Invertible Residual Rescaling Models

Jinmin Li$^1$, Tao Dai$^{2, *}$, Yaohua Zha$^{1,3}$, Yilu Luo$^1$, Longfei Lu$^1$, Bin Chen$^4$, Zhi Wang$^1$, Shu-Tao Xia$^{1,3}$, Jingyun Zhang$^5$

$^1$Tsinghua Shenzhen International Graduate School, Tsinghua University
$^2$College of Computer Science and Software Engineering, Shenzhen University
$^3$Research Center of Artificial Intelligence, Peng Cheng Laboratory
$^4$Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
$^5$WeChat Pay Lab33, Tencent

{ljm22, chayh21, luo-yl22}@mails.tsinghua.edu.cn, {daitao.edu, loneffy.lu, zhang304973926}@gmail.com, chenbin2021@hit.edu.cn, {wangzhi, xiast}@sz.tsinghua.edu.cn

Abstract

Invertible Rescaling Networks (IRNs) and their variants have witnessed remarkable achievements in various image processing tasks like image rescaling. However, we observe that IRNs with deeper networks are difficult to train, thus hindering the representational ability of IRNs. To address this issue, we propose Invertible Residual Rescaling Models (IRRM) for image rescaling by learning a bijection between a high-resolution image and its low-resolution counterpart with a specific distribution. Specifically, we propose IRRM to build a deep network, which contains several Residual Downscaling Modules (RDMs) with long skip connections. Each RDM consists of several Invertible Residual Blocks (IRBs) with short connections. In this way, RDM allows rich low-frequency information to be bypassed by skip connections and forces models to focus on extracting high-frequency information from the image. Extensive experiments show that our IRRM performs significantly better than other state-of-the-art methods with much fewer parameters and complexity. Particularly, our IRRM has respectively PSNR gains of at least 0.3 dB over HCFlow and IRN in the ×4 rescaling while only using 60% parameters and 50% FLOPs. The code will be available at https://github.com/THU-Kingmin/IRRM.

1 Introduction

Image rescaling, which aims to reconstruct high-resolution (HR) images from their corresponding low-resolution (LR) versions by forward sampling, has been widely utilized in large-size data services like storage and transmission. Recently, many flow-based generative methods [Ho et al., 2019; Kingma and Dhariwal, 2018; Nielsen et al., 2020; Xiao et al., 2020; Liang et al., 2021b] have been applied in image generation and achieved remarkable progress in image rescaling and image super-resolution (SR).

*Corresponding author: Tao Dai (daitao.edu@gmail.com)
model to model the HR-LR relationship, where the high-frequency component is hierarchically conditional on the low-frequency component of the image. Despite the success of IRN in image rescaling, such IRNs are difficult to train, especially with deep layers.

To solve this problem, in this paper, we propose a novel Invertible Residual Rescaling Model (IRRM) to facilitate training while enhancing the representational ability of models. To ease the training, we propose Residual Downscaling Module (RDM) with long skip connections, which serve as the basic block of our IRRM and long skip connection allows rich low-frequency information to be bypassed. In each RDM module, we stack several invertible residual blocks (IRB) to enhance non-linear representation and reduce model degradation with short skip connections. With long and short skip connections, abundant information can be bypassed and thus ease the flow of information. As shown in Fig. 1, our IRRM obtains better results than other state-of-the-art methods with much fewer parameters and complexity.

In summary, the main contributions are as follows:

- We propose a novel Invertible Residual Rescaling Model (IRRM) to build deep networks for highly accurate image rescaling. Our IRRM obtains much better performance than previous methods with much fewer parameters and complexity.
- We propose Residual Downscaling Module (RDM) with long skip connections, equivalent to the second-order wavelet transform, to allow the model to focus on learning the texture information of the image while easing the flow of information.
- Our proposed IRRM introduces the Invertible Residual Block (IRB), which incorporates short skip connections to enhance the model’s nonlinear representational ability. This addition significantly improves the extensibility of the model.

2 Related Work

2.1 Image Super-Resolution

Image super-resolution is a different task from rescaling, aiming to restore HR image given the LR image. However, SR combined with downsampling can be used for image rescaling task. Recently, SR networks have achieved impressive performance based on deep learning [Dai et al., 2023; Li et al., 2023; Cui et al., 2023; Zhang et al., 2023; Tang et al., 2024; Guo et al., 2024]. SRCNN [Dong et al., 2014] is the first work that applies convolutional neural networks for image SR. Later, many methods [Kim et al., 2016; Ledig et al., 2017] stack many convolutional layers with residual connections to facilitate the network training. Recently, other advanced methods like RCAN [Zhang et al., 2018], SAN [Dai et al., 2019] and SwinIR [Li et al., 2021a] build very deep networks to enhance the feature expression ability and have obtained remarkable performance with substantial computational cost. On the other hand, lightweight SR works have been proposed to relieve the problem of computational cost [Hui et al., 2019; Luo et al., 2020]. For example, the lightweight frequency-aware network (FADN) [Xie et al., 2021] is developed to reconstruct different frequency signals with various operations. Instead of designing a lightweight SR network, ClassSR [Kong et al., 2021] develops a general framework to accelerate SR networks. However, they seem to have reached the limits of performance for image generation tasks.

2.2 Image Rescaling

The goal of image rescaling is to downscale a high-resolution (HR) image into a visually satisfying low-resolution (LR) image and then restore the HR image from the LR image. Typically, image rescaling concurrently models the enlargement and reduction processes within an encoder-decoder structure to fine-tune the reduction model for subsequent enlargement tasks [Kim et al., 2018; Li et al., 2018; Li et al., 2024]. Recently, IRN [Xiao et al., 2020] has been introduced to model the rescaling process using a bijective invertible neural network. During the training phase, IRN effectively captures the high-frequency and low-frequency components. During the testing phase, the HR image can be reconstructed by the generated LR image and a randomly sampled Gaussian distribution. Besides, HCFlow [Liang et al., 2021b] assumes that the high-frequency component is dependent on the LR image and employs a hierarchical conditional framework to model the LR image from conditional distribution of the high-frequency component. Nevertheless, these models cannot extend to larger models because of the excessive focus on low frequency information and the lack of residual connections. We address these issues by introducing residual connections and residual enhanced block equivalent to the second-order wavelet transform.

3 Methodology

3.1 Model Specification

Image rescaling involves reconstructing a high-resolution (HR) image $x$ from a low-resolution (LR) image $y$ that is
obtained by downscaling \( x \). To achieve this, we employ an invertible neural network that models the LR image \( y \) and ensures that the distribution of the lost information conforms to a predetermined distribution \( z \) (denoted as \( [y, z] = f_0(x) \)), where \( z \) is sampled from \( p(z) \) and \( x \) follows the distribution \( p(x) \). Conversely, we can reconstruct \( x \) through the inverse process using \([y, z]\), denoted as \( x = f_y^{-1}(y, z) \). It is important to note that the lost information \( z \) corresponds to the high-frequency details of the image, as per the Nyquist-Shannon sampling theorem [Shannon, 1949]. Instead of explicitly preserving the high-frequency information, we can model \( z \) as a Gaussian distribution, thereby obtaining it through sampling from a Gaussian distribution.

### 3.2 Invertible Residual Rescaling Models

As shown in Fig. 3, our proposed IRRM consists of stacked Residual Downscaling Module (RDM), where RDM further contains one Haar Wavelet Transformation block and several Invertible Residual Blocks (IRBs).

**Residual Downscaling Module.** RDM utilizes Haar transformation [Mallat, 1989] to split the input into high and low frequency information. Then IRB learn to model these frequency information into the specified distribution. Specifically, given an HR image \( x \) with shape \((H, W, C)\), we first obtain the residual representation of \( x \):

\[
x_{\text{res}} = x - \text{upsample}(y)
\]

where \( y \) is the LR image and \( y = \text{downsample}(x) \). Then wavelet transformation decomposes \( x_{\text{res}} \) into global frequency features \([u_1^k, u_2^k]\):

\[
[u_1^k, u_2^k] = \text{WaveletTransform}(x_{\text{res}}^k)
\]

where \( k \) denotes the \( k\)th layer of the model. \( u_1 \) and \( u_2 \) correspond to high-frequency and low-frequency information, respectively. \( x_{\text{res}} \) facilitates the model to focus on learning high frequency representations, aiming to recover textures of the image, which is equivalent to the second-order wavelet transform (IRN [Xiao et al., 2020] is the first-order wavelet transform). Meanwhile, residual connections make the training process more stable in favour of convergence, as shown in Fig. 2. IRB extracts image features from \( u_1^k \) and \( u_2^k \) to \( u_1^{k+1} \) and \( u_2^{k+1} \). After \( n \) IRBs’ transformation, we obtain the downscaled image \( y \) with shape \((\frac{1}{2}H, \frac{1}{2}W, C)\) and Gaussian distribution \( z \) with shape \((\frac{1}{2}H, \frac{1}{2}W, 3C)\).

**Invertible Residual Block.** To be specific, wavelet features \( u_1^k \) and \( u_2^k \) are fed into stacked IRBs to the obtains LR \( y \) and latent variable \( z \). We adopt the general coupling layer for...
invertible architecture [Jacobsen et al., 2018; Behrmann et al., 2019]. The output of each IRB can be defined as:

\[
\begin{align*}
u_{k+1}^1 &= u_k^1 \cdot (EB(u_k^1)) + EB(u_k^1) \\
u_{k+1}^2 &= u_k^2 \cdot (EB(u_{k+1}^1)) + EB(u_k^1) \\
u_k^2 &= (u_{k+1}^1 - EB(u_{k+1}^1)) / (EB(u_{k+1}^1)) \\
u_k^1 &= (u_{k+1}^1 - EB(u_{k+1}^2)) / (EB(u_{k+1}^2))
\end{align*}
\]

where \(u_{k+1}^1\) and \(u_{k+1}^2\) are the output of current IRB and the input of next IRB. Note that EB can be arbitrary convolutional unit to enhance the non-linear capability of the model. In our IRRM, we employ two residual connected convolutional blocks and one plain convolutional block [He et al., 2016; Liu et al., 2021] as the enhanced block, as shown in Fig. 4. RB in EB enhances the non-linearity and the residual connections make the model training stable, solving the problem of unstable training of plain convolutional layers.

### 3.3 Loss Functions

Image rescaling aims to reconstruct exactly the HR image while generating visually pleasing LR image. Following IRN [Xiao et al., 2020], our IRRM is trained by minimizing the following loss:

\[
\mathcal{L} = \lambda_1 \mathcal{L}_{\text{back}}(x, x_{\text{back}}) + \lambda_2 \mathcal{L}_{\text{forw}}(y, y_{\text{forw}}) + \lambda_3 \mathcal{L}_{\text{latent}}(z)
\]

(4)

where \(x\) is ground-truth HR image, \(y\) is ground-truth LR image, and \(x_{\text{back}}\) is reconstructed HR image from forward LR image \(y_{\text{forw}}\) and sampling latent variable \(z\).

\(\mathcal{L}_{\text{back}}\) is the \(l_1\) pixel loss defined as:

\[
\mathcal{L}_{\text{back}}(x, x_{\text{back}}) = \frac{1}{M} \sum_{i=1}^{M} \| f_{\theta}^{-1}(y, z) - x \|_1
\]

(5)

where \(M\) is the number of pixels and \(x_{\text{back}} = f_{\theta}^{-1}(y, z)\).

Likewise, \(\mathcal{L}_{\text{forw}}\) is the \(l_2\) pixel loss defined as:

\[
\mathcal{L}_{\text{forw}}(y, y_{\text{forw}}) = \frac{1}{N} \sum_{i=1}^{N} \| f_{\theta}^y(x, z) - y \|_2
\]

(6)

where \(N\) is the number of pixels and \(y_{\text{forw}} = f_{\theta}^y(x)\) is the ground-truth LR image.

The last term \(\mathcal{L}_{\text{latent}}(z)\) is the \(l_2\) regularization on the latent variable \(z\) to ensure that \(z\) follows a Gaussian distribution. We jointly optimise the invertible architecture \(f\) by utilizing both forward and backward losses.
Figure 5: Visual results of upscaling the $4 \times$ downscaled images. The right images are $128 \times 128$ which is a patch of the left images. IRRM recovers rich textures and realistic details, leading to better recovery performance. IRRM achieves better performance with an increased PSNR of 4 dB over RCAN and 0.6 dB over IRN.

4 Experiments

4.1 Settings

Datasets. The DIV2K dataset [Agustsson and Timofte, 2017] is adopted to train our IRRM. Firstly, we prepare the HR images by cropping the original images with steps 240 to $480 \times 480$. These HR images are down-sampled with scaling factors 0.25 and 0.5 to obtain the LR images. Besides, we evaluate our models with PSNR and SSIM metrics (Y channel) on commonly used benchmarks: Set5 [Bevilacqua et al., 2012], Set14 [Yang et al., 2010], B100 [Martin et al., 2001], Urban100 [Huang et al., 2015] and DIV2K valid [Agustsson and Timofte, 2017] datasets.

Training Details. Our IRRM is composed of one or two Residual Downscaling Modules (RDMs), containing eight Invertible Residual Blocks (IRBs). These models are trained using the ADAM [Kingma and Ba, 2014] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is set to 16 for per GPU, and the initial learning rate is fixed at $2 \times 10^{-4}$, which decays by half every 10k iterations. PyTorch is used as the implementation framework. $\lambda_1$, $\lambda_2$ and $\lambda_3$ are set to 8, 8 and 1, respectively. Finally, we augment training images by flipping and rotating.

4.2 Comparison with State-of-the-art Method

Performance Comparison. We compare our IRRM with State-of-the-arts (like HAT [Chen et al., 2023], IRN [Xiao et al., 2020], HCFlow [Liang et al., 2021b] and so on) on commonly used datasets (Set5 [Bevilacqua et al., 2012], Set14 [Yang et al., 2010], B100 [Martin et al., 2001], Urban100 [Huang et al., 2015] and DIV2K valid [Agustsson and Timofte, 2017]) evaluated by Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM).

As shown in Table 1, the quantitative results of different rescaling methods are summarized. IRRM significantly outperforms previous state-of-the-art methods regarding PSNR and SSIM in five benchmark datasets. Previous SR methods have limited performance in rescaling tasks because of fixed downscaling process. IRN [Xiao et al., 2020], HCFlow [Liang et al., 2021b] and our IRRM optimize the upscaling and downscaling models by joint optimization based on the invertible architecture, further boosting the the PSNR metric about 5 dB on different benchmark datasets. Specifically, HCFlow achieves better performance with an increased PSNR of 0.2 dB over IRN in the $4 \times 4$ rescaling. Further, compared with IRN, our proposed IRRM improves 0.6 dB PSNR in the $4 \times 4$ rescaling and 2 dB PSNR in the $2 \times 2$ rescaling while only using half of the parameters. IRRM-M achieves per-
Figure 6: Visual results of Local Attribution Maps (LAM) for IRN and IRRM (ours). The LAM, calculated by diffusion index (DI), indicates the contribution of each pixel in the input image w.r.t. the patch marked with a red box in the HR image. A higher DI, illustrated in the second line, means a wider area of attention in the third lines.

Figure 7: Exploring model extensibility. As the model gets larger, IRRM without residual connections is extremely unstable in training. IRRM with residual connections and RB is stable and converges easily. PCB: plain convolution block. Res: residual connections in the Residual Downscaling Module. RB: residual block in the Invertible Residual Block. -S, -M, -L: denoted the model sizes.

4.3 Ablation Study

Residual Downscaling Module. We discard residual connections of RDM to investigate the contribution of them. As shown in Fig. 7, IRRM is more extensible with residual connections and RB. As the model gets deeper, the absence of residual connections can lead to instability of the model gradient, with problems of gradient vanishing and gradient explosion. Furthermore, IRRM with residual connections has better extensibility while the performance degrades as the network gets larger (8 RBs → 12 RBs) without residual connections, as shown below in Table 2.

Invertible Residual Block. To investigate the effects of Residual Block (EB) in the IRB, we replace it with a plain convolutional layer. The detailed structure is shown in Fig. 4.

-performance comparable to IRN with only 1/4 parameters, and IRRM-S achieves performance well beyond that of previous SR methods with less than 1M parameters. These experimental results demonstrate that our model is efficient.

Besides, we visualize the rescaling resulting images as shown in Fig. 5, including two common scenes. HR images restored by IRRM achieve better visual quality and rich textures than those of previous state-of-the-art methods. This advanced capability is attributed to residual connections in RDM, which enables the model to concentrate on learning high-frequency information, and the incorporation of Invertible Residual Block (IRB), which strengthens non-linear capacity of the model.

Interpretation with Local Attribution Maps (LAM). We analyze the diffusion index (DI) of previous methods and IRRM to better understand why IRRM works. DI is applied to measure local attribution maps (LAM [Gu and Dong, 2021]) from the input image to output image of models. When considering the same local patch, if LAM contains a greater range of pixels or encompasses a larger area, it can be inferred that the models have effectively utilized information from a larger number of pixels. As shown in Fig. 6, our IRRM utilizes a wider range of pixels for better rescaling for models. Specifically, the DI of IRRM is four times that of IRN due to the fact that we have introduced residual connections in IRRM, which makes the receptive field of the model larger.

- Performance comparable to IRN with only 1/4 parameters, and IRRM-S achieves performance well beyond that of previous SR methods with less than 1M parameters. These experimental results demonstrate that our model is efficient.

- Besides, we visualize the rescaling resulting images as shown in Fig. 5, including two common scenes. HR images restored by IRRM achieve better visual quality and rich textures than those of previous state-of-the-art methods. This advanced capability is attributed to residual connections in RDM, which enables the model to concentrate on learning high-frequency information, and the incorporation of Invertible Residual Block (IRB), which strengthens non-linear capacity of the model.

- Interpretation with Local Attribution Maps (LAM). We analyze the diffusion index (DI) of previous methods and IRRM to better understand why IRRM works. DI is applied to measure local attribution maps (LAM [Gu and Dong, 2021]) from the input image to output image of models. When considering the same local patch, if LAM contains a greater range of pixels or encompasses a larger area, it can be inferred that the models have effectively utilized information from a larger number of pixels. As shown in Fig. 6, our IRRM utilizes a wider range of pixels for better rescaling for models. Specifically, the DI of IRRM is four times that of IRN due to the fact that we have introduced residual connections in IRRM, which makes the receptive field of the model larger.

4.3 Ablation Study

Residual Downscaling Module. We discard residual connections of RDM to investigate the contribution of them. As shown in Fig. 7, IRRM is more extensible with residual connections and RB. As the model gets deeper, the absence of residual connections can lead to instability of the model gradient, with problems of gradient vanishing and gradient explosion. Furthermore, IRRM with residual connections has better extensibility while the performance degrades as the network gets larger (8 RBs → 12 RBs) without residual connections, as shown below in Table 2.

Invertible Residual Block. To investigate the effects of Residual Block (EB) in the IRB, we replace it with a plain convolutional layer. The detailed structure is shown in Fig. 4.
As shown above in Table 2, RB performs better than PCB, due to residual connections. In addition, RB+ achieves comparable performance through half of the parameters.

As illustrated in Table 3, the performance of IRRM with residual connections and RB improves as the model size increases (IRRM-S (RB, Res) \rightarrow IRRM-M (RB, Res) \rightarrow IRRM-L (RB, Res)). However, in the case of IRRM without residual connections, the performance does not improve as the model size increases (IRRM-S (PCB) \rightarrow IRRM-M (PCB) \rightarrow IRRM-L (PCB)). Notably, IRRM (PCB) achieves superior performance in this scenario.

### 4.4 Visualisation on the Influence of $z$

We also analyse the effect of different scales of $z$ on the reconstructed HR image, as shown in Fig. 8. It is observed that our proposed IRRM is insensitive to the Gaussian distribution $z$ and generates unrealistic high-frequency details only when $z$ is very large. $z$ represents the high-frequency components in an image and is typically sparse. However, changes in $z$ have a significant impact on the image details. Therefore, it is essential for the model to exhibit a high tolerance towards $z$ in order to preserve image details effectively.

### 4.5 Real-world Datasets

Real-world dataset is also applicable for our method, since the inputs of our model are high-resolution images, whether they are real-world datasets or synthetic datasets. The results are shown in Table 4.

With the recent rapid development in multimodality [Bai et al., 2024; Li et al., 2024b; Gao et al., 2023; Wang et al., 2023], we will also explore the help of textual cues in the future to better exploit the potential of the Rescaling task.

### 5 Conclusion

In this paper, we develop an efficient and light-weight framework IRRM for image rescaling. IRRM learns a bijection between HR image and LR image as well as specific distribution $z$ so that it is information lossless and invertible. Specifically, IRRM introduces residual connections in the RDM, equivalent to a second-order wavelet transform. Direct modeling of high frequency information reduces the training difficulty and allows the training gradient to trend towards brown noise rather than white noise. Besides, we propose IRB to enhance non-linear representation and reduce model degradation, while the block is lightweight and lossless. This addition significantly improves the extensibility of the model.

### Acknowledgements

This work is supported in part by the National Natural Science Foundation of China, under Grant (62302309, 62171248), Shenzhen Science and Technology Program (JCYJ2022018101014030, JCYJ2022018101012025), and the PCNL KEY project (PCL2023AS6-1), and Tencent “Rhinoceros Birds” - Scientific Research Foundation for Young Teachers of Shenzhen University.
References


