_{HF} by the forward path of the invertible network. It then substitutes a simple Gaussian sample z_{HF} for x_{HF} in reconstructing a clean image. Apparently, InvDN [8] does not leverage the possible case-specific HF information from the x_{LR}; it simply recovers the missing HFInfo from a Gaussian distribution.

3 Proposed method

3.1 System overview

As shown in Fig. 1a, the proposed INDIRECT performs dual functions in a unified and invertible way to handle image denoise and image rescaling tasks simultaneously. The dual functions include denoising x_{Noisy} and downscaling it first and then enhancing as well as upscaling its LR image back to an HR one. Specifically, the downscaling and denoising processes are coupled in the forward path while the upscaling is conducted in the backward direction of the network. Besides, to meet the needs of video transmission, the goal of


Minimize reconstruction loss
x_{Noisy}
$\tilde{x}_{Clean}$
x_{gt}
Bicubic downscaling
Discrete Invertible Denoising and Rescaling Network DDR-Net (Denoise and downscale by 2)
Discrete Invertible Denoising and Rescaling Network DDR-Net^{-1} (Denoise and upscale by 2)
Minimize L2 Loss
x_{bic}
x_{LR}
x_{HF_Noisy}
Removed for denoising
$\tilde{x}_{LR}$
$\tilde{x}_{HF}$
C
$\tilde{x}_{HF}$: Clean HFInfo
Case-dependent Texture Module (CTM)
Downscaling and Denoising
Upscaling
Loss
C Clipping(clip pixel values to the range between 0~255)
Discrete Invertible Denoising and Rescaling Network DDR-Net
Haar transform
x_{LT}
x_{HT}
DCL
DCL
DCL
ϕ
ρ
x'_{LT}
x'_{HT}
Forward
Backward
DCL Discrete Coupling Layer
Haar Transform Downscale (Forward) – Upscale (Backward) by 2
$\emptyset$, ρ Transformation Functions
Rounding

(a) The overall framework of INDIRECT, which applies DDR-Net to mitigate the quantization effect and leverages the information embedded in $\tilde{x}_{LR}$ to predict the clean missing high-frequency information $\tilde{x}_{HF}$ by CTM. By cascading more DDR-Net modules, the system can rescale the input image by a larger factor.

(b) DDR-Net is composed of several downscaling blocks. Each downscaling block comprises one Discrete Haar Transform that decomposes HR image into low-frequency and high-frequency bands and multiple discrete coupling layers.

Fig. 1 **a** The illustration of the proposed INDIRECT framework and **b** our DDR-Net structure

the noisy image rescaling task is to generate not only a clean HR image but also an informative and visually pleasing LR image in a joint optimization manner.

To this end, we introduce two modules, Discrete Invertible Denoising and Rescaling Network (DDR-Net) and Case-dependent Texture Module (CTM), to perform the noisy image rescaling task and address the challenges mentioned above. As depicted in Fig. 1a, DDR-Net and CTM work in a unified way to achieve both image rescaling and denoising. Particularly, DDR-Net is applied to remove the noise embedded in the input image and alleviate the quantization effect simultaneously, while CTM is used to effectively explore the embedded information from $\tilde{x}_{LR}$ to recover the clean HFInfo. In more detail, a noisy HR image x_{Noisy} first goes through DDR-Net forwardly to acquire its clean LR image x_{LR} and noisy HFInfo x_{HF_Noisy}. Second, x_{LR} is clipped to $\tilde{x}_{LR}$ according to the specified bit-depth and x_{HF_Noisy} is discarded to remove the noise. Finally, to reconstruct a clean HR image from $\tilde{x}_{LR}$, CTM is applied to model the missing HFInfo based on the contents of $\tilde{x}_{LR}$, with DDR-Net operated subsequently in reverse mode to achieve the upscaling process. The details of each module are elaborated in the following sections.

3.2 DDR-Net: discrete invertible denoising and rescaling network

Design and formulation: Though IRN [5] can perform the rescaling task without ill-posed issues, the downscaled image has continuous intensities, which are not suitable for digital compression and transmission. One solution is to apply intensity quantization and clipping operations on the LR image, but it may create artifacts in the reconstructed HR image. Inspired by IDF [25], we introduce a discrete invertible rescaling network (DDR-Net) as shown in Fig. 1b to tackle the quantization effect by leveraging discrete flows. The forward path of DDR-Net involves stacking multiple downscaling blocks to disentangle the noisy HR image $x_{Noisy} \in \mathbf{R}^{C \times H \times W}$ into an informative LR, denoted as x_{LR}, and its corresponding HFInfo with noise, x_{HF_Noisy}, by exploiting the information-lossless properties of the invertible network.

In the forward path of DDR-Net, each downscaling block is applied to rescale the input data horizontally and vertically by a factor of 2. This is achieved by two modules: discrete Haar transform and discrete coupling layers. Starting with the first downscaling block, we apply a discrete Haar transform to decompose the noisy input x_{Noisy} into a low-frequency component $x_{LT} \in \mathbf{R}^{C\times\frac{H}{2}\times\frac{W}{2}}$ - capturing information from the LL subband, and a high-frequency component $x_{HT} \in \mathbf{R}^{3C\times\frac{H}{2}\times\frac{W}{2}}$ - capturing information from the LH, HL, and HH sub-bands. Since x_{Noisy} is a digitized image, discrete Haar transform is applied to acquire its x_{LT} and x_{HT} discretely to alleviate rounding and quantization errors. By the same token, we replace the conventional coupling layers in IRN [5] with discrete coupling layers.

As illustrated in Fig. 1b within each discrete coupling layer, we apply additive transformations as follows to couple information between x_{LT} and x_{HT}.

$$x'_{LT} = x_{LT} + \lfloor \phi(x_{HT}) \rceil, \tag{1}$$

$$x'_{HT} = x_{HT} + \lfloor \rho(x'_{LT}) \rceil, \tag{2}$$

where $\lfloor\rceil$ is the rounding operation and ϕ, ρ are transformation functions implemented by neural networks. In this formulation, ϕ, ρ can be arbitrary deep-learning-based transformations without affecting the invertibility of the discrete coupling layer. In other words, the invertibility is guaranteed through

$$x_{HT} = x'_{HT} - \lfloor \rho(x'_{LT}) \rceil, \tag{3}$$

$$x_{LT} = x'_{LT} - \lfloor \phi(x_{HT}) \rceil. \tag{4}$$

Utilizing the aforementioned invertible formulation, we are able to implement the backward path of DDR-Net for upsampling by employing the same transformations ϕ and ρ. Ultimately, by applying the inverse discrete Haar transform, we successfully reconstruct the clean High-Resolution (HR) output.

Functionality of information embedding: Through cascading multiple discrete coupling layers and downscaling blocks, the noisy image x_{Noisy} is finally disentangled into x_{LR} and x_{HF_Noisy}. During the decoupling process, we realize two important functions for both denoise and downscaling. First, we leverage the information exchangeable properties of the invertible network to push the noise component of the input image to flow into x_{HF_Noisy} and require x_{LR} to be noise-free; second, as an important part of our strategy to achieve superior High-Resolution (HR) image reconstruction, we implicitly embed extra high-frequency components into the LR images x_{LR}. While the quality of the LR image, in terms of conventional quality metrics, may be degraded due to information embedding, they are instrumental in achieving our primary goal of enhanced HR image reconstruction by information embedding. In this way, denoising can be performed jointly with the downscaling process by discarding x_{HF_Noisy}.

The process of embedding high-frequency information into low-resolution (LR) images warrants further discussion. To illustrate this, we applied Fourier transformations to LR images obtained using different methods, thereby demonstrating how our method embeds high-frequency information (HFinfo) in LR images. As depicted in Fig. 2, the left side of Fig. 2a, b displays the LR images produced by downscaling a high-resolution (HR) image using the bicubic algorithm and the forward process of our proposed DDR-Net,

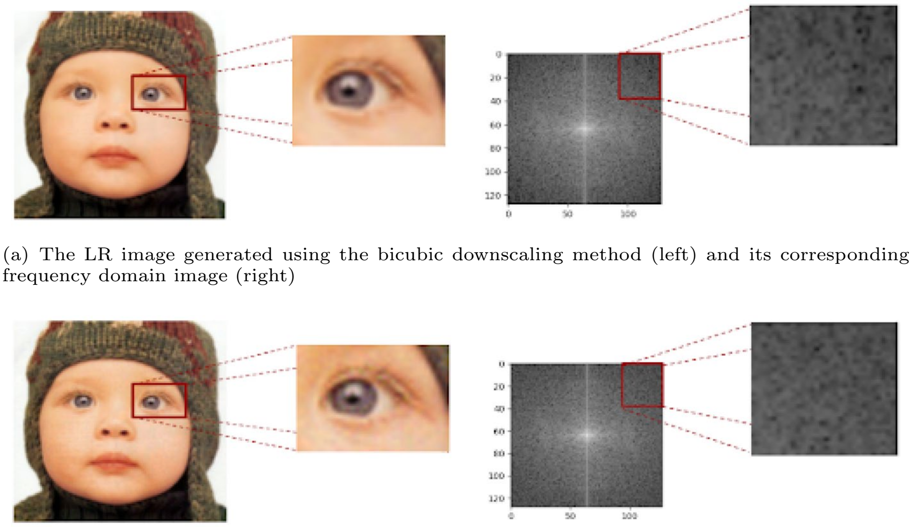


Fig. 2 Comparison of various methods for generating low-resolution (LR) images and their corresponding visualization in the frequency domain

(a) The LR image generated using the bicubic downscaling method (left) and its corresponding frequency domain image (right)

(b) The LR image generated using the our proposed method (left) and its corresponding frequency domain image (right) respectively. The right sides of these figures show their corresponding frequency responses.

A comparison of the high-frequency regions, particularly the top-right corners of the frequency responses, reveals a notable difference. Our LR image, unlike the bicubic-downscaled image, exhibits a richer presence of high-frequency information, indicated by fewer black regions. This distinction highlights that our LR image incorporates additional high-frequency details. This strategic inclusion of high-frequency information is crucial for enhancing the quality of the reconstructed HR image.

As a result of our DDR-Net's coupling structure, we have embedded the missing high-frequency contents into the LR image. This allows us to utilize the resulting $\tilde{x}_{LR}$ to accurately estimate clean High-Frequency Information (HFInfo) via our "Case-dependent" Texture Module (CTM) and significantly enhances the quality of the reconstructed High-Resolution (HR) image. The subsequent section will present a detailed description of the proposed CTM module.

3.3 CTM: case-dependent texture module

Insight explanation: Recent advancements in image reconstruction involve the integration of clean low-resolution (LR) images with a random Gaussian sample from a "unified" distribution, as demonstrated by methods like IRN [5]. This distribution is assumed to represent a portion of high-frequency information (HFinfo). However, this approach overlooks the fact that each image possesses unique HFinfo characteristics. Our innovative Case-dependent Texture Module (CTM) is designed to bridge this gap by accurately estimating the missing HFinfo, specifically tailored for each image. Unlike previous methods that employ a 'uniform' high-frequency (HF) distribution for all test images, the distinctiveness of our method lies in the 'variability' of the estimated HF distribution, which we use to generate the HFinfo for different input images.

Specifically, we employ CTM to model the variation of HF distribution for each input Low-Resolution image. Given that every image has distinct HF components and different LR structures facilitating information hiding, our approach proposes embedding case-specific High-Resolution information into $\tilde{x}_{LR}$ (i.e., defined in Sect. 3.2). The CTM module then estimates the corresponding HR distribution to facilitate the reconstruction of HFinfo. Consequently, the corresponding HR distribution is custom-tailored for each specific case, closely aligning with the characteristics of the respective input image.

Network design: Utilizing the coupling structure of DDR-Net, we strategically embed a portion of the clean high-frequency information into the LR image, denoted as $\tilde{x}_{LR}$. Besides, given that $\tilde{x}_{LR}$ is constrained to be a clean image, the embedded information within it is ideally suited for reconstructing the missing noise-free high-frequency part. This insight leads us to model the conditional distribution of clean High-Frequency Information (HFInfo) $\tilde{x}_{HF}$ given $\tilde{x}_{LR}$, rather than merely sampling HFInfo from a case-agnostic isotropic Gaussian distribution like IRN [5] or InvDN [8]. To effectively model the conditional distribution $p(\tilde{x}_{HF}|\tilde{x}_{LR})$, we have developed CTM to tailor to estimate Case-dependent high-frequency information (HFinfo) derived from the input LR image, $\tilde{x}_{LR}$.

The Case-dependent Texture Module (CTM) is structured as an encoder-decoder framework. This design enables the module to learn a probabilistic mapping from input data to a latent space via its encoder. Subsequently, by sampling points within this latent space, the CTM decoder facilitates the estimation of High-Frequency Information (HFInfo), denoted as ($\tilde{x}_{HF}$). To accurately model the LR-to-HR one-to-many mapping, the CTM's encoder, which comprises residual blocks, is tasked with modeling the conditional Gaussian distribution. This distribution characterizes the high-frequency latent code $z|\tilde{x}_{LR}$ based on the LR image $\tilde{x}_{LR}$.

As illustrated in Fig. 3, the conditional distribution of z is designed to vary with $\tilde{x}_{LR}$, allowing for the drawing of a case-specific high-frequency latent sample z from its content-dependent Gaussian distribution. The decoder of CTM, which is trained to transform this Gaussian distribution into one that more accurately characterizes HFInfo, utilizes the latent code z to estimate the missing high-frequency components suitable for each specific input, thereby yielding precise HFInfo.

Furthermore, since each sample z represents a specific realization of HFInfo, the CTM also models the uncertainty inherent in the one-to-many mapping from LR to HR. The estimated case-specific HFInfo $\tilde{x}_{HF}$, representing the missing clean HFInfo, is then combined with the LR image $\tilde{x}_{LR}$ to recover the clean HR image $\tilde{x}_{Clean}$ via the backward (i.e., inverse) operation of DDR-Net. This process effectively enhances the reconstruction quality of the HR output.

3.4 Loss functions

In designing our loss function, we first require that x_{LR} should be informative and noiseless. We follow the idea of [5, 8] to guide its visual quality by a bicubically downsampled clean image. Specifically, let x_{bic} be the LR image downscaled with a bicubic kernel from the ground truth x_{gt}, which is a clean HR image. The visual loss is defined as

$$L_{vis} = \frac{1}{N}\sum_{i=1}^{N} \|x_{LR}^{i} - x_{bic}^{i}\|^2. \tag{5}$$

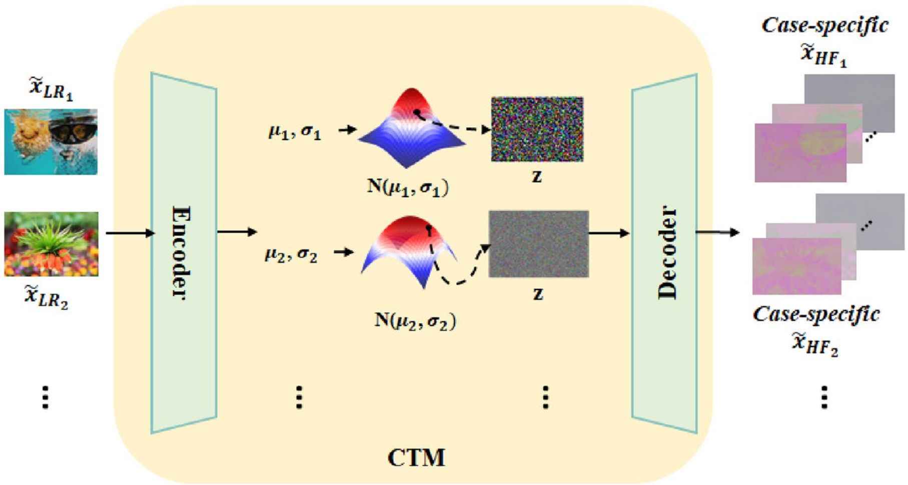


Fig. 3 The architecture of the case-dependent texture module. CTM takes the clipped LR image from DDR-Net as its input and generates its corresponding missing textures

Next, we utilize the Charbonnier loss [14] to minimize the distortion between the clean HR image x_{gt} and the reconstructed image $\tilde{x}_{Clean}$:

$$L_{rec} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\|\tilde{x}^i_{Clean} - x^i_{gt}\|^2 + \varepsilon^2}. \tag{6}$$

Lastly, we optimize our INDIRECT model by minimizing a weighted sum of the visual loss L_{vis} and the reconstruction loss L_{rec}:

$$L_{INDIRECT} = \alpha L_{vis} + \beta L_{rec}, \tag{7}$$

where α and β are hyper-parameters used to balance the quality between the LR and HR images.

In the case of optimizing the perceptual quality of $\tilde{x}_{Clean}$, we introduce a perceptual loss [26] L_{per} to encourage natural and perceptually-pleasing results by enhancing the feature similarities between the clean HR image x_{gt} and the reconstructed image $\tilde{x}_{Clean}$. This constraint is imposed in the feature space, with the semantic features extracted by a pre-trained model. The perceptual loss is given by

$$L_{per} = \frac{1}{N} \sum_{i=1}^{N} \|\psi_j(\tilde{x}^i_{Clean}) - \psi_j(x^i_{gt})\|^2, \tag{8}$$

where N is the batch size, ψ is the pre-trained feature extractor model, and j is the index of layers used to evaluate the perceptual loss. Based on the perceptual loss, we train another model, called INDIRECTp, based on the following loss:

$$L_{INDIRECTp} = \alpha L_{vis} + \beta L_{rec} + \gamma L_{per}. \tag{9}$$

3.5 Comparison with IRN and InvDN

Similar to IRN and InvDN, INDIRECT is built upon the invertible neural networks. Although they are all flow-based models that can perform the perfect invertible bijective transformation, they differ in the applications and their intrinsic structures. For applications, rather than addressing a single task like IRN and InvDN, INDIRECT deals with a more challenging hybrid task: denoise and rescale an image in a unified way. Consequently, its forward path, disentangling the LT/HT components while separating noises from the input image, works bifunctionally. Regarding the intrinsic structure, we introduce the discrete flow scheme for image compression and transmission. In particular, we redesign the transformation functions as discrete ones to alleviate the quantization effect. Moreover, the prediction of the missing HFInfo is also significantly different from IRN and InvDN. Instead of randomly sampling from a case-agnostic distribution, we propose novel CTM to generate case-specific and more precise high-frequency components to recover an image with finer details. These non-trivial inventions are indispensable for the new and practical task, noisy image rescaling, that cannot be solved efficiently by the existing methods.

4 Experimental results

This section starts with an introduction to the datasets, evaluation metrics, and implementation details. Next, we report experimental results not only for (1) the noisy image rescaling task but also for (2) the standalone image rescaling and (3) the standalone image denoising tasks. Moreover, we present in-depth analyses of the effectiveness of CTM, how the quantization effect is mitigated and how visually pleasing LR images are achieved. Finally, we conclude this section with ablation studies to analyze the effect of each component in our scheme.

4.1 Datasets and evaluation metrics

Datasets: To evaluate our scheme on the three tasks and make a fair comparison with the competing methods, we follow the common test protocols for (2) the standalone image rescaling and (3) the standalone image denoising tasks, respectively. In particular, (1) the noisy image rescaling is a new task and will be analyzed based on the denoising framework. In detail, for the tasks of (1) noisy image rescaling and (3) image denoising, SIDD [27] dataset is used for training. This dataset is collected by five representative smartphone cameras rather than DSLR, making the denoising task more challenging. Specifically, at training time, we use the medium version of SIDD [27] dataset, which consists of 320 clean-noisy image pairs for training and 40 pairs for validation. As regards the testing phase, the evaluation is performed on SIDD validation set [27] and DND dataset [28]. DND [28] dataset is captured by four consumer cameras with different sensor sizes and consists of 50 real-world noisy-clean image pairs.

Concerning (2) the image rescaling task, we apply DIV2K datasets [29] for training, like the conventional methods. DIV2K has 1000 HR images with diverse contents and is composed of 800 pairs of training data, among which 100 pairs are for validation and another 100 pairs are for testing. The evaluation is then performed on widely used datasets: Set5 [30], Set14 [31], BSDS100 [32] which involves large natural scenes, Urban100 [33] which consists of urban scenes, Manga109 [34] which is composed of manga drawn by Japanese professional manga artists and DIV2K validation set [29].

Evaluation metrics: When conducting quantitative analyses, we utilize Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [35] as the main evaluation metrics for (1) noisy image rescaling, (2) the standalone image rescaling task, (3) image denoising. In addition, since most of the previous methods did not use human visual perception metrics for comparisons, we also adopt the perceptual quality metric: LPIPS [36] with VGG features to measure the quality of the reconstructed images. Natural Image Quality Evaluator (NIQE) [37] is also used for some tasks.

4.2 Implementation details

Our INDIRECT comprises two DDR-Net modules, each of which consists of eight discrete coupling layers. Adam parameter [38] is used for optimization with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is 2×10^{-4} and is reduced by half every 100k iterations. Pytorch is the major framework for our implementation and the training is performed on 2080-Ti GPU. For data augmentation, we apply horizontal and vertical flipping and 90° rotation.

4.3 Results on noisy image rescaling

We compare our method for (1) the noisy image rescaling task with two compound schemes: (A) image denoising + image super-resolution (SR) and (B) image denoising + image rescaling. For (A), we first denoise the noisy input image by state-of-the-art (SOTA) denoising methods [7, 8], downscale the denoised image by the bicubic kernel, and then upscale the LR image by SR methods [16]; for (B), the denoising is performed in a similar way to (A) while SOTA rescaling methods [1, 5] are applied instead. Besides, the individual modules in (A) and (B) are trained separately and then combined to carry out the hybrid task. In contrast, since InvDN [8] alone can implicitly perform noisy image rescaling, we re-train its network for a fair comparison and denote the re-trained version as InvDN-r.

As shown in Table 1, we observe that the compound scheme (B) generally achieves better results than (A) since the former has a more flexible downscaling design. Moreover, for the type that can jointly perform image rescaling

Table 1 The quantitative result of noisy image rescaling results (downscale and upscale by 4) in terms of PSNR(↑)/SSIM(↑)/LPIPS(↓)

Denoising	Downscaling	Upscaling	Scheme	Params	SIDD	DND
MIRNet [7]				47.18M	38.93/0.9092/0.340	38.98/0.9447/–
InvDN [8]	Bicubic	RCAN [16]	(A) Denoising + SR	18.24M	38.76/0.9071/0.351	38.59/0.9406/–
MIRNet [7]				84.58M	39.05/0.9126/0.330	*39.20/0.9473*/–
InvDN [8]	CAR [1]	EDSR [15]		55.44M	38.71/0.9095/0.347	38.80/0.9434/–
MIRNet [7]				36.13M	39.10/**0.9150**/*0.325*	**39.42**/*0.9510*/–
InvDN [8]	IRN [5]	IRN [5]	(B) Denoising + Rescaling	6.99M	38.93/0.9122/0.344	38.92/0.9467/–
InvDN-r(re-trained version of InvDN [8])				*2.64M*	*39.15/0.9129*/0.332	38.93/0.9467/–
INDIRECT			Noisy image rescaling	**2.56M**	**39.34**/*0.9147*/**0.323**	39.13/0.9483/–

Numbers in bold and italic indicate the best and the second-best performance, respectively. Note that LPIPS can not be evaluated on the DND dataset because the ground truth has not been published

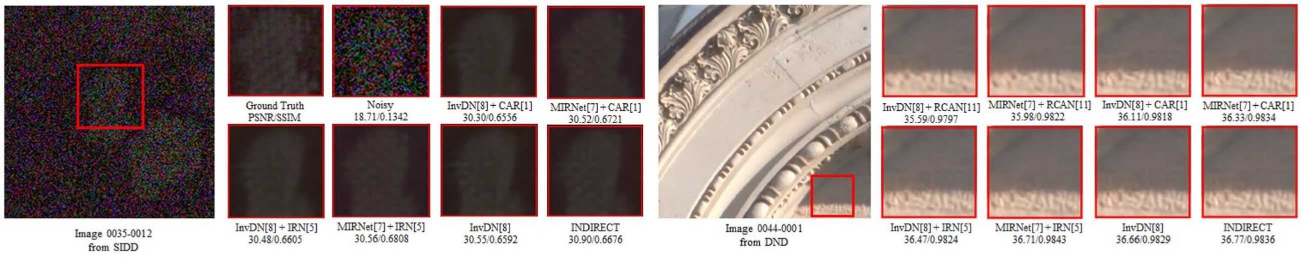


Fig. 4 Qualitative result comparison for noisy image rescaling (PSNR/SSIM provided)

and image denoising, the results show that our method outperforms InvDN [8] and achieves significant improvement. In addition, INDIRECT not only achieves better PSNR and SSIM results than most of the other methods on both SIDD and DND datasets but also yields superior LPIPS results with fewer model parameters. It is pertinent to note that the DND dataset does not release the ground truth data for its testing set. Therefore, we obtained PSNR and SSIM metrics by submitting our results to the DND benchmark website [28]. Given that the website does not provide LPIPS results, we have left the LPIPS results for the DND dataset blank.

As shown in Fig. 4, the compound scheme (A) suffers from the blurriest results while INDIRECT provides sharper details than InvDN [8], such as the textures at the bottom and the black line on the upper right side. Besides, INDIRECT can achieve comparable results to the compound scheme (B) with much fewer model parameters.

4.4 Results on image rescaling

Regarding (2) the standalone image rescaling task, we compare our method with two types of baselines: (A) compound methods that append the bicubic downscaling with SOTA image SR methods [15, 16, 39], and (B) image rescaling methods that include encoder-decoder frameworks [1, 2] and invertible frameworks [5, 40].

As shown in Table 2, the compound methods (A) achieve similar performance. In comparison, with learnable downscaling methods, most of the compound methods (B) benefit from the joint optimization of the downscaling and upscaling processes, showing much better performance than the methods in (A). Furthermore, invertible frameworks [5, 40] outperform the encoder-decoder methods [1, 2] by exploiting the invertible architecture, which mitigates the ill-posed problem. Particularly, we address the quantization issue in the rescaling task and exploit the informative LR image to estimate more precise HFInfo, making INDIRECT outperform all the competing methods.

As shown in Table 3, INDIRECTp is involved in this test to demonstrate the effectiveness of the perceptual loss, and it also outperform INDIRECT which already achieves better perceptual quality than the other SOTA image rescaling methods both on LPIPS and NIQE.

Figure 5 shows the qualitative comparison between SOTA image rescaling methods and ours. We see that INDIRECT produces more visually pleasing results, as shown in the red-cropped region. Especially, INDIRECT recovers more high-frequency details than IRN [5] and HCFlow [40], both estimating the missing HFInfo by drawing samples from a

Table 2 Comparison in terms of PSNR-Y(↑)/SSIM-Y(↑)(on the Y channel only) on different datasets for image rescaling results (downscale and upscale by 4)

Downscaling	Upscaling	Param	Set5	Set14	BSD100	Urban100	Manga109	DIV2K
	Bicubic	–	28.42/0.810	26.00/0.703	25.96/0.668	23.14/0.658	24.89/0.787	26.66/0.852
	EDSR [15]	43.1M	32.62/0.898	28.94/0.790	27.79/0.744	26.86/0.808	31.02/0.915	29.38/0.903
	RCAN [16]	15.6M	32.63/0.900	28.87/0.789	27.77/0.744	26.82/0.809	31.21/0.917	30.77/0.846
Bicubic	SAN [39]	15.7M	32.64/0.900	28.92/0.789	27.78/0.744	26.79/0.807	31.18/0.917	–
TAD [2]	TAU [2]	–	31.81/–	28.63/–	28.51/–	26.63/–	–	31.16/–
CAR [1]	EDSR [15]	52.8M	33.88/0.917	30.31/0.838	29.15/0.800	29.28/0.871	33.89/0.941	32.82/0.884
IRN [5] (invertible)		*4.4M*	36.19/*0.945*	32.67/0.902	31.64/0.883	31.41/0.916	*35.94/0.962*	35.07/*0.932*
HCFlow [40] (invertible)		*4.4M*	*36.29*/**0.947**	*33.02*/**0.907**	*31.74/0.886*	*31.62/0.921*	*–/–*	*35.23*/**0.935**
INDIRECT (invertible)		**3.7M**	**36.40**/*0.945*	**33.07**/*0.904*	**31.95/0.887**	**32.13/0.923**	**36.65/0.964**	**35.43/0.935**

Numbers in bold and italic indicate the best and the second-best performance, respectively. Evaluations that are not provided in the original paper are denoted as "–". For the coupling layers, we apply the dense block as IRN [5]

Table 3 Quantitative result of perceptual quality comparison with different state-of-the-art rescaling methods for image rescaling results (downscale and upscale by 4) in terms of LPIPS(↓)/NIQE(↓)

Downscaling	Upscaling	Param	Set5	Set14	BSD100	Urban100	Manga109	DIV2K
CAR [1]	EDSR [15]	52.8M	0.131/4.170	0.201/4.759	0.258/5.153	0.134/5.088	0.070/4.634	0.193/4.547
IRN [5] (invertible)		*4.4M*	0.077/3.865	0.123/4.280	0.166/4.433	0.084/4.470	0.043/4.354	0.119/3.935
HCFlow [40] (invertible)		*4.4M*	0.080/3.765	0.120/4.480	0.169/4.469	0.081/4.532	0.042/4.474	0.119/4.134
INDIRECT (invertible)		**3.7M**	*0.076/3.571*	*0.115/4.223*	*0.158/4.329*	*0.075/4.335*	*0.039/4.254*	*0.109/3.769*
INDIRECTp (invertible)		**3.7M**	**0.059/3.134**	**0.080/3.272**	**0.070/3.163**	**0.047/3.807**	**0.025/3.742**	**0.050/3.105**

Numbers in bold and italic indicate the best and the second-best performance, respectively

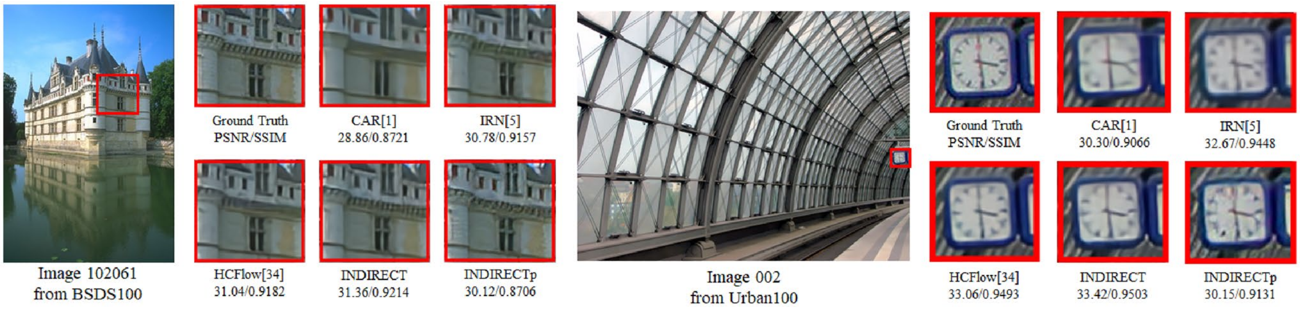


Fig. 5 Qualitative result comparison for image rescaling (PSNR-Y/SSIM-Y provided)

case-agnostic Gaussian distribution. Furthermore, with the perceptual loss, INDIRECTp generates much sharper and more realistic images.

4.5 Results on image denoising

In this subsection, we focus on the standalone image denoising task. Here, we apply the original continuous invertible networks since our DDR-Net is designed primarily to address the quantization effects in generating the LR image for rescaling.

In the standalone denoising task, the LR image serves as a latent signal not to be consumed by humans. Table 4 shows that as compared with InvDN [8], our proposed method achieves better PSNR/SSIM results on the SIDD dataset by effectively exploiting the embedded information in $\tilde{x}_{LR}$ to estimate the clean HFInfo in a case-specific fashion. Additionally, we conducted extra evaluations using the LPIPS metric to demonstrate the perceptual quality performance. Specifically, MIRNet [7], with a model size of 31.78M, secured the best performance with an LPIPS score of 0.307. Our INDIRECT method, significantly smaller at 3.44M, achieved the second-best result with an LPIPS score of 0.342. InvDN [8] and DANet [6] also demonstrated competitive performances with scores of 0.346 and 0.366, respectively. Although INDIRECT did not attain the top LPIPS result in the image denoising task, its considerably smaller model size than MIRNet [7] underscores its efficiency and effectiveness.

Table 4 Comparison in terms of PSNR(↑)/SSIM(↑) with different SOTA image denoising methods

Method	Param		SIDD	DND
BM3D [22]	–		25.29/0.412	34.51/0.8507
DnCNN [41]	< 10M	0.56M	38.56/0.910	32.43/0.7900
GradNet [42]		1.60M	38.34/0.946	39.44/0.9543
InvDN [8]		2.64M	***39.23/0.914***	***39.19***/0.9484
INDIRECT		3.44M	*39.35/0.915*	39.22/0.9481
CBDNet [43]		4.34M	38.68/0.909	38.06/0.9421
VDN [44]		7.81M	39.29/0.911	39.38/0.9518
AINDNet [45]	≥10 M	13.76M	39.08/0.953	39.53/0.9561
MIRNet [7]		31.78M	**39.72/0.959**	**39.88/0.9563**
DANet [6]		63.01M	39.30/0.916	*39.58/0.9545*

Numbers in bold and italic indicate the best and the second-best performance, respectively. Numbers in bolditalic indicate the performance evaluated upon the source code. For the coupling layers, we apply the residual block as InvDN [8]

Figure 6 presents the qualitative results. As shown, although our method could not achieve better PSNR/SSIM results than MIRNet [7], it shows comparable subjective results, being able to remove the noise effectively. Moreover,

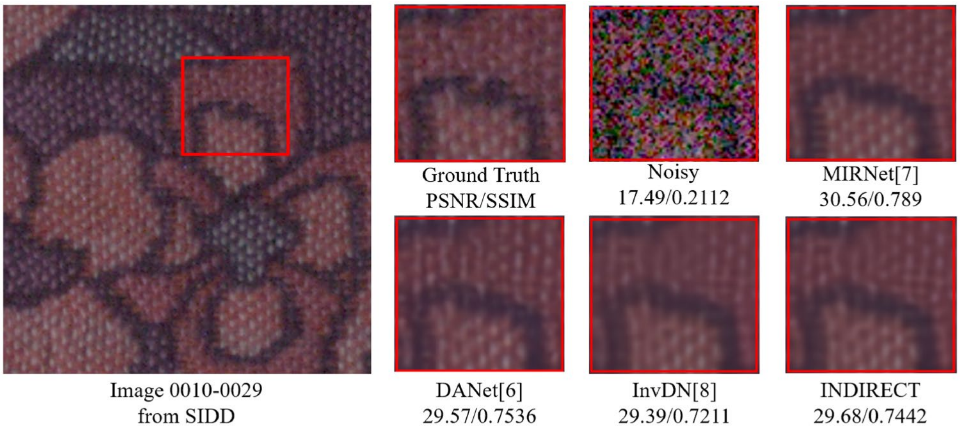


Fig. 6 Qualitative result comparison for image denoising

it reconstructs a clearer image than InvDN [8] and is comparable to DANet [6].

4.6 The effectiveness of CTM

As discussed in Sect. 3.3, the possible advantage of the proposed CTM over the traditional methods of using a case-independent distribution to model the HF distribution is that it allows for a more accurate and content-specific estimation of the missing HF information. By doing so, the CTM module facilitates a more precise reconstruction of the HR images.

As evidenced in Table 10, the integration of the IRN [5] module with the CTM for HFinfo estimation instead of using a case-independent Gaussian sample significantly improved the model's performance. Furthermore, incorporating the CTM into our DDR-net module to form the INDIRECT network has yielded the best results, substantiating the effectiveness of the CTM in enhancing high-frequency detail in HR image reconstruction.

Moreover, to show the effectiveness of CTM, we have conducted experiments comparing PSNR/SSIM metrics between INDIRECT and case-independent methods such as IRN [5] and InvDN [8] across three primary tasks: Image Noisy Image Re-scaling, Image Re-scaling, and Image Denoising. Please refer to the Sects. 4.3, 4.4, and 4.5 for detailed discussions. The results, presented in Table 1 for the Noisy Image Re-scaling task and Table 4 for the Image Denoising task, reveal that INDIRECT outperforms both IRN [5] and InvDN [8] on the SIDD [27] dataset. Additionally, for the Image Re-scaling task, as detailed in Table 2, our INDIRECT achieves superior performance in comparisons with IRN on various datasets, including Set5, Set14, BSDS100, Urban100, Manga109, and DIV2K. Finally, in these experiments, INDIRECT achieves good results with a more efficient number of parameters compared to other methods, underscoring its efficiency and effectiveness.

4.7 Quantization effects and LR images

Quantization effects: To show that our DDR-Net can effectively alleviate the quantization effect in the LR image for the rescaling task, Fig. 7 compares the HR reconstructed images $\tilde{x}_{clean}$ produced by DDR-Net and IRN [5]. For reconstruction, the quantized LR image $\tilde{x}_{LR}$ is used along with the original HFInfo x_{HF}, generated in the downscaling stage, in place of the estimated one $\tilde{x}_{HF}$. As shown, IRN [5] suffers from severe artifacts even with access to the original HFInfo. In contrast, under the same setting, INDIRECT generates much clearer details, which underlines its ability to mitigate the quantization issue.

Visual quality of LR images: Next, we visualize our LR images optimized for the noisy image rescaling and the image rescaling tasks, respectively. We also compare them with the corresponding LR images produced by bicubic downscaling, IRN [5], and InvDN [8]. Figure 8a demonstrates that there exist some minor differences between our LR image and the bicubic downscaled LR image. This is because we can embed more helpful information in the LR image to predict HFInfo.

4.8 Ablation studies

Our ablation studies analyze the effectiveness of DDR-Net and CTM. We use PSNR and SSIM as the quality metrics, and the evaluation is performed on the image rescaling task since these two modules mainly target the challenges of this task.

Haar vs Squeeze: First, in Table 5, we compare two different decomposition schemes: one is the discrete Haar transform adopted by our current design, and the other is Squeeze, which is very popular in normalizing flow models [18–20]. We observe that the discrete Haar transform

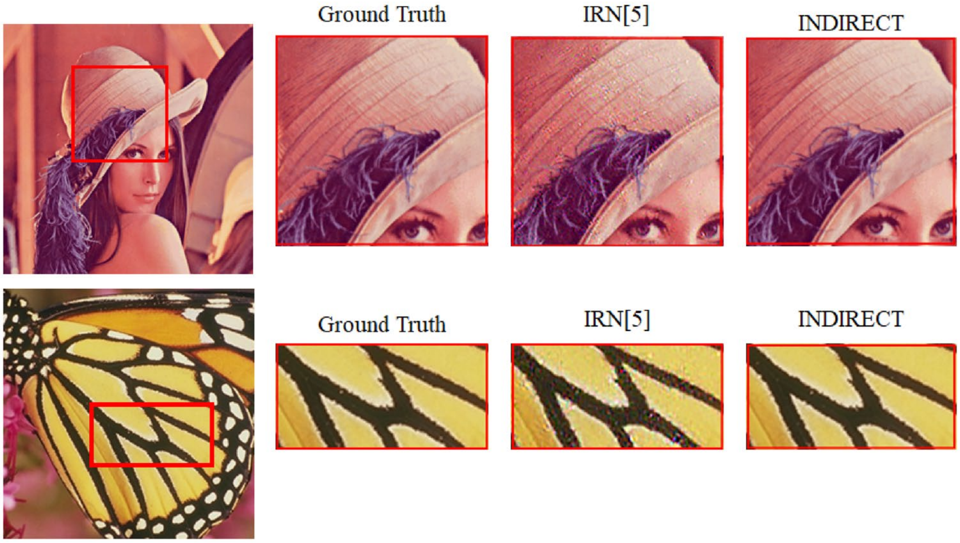


Fig. 7 Comparison of quantization effect on IRN [5] and INDIRECT

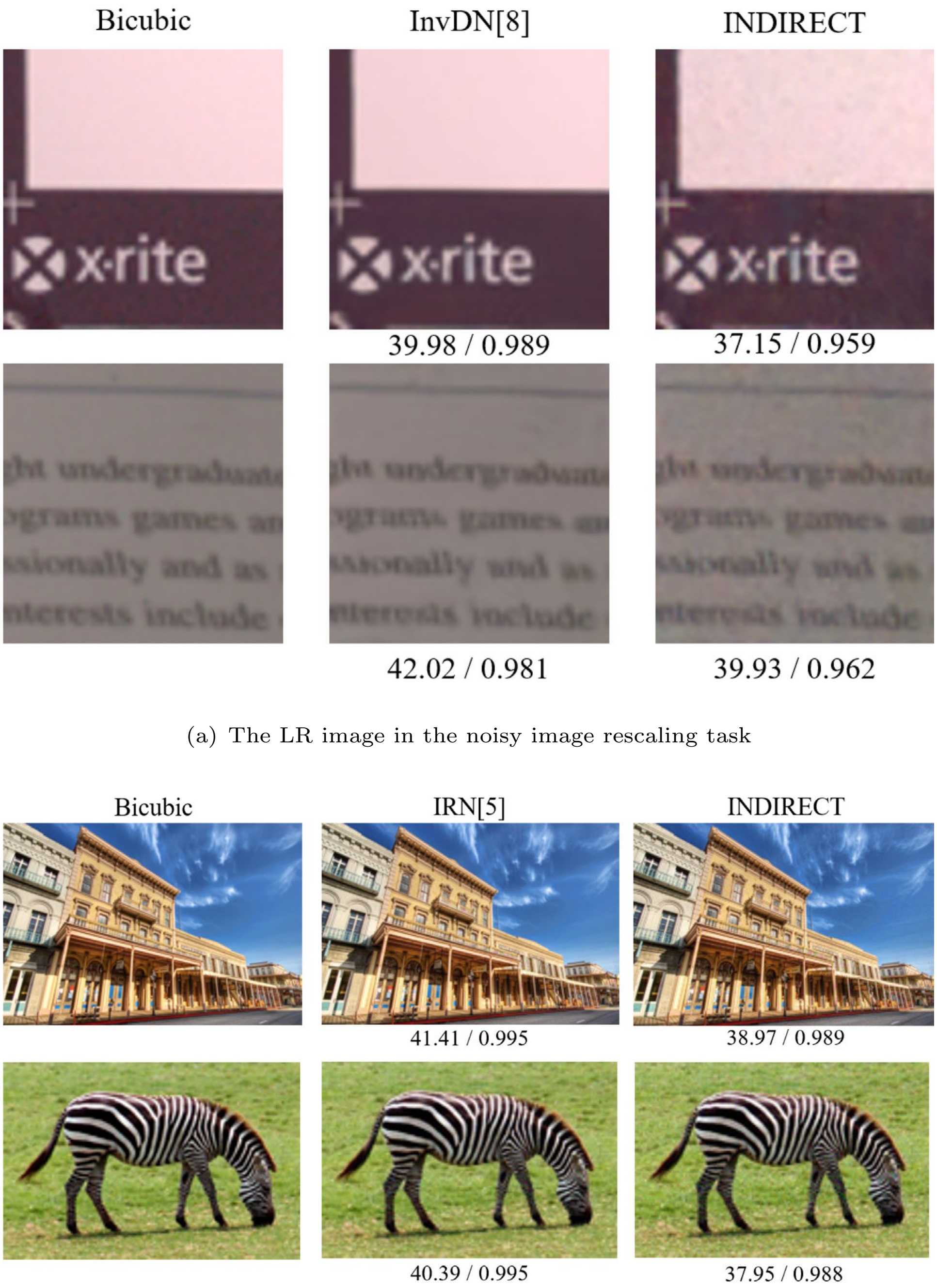
achieves better performance than Squeeze since the former implements the much desired inductive bias, which disentangles an input image into the low-frequency and high-frequency bands.

Numbers of discrete coupling layers: Next, we study the performance of INDIRECT with varying numbers of discrete coupling layers in the downscaling blocks. According to Table 6, we observe significant performance improvement when the number of coupling layers increases from 4 to 8. However, beyond 8 layers, there is a drop in performance.

Stochastic versus Deterministic CTM: We analyze the design of our CTM module, which can be implemented either in a stochastic manner or a deterministic way. In the deterministic implementation, a CNN-based prediction module is applied to estimate the case-dependent HFInfo directly. In the stochastic alternative, as shown in Fig. 3, we start with an encoder to embed and model the missing HFInfo into a feature space using a Gaussian distribution. Then, after drawing a sample from the distribution, we build the decoder upon a CNN network to recover HFInfo. From the results presented in Table 7, we observe that the deterministic approach performs worse than the stochastic design since it fails to model the one-to-many mapping uncertainty about HFInfo.

Design of objective functions: Next, different objective functions are compared. We analyze not only how L1/L2 loss affects the performance but also the necessity of the constraint on HFInfo. Specifically, the HFInfo $\tilde{x}_{HF}$ generated by CTM can be further restrained by the “true" high-frequency component x_{HF} disentangled from a noiseless HR image by DDR-Net. As shown in Table 8, we observe that INDIRECT shows the best performance when we constrain x_{LR} with L2 loss and $\tilde{x}_{HR}$ with L1 loss. Besides, Table 8 also shows that additionally imposing a loss, L1 or L2, on HFInfo does not bring further gain. Also, we conjecture that L2 loss is more preferred for L_{vis} because some HFInfo needs to be embedded in the LR image. Regarding the reconstruction loss, L1 loss is preferable since a clean and faithful HR image is our goal.

Effectiveness of hyper-parameters: In this analysis, we investigate the impact of the hyper-parameters, which helps the model focus either on the HR images or achieves a balance in quality between LR and HR images. In most of our experiments, we primarily focused on optimizing the parameters α and β as outlined in Equation 7. Particularly in tasks like Image Re-scaling, γ was set to 1 in our experiments to concentrate on the impacts of α and β. Accordingly, we set α to 1 and incrementally adjust β through 1, 2, 5, 10, and 15 to conduct the experiments. As depicted in Table 9, increasing the β results in performance improvement for HR images. However, this improvement comes at the cost of reducing the quality of LR images. This trade-off is because a higher weight on the reconstruction loss compels the model to incorporate more high-frequency information into the LR image, consequently degrading the LR quality. Furthermore, we observed that beyond a certain threshold of β, the quality of both LR and HR images declined significantly. This deterioration is marked when β is set to 15, leading to a considerable decrease in the quality of both LR and HR images. This result highlights the importance of carefully balancing these hyper-parameters to prevent adverse effects on image quality.



(a) The LR image in the noisy image rescaling task

(b) The LR image in the image rescaling task

Fig. 8 The visual quality comparison between our downscaled image and bicubic LR image (PSNR-Y/SSIM-Y provided)

Effectiveness of each submodule: Finally, we analyze how DDR-Net and CTM contribute to the final performance. In this analysis, IRN [5] is treated as our baseline model. From Table 10, when we deploy the CTM module in IRN to estimate a case-dependent HFInfo rather than drawing a case-agnostic Gaussian sample, the performance is improved. Moreover, the results of DDR-Net also evidence its effectiveness in alleviating the quantization issue. The improvement becomes even more significant when DDR-Net and CTM are combined to form our INDIRECT.

Table 5 Ablation study of various decomposition approaches in downscaling blocks

Dataset	Type of decomposition	
	(PSNR-Y(↑)/SSIM-Y(↑))	
	Haar	Squeeze
Set5	**36.40/0.9448**	36.36/0.9436
Set14	**33.07/0.9037**	**33.07**/0.9032
BSDS100	**31.95/0.8873**	31.91/0.8848
Urban100	**32.13/0.9227**	32.06/0.9198
Manga109	**36.65/0.9637**	36.58/0.9610
DIV2K	**35.43/0.9347**	35.29/0.9315

Numbers in bold indicate the best performance

Table 6 Ablation study of different decomposition approaches in downscaling blocks

Dataset	Number of discrete coupling layers		
	(PSNR-Y(↑)/SSIM-Y(↑))		
	4	8	12
Set5	35.88/0.9401	**36.40/0.9448**	*36.08/0.9409*
Set14	32.34/0.8912	**33.07/0.9037**	*32.61/0.8946*
BSDS100	31.35/0.8746	**31.95/0.8873**	*31.60/0.8778*
Urban100	31.07/0.9087	**32.13/0.9227**	*31.65/0.9147*
Manga109	35.59/0.9576	**36.65/0.9637**	*36.13/0.9594*
DIV2K	34.74/0.9268	**35.43/0.9347**	*35.00/0.9286*

Numbers in bold and italic indicate the best and the second-best performance, respectively

Table 7 Ablation study of different designs of the CTM module

Dataset	Type of CTM	
	(PSNR-Y(↑)/SSIM-Y(↑))	
	Deterministic	Stochastic (CNN-based decoder)
Set5	36.30/0.9446	**36.40/0.9448**
Set14	32.90/0.9026	**33.07/0.9037**
BSDS100	31.86/0.8865	**31.95/0.8873**
Urban100	31.84/0.9201	**32.13/0.9227**
Manga109	36.39/0.9631	**36.65/0.9637**
DIV2K	35.31/0.9342	**35.43/0.9347**

Numbers in bold indicate the best performance

Table 8 Ablation study of different objective functions of INDIRECT in terms of PSNR-Y(↑)/SSIM-Y(↑)

L_{vis}	L_{rec}	HFInfo Loss	Set5	Set14	BSD100	Urban100	Manga109	DIV2K
L1	L1	No	35.52/0.936	32.27/0.889	31.19/0.867	31.39/0.911	35.45/0.954	34.53/0.921
L1	L2	No	35.15/0.903	30.13/0.832	29.23/0.796	28.69/0.855	32.30/0.922	32.20/0.876
L2	L1	No	**36.40/0.945**	**33.07/0.904**	**31.95/0.888**	**32.13/0.923**	**36.65/0.964**	**35.43/0.935**
L2	L2	No	*36.17*/0.940	*32.92*/*0.896*	*31.94*/0.880	*31.93*/0.915	*36.29*/0.958	*35.29*/0.930
L2	L1	L1	35.58/0.936	32.00/0.885	31.19/0.869	31.10/0.869	35.53/0.958	34.61/0.924
L2	L1	L2	36.04/*0.941*	32.56/*0.896*	31.67/*0.883*	*31.58*/0.917	36.10/*0.962*	35.11/*0.932*

Numbers in bold and italic indicate the best and the second-best performance, respectively

Table 9 Ablation study of Effectiveness of alpha parameter

α	β	LR images (PSNR-Y(↑)/ SSIM-Y(↑)	HR images PSNR-Y(↑)/ SSIM-Y(↑)
1	1	**37.01/0.8866**	38.10/0.8963
1	2	*35.86/0.8588*	38.12/0.8968
1	5	33.97/0.7991	*38.22/0.8977*
1	10	32.81/0.8734	**38.40/ 0.9311**
1	15	30.80/0.8565	38.35/0.9303

Numbers in bold and italic indicate the best and the second-best performance, respectively

Table 10 Ablation study of different modules of INDIRECT in terms of PSNR-Y(↑)/SSIM-Y(↑)

Dataset	Method			
	IRN [5]	IRN [5] + CTM	DDR-Net	INDIRECT
Set5	36.19/*0.945*	*36.36*/**0.946**	36.34/*0.945*	**36.40**/*0.945*
Set14	32.67/0.902	32.87/*0.903*	*32.90*/0.902	**33.07/0.904**
BSD100	31.64/0.883	31.77/*0.885*	*31.87*/**0.887**	**31.95/0.887**
Urban100	31.41/0.916	31.64/0.918	*31.84/0.920*	**32.13/0.923**
Manga109	35.94/0.962	36.06/0.962	*36.45/0.963*	**36.65/0.964**
DIV2K	35.07/0.932	35.22/0.933	*35.31/0.934*	**35.43/0.935**

Numbers in bold and italic indicate the best and the second-best performance, respectively

5 Conclusion

This paper proposes an effective noisy image rescaling network that can jointly perform denoising and rescaling tasks with relatively few parameters.. The network, INDIRECT, not only addresses the quantization issue but also exploits the embedded case-dependent information in the LR image for better reconstruction. Specifically, DDR-Net is first designed to perform the twofold tasks efficiently and alleviate the quantization issue. Then we introduce CTM to generate more precise HFInfo while mitigating the ill-posed problem. Besides, we also introduce a perceptual version of INDIRECT to enhance the perceptual quality of the reconstructed HR image. The experimental results show that our method can achieve performance improvement both on the noisy image rescaling and image rescaling tasks. Regarding the stand-alone image denoising task, INDIRECT spends much fewer model parameters to attain a comparable result than SOTA methods. Moreover, the reconstructed clean HR images and the LR images show visually pleasing qualities with finer details.

Appendix: Symbol comparison table

We provide a table for symbol comparison as follows:

Mathematical notation	**Meaning**
x_{HR}	The input high-resolution image for the image rescaling task of IRN
$\tilde{x}_{HR}$	The reconstructed high-resolution image for the image rescaling task of IRN
x_{Noisy}	The input noisy high-resolution image for the noisy image rescaling task of INDIRECT
$\tilde{x}_{Clean}$	The reconstructed clean high-resolution image for the noisy image rescaling task of INDIRECT
x_{bic}	The bicubically downsampled image from a clean and high-resolution image (the ground truth image)
x_{LR}	The output low-resolution image from the downscaling path of IRN/INDIRECT
$\tilde{x}_{LR}$	The quantized and clipped version (clipped to 0~ 255 for each RGB channel) of x_{LR}
x_{HF}	The output high-frequency component from the downscaling path of IRN
x_{HF_Noisy}	The output noisy high-frequency component from the downscaling path of INDIRECT
$\tilde{x}_{HF}$	The generated high-frequency components from CTM
z	a random sample drawn from Gaussian distribution $p(Z)$
x_{gt}	The ground truth image that is clean and high-resolution for the noisy image rescaling task
ϕ and ρ	The transform functions in the discrete coupling layers. Their rounding operations are denoted as $\lfloor\phi()\rceil$ and $\lfloor\rho()\rceil$
x_{LT}	The low-frequency component that captures the information from the LL subband of the Haar transformation result
x_{HT}	The high-frequency component that captures the information from the LH, HL, and HH subbands of the Haar transformation result
x'_{LT}	The transformation result of x_{LT} by applying $\lfloor\phi()\rceil$
x'_{HT}	The transformation result of x_{HT} by applying $\lfloor\rho()\rceil$
N	The batch size
L_{vis}	The visual loss that constrains x_{LR} by the downscaled ground truth image x_{bic}
L_{rec}	The reconstruction loss that constrains $\tilde{x}_{Clean}$ by x_{gt}
L_{per}	The perceptual loss that constrains the perceptual quality of $\tilde{x}_{Clean}$ by the feature similarities to x_{gt}

Author contributions All authors have contributed equally. All authors reviewed the manuscript.

Funding Open Access funding enabled and organized by National Yang Ming Chiao Tung University. Funding was provideb by National Science and Technology Council, Taiwan (Grant Nos. NSTC-112-2221-E-A49-089-MY3, NSTC-110-2221-E-A49-066-MY3, NSTC-111- 2634-F-A49-010, NSTC-112-2425-H-A49-001).

Data availability The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Sun, W., Chen, Z.: Learned image downscaling for upscaling using content adaptive resampler. IEEE Trans. Image Process. **29**, 4027–4040 (2020)
2. Kim, H., Choi, M., Lim, B., Lee, K.M.: Task-aware image downscaling. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 399–414 (2018)
3. Li, Y., Liu, D., Li, H., Li, L., Li, Z., Wu, F.: Learning a convolutional neural network for image compact-resolution. IEEE Trans. Image Process. **28**(3), 1092–1107 (2018)
4. Chen, Y., Xiao, X., Dai, T., Xia, S.-T.: Hrnet: Hamiltonian rescaling network for image downscaling. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 523–527. IEEE (2020)
5. Xiao, M., Zheng, S., Liu, C., Wang, Y., He, D., Ke, G., Bian, J., Lin, Z., Liu, T.-Y.: Invertible image rescaling. In: European Conference on Computer Vision, pp. 126–144. Springer (2020)
6. Yue, Z., Zhao, Q., Zhang, L., Meng, D.: Dual adversarial network: toward real-world noise removal and noise generation. In: European Conference on Computer Vision, pp. 41–58. Springer (2020)
7. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H., Shao, L.: Learning enriched features for real image restoration and enhancement. In: Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16, pp. 492–511. Springer (2020)
8. Liu, Y., Qin, Z., Anwar, S., Ji, P., Kim, D., Caldwell, S., Gedeon, T.: Invertible denoising network: a light solution for real noise removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13365–13374 (2021)
9. Hu, Y., Li, J., Huang, Y., Gao, X.: Image super-resolution with self-similarity prior guided network and sample-discriminating learning. IEEE Trans. Circuits Syst. Video Technol. **32**(4), 1966–1985 (2022). https://doi.org/10.1109/TCSVT.2021.3093483
10. Yue, H., Liu, J., Yang, J., Sun, X., Nguyen, T.Q., Wu, F.: Ienet: internal and external patch matching convnet for web image guided denoising. IEEE Trans. Circuits Syst. Video Technol. **30**(11), 3928–3942 (2020). https://doi.org/10.1109/TCSVT.2019.2930305
11. Jiang, B., Lu, Y., Wang, J., Lu, G., Zhang, D.: Deep image denoising with adaptive priors. IEEE Trans. Circuits Syst. Video Technol. **32**(8), 5124–5136 (2022)
12. Xu, J., Xu, L., Gao, Z., Lin, P., Nie, K.: A denoising method based on pulse interval compensation for high-speed spike-based image sensor. IEEE Trans. Circuits Syst. Video Technol. **31**(8), 2966–2980 (2021)
13. Zhang, X., Zheng, J., Wang, D., Zhao, L.: Exemplar-based denoising: a unified low-rank recovery framework. IEEE Trans. Circuits Syst. Video Technol. **30**(8), 2538–2549 (2020)
14. 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, pp. 624–632 (2017)
15. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144 (2017)
16. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301 (2018)
17. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: ESRGAN: enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)
18. Dinh, L., Krueger, D., Bengio, Y.: NICE: non-linear independent components estimation (2014). arXiv preprint arXiv:1410.8516
19. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp (2016). arXiv preprint arXiv:1605.08803
20. Kingma, D.P., Dhariwal, P.: Glow: generative flow with invertible 1x1 convolutions (2018). arXiv preprint arXiv:1807.03039
21. Buades, A., Coll, B., Morel, J.-M.: A non-local algorithm for image denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2, pp. 60–65. IEEE (2005)
22. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Trans. Image Process. **16**(8), 2080–2095 (2007)
23. Chen, Y.-A., Hsiao, C.-C., Peng, W.-H., Huang, C.-C.: Direct: discrete image rescaling with enhancement from case-specific textures. In: 2021 International Conference on Visual Communications and Image Processing (VCIP), pp. 1–5 (2021)
24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241. Springer (2015)
25. Hoogeboom, E., Peters, J.W., Berg, R.v.d., Welling, M.: Integer discrete flows and lossless compression (2019). arXiv preprint arXiv:1905.07376
26. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, pp. 694–711. Springer (2016)
27. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1692–1700 (2018). https://www.eecs.yorku.ca/~kamel/sidd/dataset.php
28. Plotz, T., Roth, S.: Benchmarking denoising algorithms with real photographs. In: Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition, pp. 1586–1595 (2017). https://noise.visinf.tu-darmstadt.de

29. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 126–135 (2017). https://data.vision.ee.ethz.ch/cvl/DIV2K/
30. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding (2012)
31. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: International Conference on Curves and Surfaces, pp. 711–730. Springer (2010)
32. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, vol. 2, pp. 416–423. IEEE (2001)
33. Huang, J.-B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5197–5206 (2015). https://opendatalab.com/OpenDataLab/Urban100
34. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. Multimed. Tools Appl. **76**(20), 21811–21838 (2017). https://www.manga109.org/en/index.html
35. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
36. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
37. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind ’ ’ image quality analyzer. IEEE Signal Process. Lett. **20**(3), 209–212 (2012)
38. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2014). arXiv preprint arXiv:1412.6980
39. Dai, T., Cai, J., Zhang, Y., Xia, S.-T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11065–11074 (2019)
40. Liang, J., Lugmayr, A., Zhang, K., Danelljan, M., Van Gool, L., Timofte, R.: Hierarchical conditional flow: a unified framework for image super-resolution and image rescaling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4076–4085 (2021)
41. 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. **26**(7), 3142–3155 (2017)
42. Liu, Y., Anwar, S., Zheng, L., Tian, Q.: GradNet image denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 508–509 (2020)
43. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1712–1722 (2019)
44. Yue, Z., Yong, H., Zhao, Q., Zhang, L., Meng, D.: Variational denoising network: toward blind noise modeling and removal (2019). arXiv preprint arXiv:1908.11314
45. Kim, Y., Soh, J.W., Park, G.Y., Cho, N.I.: Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3482–3492 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.