Effective Invertible Arbitrary Image Rescaling

Zhihong Pan\textsuperscript{1}  Baopu Li\textsuperscript{2*}  Dongliang He\textsuperscript{3}  Wenhao Wu\textsuperscript{3}  Errui Ding\textsuperscript{3}

\textsuperscript{1}Baidu Research (USA)  \textsuperscript{2}Oracle Health and AI  \textsuperscript{3}Baidu VIS

zhihongpan@baidu.com, baopu.li@oracle.com, \{hedongliang01, wuwenhao01, dingerrui\}@baidu.com

Abstract

Great successes have been achieved using deep learning techniques for image super-resolution (SR) with fixed scales. To increase its real world applicability, numerous models have also been proposed to restore SR images with arbitrary scale factors, including asymmetric ones where images are resized to different scales along horizontal and vertical directions. Though most models are only optimized for the unidirectional upscaling task while assuming a predefined downscaling kernel for low-resolution (LR) inputs, recent models based on Invertible Neural Networks (INN) are able to increase upscaling accuracy significantly by optimizing the downsampling and upscaling cycle jointly. However, limited by the INN architecture, it is constrained to fixed integer scale factors and requires one model for each scale. Without increasing model complexity, a simple and effective invertible arbitrary rescaling network (IARN) is proposed to achieve arbitrary image rescaling by training only one model in this work. Using innovative components like position-aware scale encoding and preemptive channel splitting, the network is optimized to convert the non-invertible rescaling cycle to an effectively invertible process. It is shown to achieve a state-of-the-art (SOTA) performance in bidirectional arbitrary rescaling without compromising perceptual quality in LR outputs. It is also demonstrated to perform well on tests with asymmetric scales using the same network architecture.

1. Introduction

Recent deep learning based image super-resolution (SR) methods have advanced the performance of image upscaling significantly but they are often limited to fixed integer scale factors and pre-determined downscaling degradation kernels. To work in real world applications where an image is commonly rescaled to arbitrary sizes, additional image resizing is often needed, which leads to degradation in both performance and efficiency. Lately there are growing interests in SR models that support arbitrary scale factors and great successes have been achieved in recent works [5, 7, 21]. However, they are only optimized for the unidirectional upscaling process with the LR inputs either synthesized from a predefined downscaling kernel or in its native resolution. Considering the potential mutual benefits between downsampling and the inverse upscaling, some image rescaling models [9, 19, 24] are developed to optimize these two processes jointly and significant improvements in upscaling accuracy are achieved comparing to unidirectional SR models of the same scale factors. The state-of-the-art (SOTA) performance for such bidirectional image rescaling is set by the invertible rescaling net (IRN) as proposed by Xiao et al. [24]. As shown in Fig. 1, it is able to achieve the best performance so far since both the Haar transformation and the invertible neural network (INN) [2] backbone are invertible processes, and its forward and invertible backward operations can model the downsampling and inverse upscaling cycle naturally. Denoting the forward downsampling process as $(y_L, z_L) = f(x_H)$, the HR image can be fully restored as $x_H = f^{-1}(y_L, z_L)$ if the latent variable $z_L$ is preserved. When the network is optimized to store as much information as allowed in $y_L$ and convert $z_L$ as input-independent random variables, the optimal HR output $x_H$ can be restored as $f^{-1}(y_L, \hat{z}_L)$ where $\hat{z}_L$ is randomly sampled with minimized loss in restoration accuracy. However, limited by the nature of INN architecture that the number of pixels must be equivalent when differences in resolution are accounted for (LR features have more channels), the applicable scale factors are limited to integers.

To overcome the above issues, we propose a new invertible arbitrary rescaling network (IARN) that is able to implement bidirectional arbitrary rescaling using similar INN backbone and it is able to train one model for a range of arbitrary scales, including asymmetric ones. As shown in Fig. 1, it replaces the Haar wavelet transformation blocks in IRN with a preemptive channel-splitting step denoted as

\begin{equation}
\begin{aligned}
x_H^{\text{if}} &= s(x_H) \\
x_H^{\text{ii}} &= x_H - x_H^{\text{if}}
\end{aligned}
\end{equation}

where subscript $H$ is used to specify that all images are preserved in original high-resolution. Note that this step is
Invertible Neural Network

Haar Transform

Downscale Upscale

Invertible Neural Network Channel Split

Rescaling Invertible

IRN IARN

Figure 1: Comparison between IRN [24] and the proposed IARN. \( x_H, y_L \) and \( z_{H,L} \) denote input HR image, output LR image and the latent variables. Note that subscripts \( H, L \) refer to high- and low-resolution respectively, and superscripts \( hf, lf \) refer to high and low frequency channels respectively. The main differences in IARN include channel-splitting in high-resolution in place of Haar transform, the additional near-invertible rescaling between \( y \) to high and low frequency channels respectively. The main differences in IARN include channel-splitting in high-resolution and the latent variables. Note that subscripts refer to high- and low-resolution respectively, and superscripts \( hf, lf \) refer to high and low frequency channels respectively.

In summary, the main contributions of our work include:

- A preemptive channel-splitting step, which separates the rescaling-invertible component from the input HR image, is advanced to make the learning of invertible arbitrary rescaling more effective.
- A position-aware scale encoding, which is independent of input image size and compatible with asymmetric rescaling, is proposed to further boost model performance in a large range of arbitrary scales.

2. Related Works

Arbitrary Scale Super-Resolution. Single image super-resolution, as a form of image rescaling with a fixed integer scale factor like \( \times 2 \) and \( \times 4 \), has been studied extensively. For the last few years, deep learning based methods like [6, 10, 15, 28, 27] have brought great successes in this field, but these methods commonly train one model for each scale factor. More recent models proposed by Lim et al. [15] and Li et al. [12] are capable of training one model for multiple scaling factors but only limited to integer ones. Inspired by the weight prediction techniques in meta-learning [11], a single Meta-SR model was proposed by Hu et al. [7] to solve image rescaling of arbitrary scale factors, by predicting weights of convolutional layers for arbitrary scale factors within certain range. Alternatively, Behjati et al. [3] proposed the OverNet to generate over-upscaled maps from which HR images of arbitrary scales can be recovered using optimal downsampling. In the latest ArbSR, Wang et al. [21] proposed a plug-in module that can optimize existing SR models for arbitrary asymmetric rescaling, where scale factors along horizontal and vertical directions could be different. While these methods are often limited to a maximum scale factor like \( \times 4 \) to maintain high performance, Chen et al. [5] proposed recently to use learned pixel representation features to replace pixel value features in previous methods. Using the innovative local implicit image function (LIIF), this model can extrapolate well to out-of-distribution large scales that are not seen in training. Different from the above models which are optimized for upsampling reconstruction only, we consider the bidirectional arbitrary rescaling of downscaling and upsampling as one process in this work.
Bidirectional Image Rescaling. To take advantage of the potential mutual benefits between downsampling and the inverse upscaling, Kim et al. [9] proposed an auto-encoder framework to jointly train image downsampling and upsampling together. Similarly, Sun et al. [19] suggested a new image downscaling method using a content adaptive-resampler, which can be jointly trained with any existing differentiable upsampling (SR) models. More recently, Xiao et al. [24] advanced an invertible rescaling net (IRN) that has achieved SOTA performance for learning based bidirectional image rescaling. Based on the invertible neural network (INN) [2], IRN learns to convert HR input to LR output and an auxiliary latent variable $z$. By mapping $z$ to a case-agnostic normal distribution during training, inverse image upscaling is implemented by randomly sampling $\hat{z}$ from the normal distribution without need of the case specific $z$. While the above bidirectional image rescaling methods are limited to a fixed integer scale factor like $\times 4$, Pan et al. [18] proposed the BAIRNet as the first to solve bidirectional arbitrary rescaling by utilizing local implicit image functions for both downscaling and upsampling with better overall cross-scale performance over IRN. Most recently, Xing et al. [25] proposed an encoder-decoder network (AIDN) to tackle the same challenge with consistent improvements over both IRN and BAIRNet. Instead of using separate encoder and decoder to model image downscaling and upscaling respectively, we propose here to model the downscaling and upsampling processes as forward and backward operations of one INN backbone. While very similar, their performances in the upscaling task are consistently lower than ours with very few exceptions in certain scales. Additionally, LR outputs from our IARN have better perceptual quality comparing to AIDN in both blind or non-blind image quality assessments. As it is shown in [13] that there is a performance trade-off between the generated LR and restored HR, our IARN is obviously the superior one comparing to AIDN when both downscaling and upsampling tasks are considered.

3. Proposed Method

3.1. Network Architecture

Elaborated from the general pipeline shown in Fig. 1, the detailed architecture of the proposed IARN is illustrated in Fig. 2. For the forward downscaling process to generate LR output $y_L$ from HR input $x_H$, the overall process is summarized in the following four steps:

$$
x_{H}^{lf} = s(x_{H})$$
$$x_{H}^{hf} = x_{H} - x_{H}^{lf}$$
$$y_{H}, z_{H}, p_{H}^{lf}, p_{H}^{hf} = f(x_{H}^{lf}, x_{H}^{hf}, p_{H}^{lf}, p_{H}^{hf})$$
$$y_{L} = d(y_{H})$$

where $s(\cdot)$ is the preemptive channel splitting function detailed below, $f(\cdot)$ is the forward function of the INN backbone and $d(\cdot)$ is a downscaling function like nearest neighbour or bicubic interpolation. For variables, $x$ is the input in forward direction, $y$ is the output, and $p$ and $z$ are the position-aware scale encoding at the input and output side respectively. Details about this position-aware scale encoding will be introduced in later subsection. For subscripts and superscripts, $H$ and $L$ means high- and low-resolutions, while $hf$ and $lf$ represents high and low frequency components respectively. Similarly the inverse upsampling process consists of the following three steps:

$$y_{H} = u(y_{L})$$
$$x_{H}^{lf}, x_{H}^{hf}, p_{H}^{lf}, p_{H}^{hf} = f^{-1}(y_{H}, z_{H}, p_{H}^{lf}, p_{H}^{hf})$$
$$x_{H} = x_{H}^{lf} + x_{H}^{hf}$$

where $u(\cdot)$ is an upsampling function and $f^{-1}(\cdot)$ is the inverse function of $f(\cdot)$.

The process presented in Equation 5 is an ideal inverse process, where $y_{H}$ is perfectly restored after the downscaling and upsampling cycle, and $z_{H}, p_{H}^{lf}, p_{H}^{hf}$ are all preserved. This ideal situation does not exist in real application where only $y_{L}$ is saved and $y_{H} \neq y_{H}$. Similar to the previous studies, a generated $\hat{z}_{H}$ is used in place of $z_{H}$. For scale encoding, as they are utilized to carry position and scale factor information for each pixel for both forward and backward directions, and they are independent of low and high-frequency branches, one general $p$ is used for $p_{H}^{lf}$ and $p_{H}^{hf}$ during forward process and the same for $\hat{p}_{H}^{lf}$ and $\hat{p}_{H}^{hf}$ during backward upsampling. As a result, the restored HR output $\hat{x}_{H}$ is calculated as

$$\hat{y}_{H} = u(y_{L})$$
$$\hat{x}_{H}^{lf}, \hat{x}_{H}^{hf} = f^{-1}(\hat{y}_{H}, \hat{z}_{H}, \hat{p}_{H})$$
$$\hat{x}_{H} = \hat{x}_{H}^{lf} + \hat{x}_{H}^{hf}$$

While the primary goal is to minimize the restoration error between $\hat{x}_{H}$ and $x_{H}$, the overall loss used to train the network includes multiple losses for different objectives as shown below

$$L = \lambda_{1} L_{r} + \lambda_{2} L_{g} + \lambda_{3} L_{d} + \lambda_{4} L_{l}.$$ (7)

Here $L_{r}$ is the L1 reconstruction loss for upscaled HR output $\hat{x}_{H}$ and $L_{g}$ is the L2 guidance loss for downscaled LR output $y_{L}$ in reference to a downscaled LR reference $\tilde{y}_{L}$ using bicubic interpolation. For $L_{d}$, it is similar to the distribution regulation loss of latent variable $z$ as in IRN [24]. However, with the newly introduced position-aware scale encoding, in place of the L2 loss between $z_{H}$ and $\hat{z}_{H}$, it is calculated between $(z_{H}, \hat{p}_{H}^{lf}, \hat{p}_{H}^{hf})$ and reference $(z_{H}, \hat{p}_{H}^{lf}, \hat{p}_{H}^{hf})$. Here $z_{H}^{0}$ is a zero tensor, following
the practice in IRN for the surrogate distribution matching loss $L_{dist}$. The last term $L_s$ is a novel rescaling-invertible loss that aims to make the full process effectively invertible. As discussed earlier, the proposed process is only partially invertible since the downsampling and upsampling cycle is not invertible. In general cases, $\hat{y}_H = u(d(y_H)) \neq y_H$. However, given specific pair of $d(\cdot)$ and $u(\cdot)$, there are certain subset of rescaling invertible images $y_s^i$ satisfies $y_s^i_H = u(d(y_s^i_H))$. For example, as explained later in Section 3.2, for NN resampling, we can generate a rescaling-invertible $x_s^i$ from any input image $x_H$. The rescaling-invertible loss $L_s$ is introduced here as the $L_2$ loss between $\hat{y}_H$ and $y_H$. When the loss is zero, then $y_H$ is rescaling invertible.

Note that while the layers of Invertible Blocks (InvBlock) in our INN backbone is similar to the ones in IRN [22], there are some key differences. First, position-aware scale encoding is introduced to both the lower low-frequency (LF) and upper high-frequency (HF) branches to enable the network adaptive to arbitrary scales. Secondly, for transformation functions between the upper and lower branches, including $\phi(\cdot)$, $\rho(\cdot)$ and $\eta(\cdot)$, the original densely connected convolutional block in IRN is enhanced with dilated convolutions where the dilation varies from 1 to 1, which is the number of layers in each block. $l$ is simplified as 2 here in Fig. 2. This enhanced Dense Atrous Block is introduced to increase effective receptive fields of the network with mixed dilation larger or equal to one. As the process in the INN backbone is applied to original high-resolution images, the receptive fields of our IARN is the same as IRN in pixel units, but would be smaller in terms of image regions given the resolution difference if the network structure and depth remains the same. Using a mixture of dilated convolutions in place of the original ones, it can increase the receptive field without changes in the number of model parameters and model complexity.

### 3.2. Preemptive Channel Splitting

As seen in the Fig. 2, the HR output $y_H$ is located in the lower branch of INN backbone and matches the low-frequency HR input $x_s^i_H$ in channel orders. While the transformation functions in the INN backbone can transfer features between lower and upper branches quite efficiently, it may be beneficial to pre-process the inputs so that $x_s^i$ is close to $y_H$ even before feeding into the INN backbone, making the learning task of transformation functions even more efficient. As discussed above, it is ideal for $y_H$ to be rescaling invertible for a given pair of resampling operation $d(\cdot)$ and $u(\cdot)$. Here a preemptive channel splitting function is proposed as $s(\cdot) = u(d(\cdot))$. As pointed out in BAIR-Net [18], in the case of integer scale factors, $s(\cdot)$ would be an idempotent operation when using bilinear downscaling and nearest-neighbour upscaling. That is, for any image $x$, $s(s(x)) = s(x)$. In other words, $s(x)$ is rescaling-invertible. For arbitrary scales here, $s(x)$ is only rescaling invertible when using nearest-neighbour for both downsampling and upsampling. Thus the preemptive channel splitting, which splits the input image to two branches, is proposed here as

$$
\begin{align*}
x_s^i_H &= u_N(d_N(x_H)) \\
x_s^i_H &= x_H - x_s^i_H
\end{align*}
$$

where $N$ refers to nearest-neighbour resampling. Note that the pair of $d(\cdot)$ and $u(\cdot)$ must be the same for both channel splitting in the front and rescaling at the end. Experiments are also conducted to demonstrate the advantage of nearest-neighbour over other choices of resampling choices.
3.3. Position-Aware Scale Encoding

Both the preemptive channel splitting and rescaling invertible loss help the network transform \( y_H \) to induce minimum losses during downscaling and rescaling steps and lead to optimal restoration of \( x_H \) consequently. As this transformation is sensitive to scale factors, to enable the model robust to a large range of arbitrary scales, scale information are necessary during the transformation between the lower and upper branches. Additionally, depending on the scale factor and rescaling method, the position of the pixel inside the image is also needed to determine operations applied to the pixel itself. Using nearest-neighbour as an example, some pixels are resampled from neighbours located to the top-left direction while others resampled from different directions. The naive option of using absolute pixel locations would result in training in a limited image size while testing in larger out-of-distribution image size. To account for these factors, a position-aware and image size independent scale encoding \( p \) is defined as \((s_h, s_v, d_h, d_v)\). Here \( s_h \) and \( s_v \) are the scale factors along horizontal and vertical directions respectively, to accommodate asymmetric scales. For \( d_h \) and \( d_v \), they are the relative horizontal and vertical distances from the input pixel to the closest resampled pixel to its bottom-right direction and can be calculated as below

\[
\begin{align*}
    d_h(i,j) &= \min_{i',i' \leq i} \min_{s_h \leq 0} i's_h - i_s h \\
    d_v(i,j) &= \min_{j',j' \leq j} \min_{s_v \leq 0} j's_v - j_s v
\end{align*}
\]

\tag{9}

where \((i,j)\) and \((i',j')\) are indices of the input pixel and resampled pixel respectively, \( s_h \) and \( s_v \) are pixel sizes along horizontal and vertical axes for input image, and \( s_h' \) and \( s_v' \) are pixel sizes of resampled image. Experiments conducted later also demonstrate that the position-aware scale encoding is greatly beneficial to network performance and it is the best to include it in both lower and upper branches.

4. Experiments

4.1. Data and Settings

For fair comparisons with recent relevant works like IRN and BAIRNet, the DIV2K [1] training set is used for baseline training. Another dataset, Flickr2K [20], is also included for training the final model. For quantitative evaluation, we use HR images from six commonly used datasets for comprehensive comparison, including Set5 [4], Set14 [26], BSD100 [16], Urban100 [8], Manga109 [8], and the DIV2K validation set. Following previous practices, we take the peak noise-signal ratio (PSNR) and SSIM [23] on the luminance channel for all test sets, with the exception of DIV2K which uses average of RGB channels.

To accommodate the large range of arbitrary scale factors, a total of 20 layer of invertible blocks are included in the INN backbone, less than the total of 24 used in IRN×2 and IRN×4. For the Dense Atrous Block, 4 layers of dilated convolution are included with dilation setting as 1 to 4 consecutively. Mini-batches of 16 \( 144 \times 144 \) patches are initially used, each with a random scale sampled from an uniform distribution of \( U(1, 4) \). It is upgraded to 24 \( 192 \times 192 \) patches for the final model when trained using both DIV2K and Flickr2K. There are two stages of training for the final models, each with 250k iterations where the learning rate is reduced by half after each 50k iterations. Settings of the two stages are the same except for the starting learning rate, which is \( 2 \times 10^{-4} \) for the first and changed to \( 1 \times 10^{-4} \) when resuming the second stage training. The weights of losses are set at empirically at 1, 16 and 2 for \( L_r, L_y \) and \( L_i \) respectively. As we have found out in experiments, and similar to the findings by Li et al. [13], setting \( z \) as zero for training and inference achieves equivalent or better performance comparing to randomly sampled \( z \). As a result, \( L_d \) is set as zero for our experiments as it has little impact on learning.

4.2. Arbitrary Rescaling Performance

To assess the performance of our proposed method for arbitrary rescaling in symmetric scales, we compare restoration quality of rescaled HR images in a set of arbitrary scales and the results are included in Table 1. Similar to BAIRNet, models trained for fixed integer scales are evaluated for arbitrary scales using additional bicubic interpolations to keep the resolution of LR images fed into respective models identical across different methods. For models capable of arbitrary scales on their own, most of them are optimized for upscaling only so they are only assessed here for HR outputs. Other than our IARN, BAIRNet is the only one that is trained for bidirectional arbitrary image rescaling.

As shown in Table 1, results from our IARN are the best for all test cases except for a couple of cases where AIDN is slightly better. Using the challenging Manga109 test set as an example, it leads over BAIRNet in PSNR by +3.04, +1.16 and +0.89 for \( \times 1.5, \times 2.5 \) and \( \times 3.5 \) respectively. As shown in visual examples in Fig. 3, bidirectional models like IRN, BAIRNet, AIDN and our IARN outperform others significantly. Among the four, IARN is able to restore details more accurately, making it easier to recognize details like the number 00. The IARN model used for comparisons here uses an asymmetric pre-training for the first stage. For intra-model comparison, two other variants, IARN† and IARN‡, are also included in Table 1 with \( \times 3.5 \) as an example. IARN† takes the same pre-training strategy but only uses DIV2K with baseline batch and patch sizes, resulting a slightly worse performance behind IARN. IARN‡ is the same as IARN† in data settings but it is exclusively trained using symmetric scales for both stages. It is consistently worse than IARN, proving that pre-training
Table 1: Quantitative comparisons of SOTA SR and rescaling methods with the best two results highlighted in red and blue respectively (IARN\textsuperscript{1} and IARN\textsuperscript{†} are used for intra-model comparisons only and not ranked, and methods in bold means requirement of multiple models and additional interpolations to conduct tests in arbitrary scales).

<table>
<thead>
<tr>
<th>Method</th>
<th>Scale</th>
<th>Set5</th>
<th>Set14</th>
<th>BSD100</th>
<th>Urban100</th>
<th>Manga109</th>
<th>DIV2K</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCAN [27]</td>
<td>1.5</td>
<td>40.97/0.9767</td>
<td>37.05/0.9578</td>
<td>35.59/0.9516</td>
<td>35.93/0.9660</td>
<td>42.33/0.9889</td>
<td>38.47/0.9701</td>
</tr>
<tr>
<td>Meta-SR [7]</td>
<td>1.5</td>
<td>41.47/0.9785</td>
<td>37.52/0.9601</td>
<td>35.86/0.9543</td>
<td>36.91/0.9696</td>
<td>43.17/0.9904</td>
<td>38.88/0.9718</td>
</tr>
<tr>
<td>LIIF [5]</td>
<td>1.5</td>
<td>41.23/0.9774</td>
<td>37.37/0.9591</td>
<td>35.76/0.9536</td>
<td>36.70/0.9684</td>
<td>42.84/0.9894</td>
<td>38.82/0.9717</td>
</tr>
<tr>
<td>ArbSR [21]</td>
<td>1.5</td>
<td>41.47/0.9786</td>
<td>37.51/0.9603</td>
<td>35.86/0.9547</td>
<td>36.92/0.9697</td>
<td>43.12/0.9904</td>
<td>38.84/0.9719</td>
</tr>
<tr>
<td>CAR [19]</td>
<td>1.5</td>
<td>40.50/0.9763</td>
<td>37.08/0.9596</td>
<td>35.72/0.9535</td>
<td>34.70/0.9635</td>
<td>40.90/0.9881</td>
<td>37.93/0.9683</td>
</tr>
<tr>
<td>IRN [24]</td>
<td>1.5</td>
<td>43.55/0.9891</td>
<td>39.52/0.9795</td>
<td>39.28/0.9833</td>
<td>36.52/0.9811</td>
<td>42.64/0.9936</td>
<td>40.18/0.9838</td>
</tr>
<tr>
<td>BAIRNet [18]</td>
<td>1.5</td>
<td>47.13/0.9849</td>
<td>43.12/0.9760</td>
<td>46.63/0.9959</td>
<td>44.01/0.9946</td>
<td>45.49/0.9948</td>
<td>44.99/0.9920</td>
</tr>
<tr>
<td>AIDN [25]</td>
<td>1.5</td>
<td>50.61/0.9961</td>
<td>46.70/0.9920</td>
<td>49.82/0.9983</td>
<td>46.26/0.9967</td>
<td>47.01/0.9953</td>
<td>47.01/0.9953</td>
</tr>
<tr>
<td>IARN\textsuperscript{†}</td>
<td>1.5</td>
<td>51.02/0.9968</td>
<td>47.25/0.9938</td>
<td>50.91/0.9986</td>
<td>47.58/0.9975</td>
<td>48.58/0.9975</td>
<td>46.74/0.9949</td>
</tr>
<tr>
<td>RCAN [27]</td>
<td>2.5</td>
<td>36.05/0.9436</td>
<td>31.69/0.8815</td>
<td>30.47/0.8508</td>
<td>30.42/0.8990</td>
<td>36.59/0.9634</td>
<td>32.72/0.9079</td>
</tr>
<tr>
<td>Meta-SR [7]</td>
<td>2.5</td>
<td>36.18/0.9441</td>
<td>31.90/0.8814</td>
<td>30.47/0.8508</td>
<td>30.57/0.9003</td>
<td>36.55/0.9639</td>
<td>32.77/0.9086</td>
</tr>
<tr>
<td>LIIF [5]</td>
<td>2.5</td>
<td>35.98/0.9434</td>
<td>31.64/0.8813</td>
<td>30.45/0.8510</td>
<td>30.42/0.8992</td>
<td>36.39/0.9630</td>
<td>32.78/0.9091</td>
</tr>
<tr>
<td>ArbSR [21]</td>
<td>2.5</td>
<td>36.21/0.9448</td>
<td>31.99/0.8830</td>
<td>30.51/0.8536</td>
<td>30.68/0.9027</td>
<td>36.67/0.9646</td>
<td>32.77/0.9093</td>
</tr>
<tr>
<td>CAR [19]</td>
<td>2.5</td>
<td>37.33/0.9548</td>
<td>33.78/0.9169</td>
<td>32.53/0.9020</td>
<td>32.19/0.9301</td>
<td>37.63/0.9717</td>
<td>34.32/0.9310</td>
</tr>
<tr>
<td>IRN [24]</td>
<td>2.5</td>
<td>39.78/0.9742</td>
<td>36.39/0.9553</td>
<td>35.56/0.9540</td>
<td>33.99/0.9589</td>
<td>39.33/0.9836</td>
<td>36.60/0.9607</td>
</tr>
<tr>
<td>BAIRNet [18]</td>
<td>2.5</td>
<td>40.11/0.9664</td>
<td>36.62/0.9469</td>
<td>36.29/0.9563</td>
<td>36.62/0.9679</td>
<td>40.26/0.9830</td>
<td>37.46/0.9627</td>
</tr>
<tr>
<td>AIDN [25]</td>
<td>2.5</td>
<td>40.77/0.9750</td>
<td>36.72/0.9588</td>
<td>36.65/0.9593</td>
<td>37.10/0.9710</td>
<td>37.88/0.9659</td>
<td>37.88/0.9659</td>
</tr>
<tr>
<td>IARN\textsuperscript{†}</td>
<td>2.5</td>
<td>40.93/0.9756</td>
<td>37.78/0.9598</td>
<td>36.81/0.9607</td>
<td>36.95/0.9703</td>
<td>41.38/0.9862</td>
<td>37.92/0.9662</td>
</tr>
</tbody>
</table>

Figure 3: Visual examples of arbitrary rescaling at two scales: ×2.5 and ×3.5 (Zoom in to see improved sharpness in IARN results comparing to AIDN, including the area marked by red arrows).

with asymmetric scales can help the final model performs better in tests of symmetric scales.

Results of continuous scales between ×1.1 and ×4 (increasing by 0.1 at each sampling point) are also illustrated in Fig. 5 for a more thorough assessment. As explained in BAILNet, there is a significant performance boost when the models are trained jointly for both downscaling and upscaling. Among this group, BAILNet is obviously better than CAR and IRN across the range overall, except trailing slightly behind IRN for a limited choices of scales. Both AIDN and our IARN model are able to outperform other models across the full range of arbitrary scales, surpassing IRN even at fixed scales like ×2 and ×4. Between the two,
4.3. LR Image Quality

For the group of six bidirectional rescaling methods, LR outputs are also generated from the learning based models respectively. Without a known ground-truth reference, the goal is to generate visually pleasant images which look similar to the LR reference that is downsampled from the HR input using conventional bicubic interpolation. As shown in Table 2, the SSIM calculated in relation to the LR references and the blind image quality metric NIQE [17] are used for comparison. To accommodate the native integer scales in CAR and IRN, ×2 and ×4 are used for assessments. It is clearly shown that our IARN is the best overall, slightly ahead of IRN as the lose second best, while AIDN is trailing behind as the third best. An example of ×4 LR images with the most visible artifacts are included in Fig. 4 to compare different models. While they are not much different from the bicubic reference in overall view, the CAR one is noticeably brighter. From zoomed-in views on the first row, HCFlow and BAIRNet are obviously the worst two with obvious false color artifacts. While the remaining three are similar visually, the artifacts are slightly more noticeable in AIDN, as marked by red arrows in the left zoomed-in window in particular.

Table 3: Ablation study for different modules (Color-highlighted group in one column means they share the same setting in all other columns).

<table>
<thead>
<tr>
<th>Channel</th>
<th>Scale</th>
<th>Atrous</th>
<th>Rescaling #</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Splitting Encoding Convolution Method</td>
<td>InvBlock</td>
<td>Urban100</td>
<td>×4</td>
<td></td>
</tr>
<tr>
<td>✗ Dual ✗ Bilinear</td>
<td>16</td>
<td>26.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✔ Dual ✗ Bilinear</td>
<td>16</td>
<td>27.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✔ HF ✗ Bilinear</td>
<td>16</td>
<td>26.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✔ LF ✗ Bilinear</td>
<td>16</td>
<td>27.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✗ None ✗ Bilinear</td>
<td>16</td>
<td>26.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✗ Dual ✔ Bilinear</td>
<td>16</td>
<td>27.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✗ Dual ✔ NN</td>
<td>16</td>
<td>27.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✗ Dual ✔ Bicubic</td>
<td>16</td>
<td>27.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✗ Dual ✔ NN</td>
<td>20</td>
<td>27.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✗ Dual ✔ NN</td>
<td>24</td>
<td>27.79</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.4. Ablation Study

To show the effects of different modules proposed here, we conduct a comprehensive ablation study for different combinations using Urban100 ×4 as the testing benchmark. For illustrative purpose, all models were trained for 100k iterations and the differences in performance is clearly shown in Table 3. For easy cross-reference, colored highlight are used to associate mini-groups for comparison. Two of the
most critical modules to boost performance are preemptive channel splitting and position-aware scale encoding. As highlighted in red, when channel-splitting is activated while keeping other configurations identical, the average PSNR increases by +0.43. Note that when no channel-splitting is applied, original RGB channels are fed to the LF branch while inputs to HR are zeros. For scale encoding options as highlighted in blue, it is the best to include it for both branches, boosting PSNR by +0.52. When atrous convolution is applied, another +0.19 is added in PSNR. For rescaling, NN is the best choice while bicubic is far behind the other two. This corroborates the methodology discussed earlier that channel-split function $x(\cdot)$ makes $x/2$ rescaling-invertible when using NN resampling. While the final PSNR is always better when the number of InvBlock layers increases, it is set at 20 in other experiments for better efficiency with minimum performance loss.

Table 4: PSNR for symmetric and asymmetric scale factors.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scale</th>
<th>Param</th>
<th>Downscaling</th>
<th>Upscaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArbSR [21]</td>
<td>$x \times 3$</td>
<td>30.58</td>
<td>51.1M</td>
<td>0.004s</td>
</tr>
<tr>
<td>IRN [24]</td>
<td>$x \times 3$</td>
<td>34.87</td>
<td>52.8M</td>
<td>0.004s</td>
</tr>
<tr>
<td>BAIRNet [18]</td>
<td>$x \times 3$</td>
<td>36.64</td>
<td>54.23</td>
<td>0.018s</td>
</tr>
<tr>
<td>IARN*</td>
<td>$x \times 3$</td>
<td>36.96</td>
<td>54.23</td>
<td>0.018s</td>
</tr>
</tbody>
</table>

4.5. Asymmetric Arbitrary Rescaling

As discussed earlier, our IARN model can be optimized for asymmetric scales using the exact same network architecture. For fair comparisons with ArbSR and BAIRNet, another model denoted as IARN* is trained using asymmetric scale in both stages. As shown in Table 4, two challenging large benchmark test sets are used for assessment in three asymmetric and one symmetric scales. Results from ArbSR, IRN and BAIRNet are used for comparisons, using the same evaluation protocol as in BAIRNet. It is demonstrated that, both in quantitative metrics and visual examples as included in Fig. 6, our IARN* holds a clear advantage over other peers in restoring more accurate details.

4.6. Model Complexity and Efficiency

To assess the efficiency of our proposed IARN, it is compared with others in terms of model size and average inference time per image for the BSD100 test set. As illustrated in Table 5, three most relevant models which are optimized jointly for downscaling and upscaling are included so the inference time includes both downscaling and upscaling and listed separately. Note that for CAR and IRN, they have models of different sizes depending on the scale factor, while for BAIRNet and our IARN, only one model is needed for both tests. It is shown that our IARN is comparable to IRN $\times 4$ in terms of number of parameters, much more smaller than CAR and BAIRNet. For inference speed, it is slower than CAR and IRN. But for the only two models capable of bidirectional arbitrary rescaling (AIDN not yet available for assessment), our model is able to outperform the previous SOTA BAIRNet consistently while using less than 20% parameters and reducing inference time by 66% and 60% for $x \times 2$ and $x \times 4$ respectively.

Table 5: Model size and inference time comparison.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scale</th>
<th>Param</th>
<th>Downscaling</th>
<th>Upscaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAR [19]</td>
<td>$x \times 2$</td>
<td>51.1M</td>
<td>0.004s</td>
<td>0.005s</td>
</tr>
<tr>
<td>IRN [24]</td>
<td>$x \times 4$</td>
<td>52.8M</td>
<td>0.004s</td>
<td>0.005s</td>
</tr>
<tr>
<td>BAIRNet [18]</td>
<td>$x \times 4$</td>
<td>22.4M</td>
<td>0.506s</td>
<td>0.129s</td>
</tr>
<tr>
<td>IARN</td>
<td>$x \times 4$</td>
<td>4.32M</td>
<td>0.058s</td>
<td>0.154s</td>
</tr>
</tbody>
</table>

5. Conclusions

Image arbitrary rescaling using deep learning is a relatively new and under explored topic in the field of low-level image processing due to its complexity. In this paper, we have presented the first invertible arbitrary image rescaling work. Based on an INN backbone enhanced with a novel preemptive channel splitting module and a new position-aware scale encoding method, the newly proposed IARN network is capable of handling bidirectional image arbitrary rescaling over a large range of scales using just one trained model. Extensive experiments on a comprehensive set of benchmark datasets validate a much better performance of arbitrary image rescaling over the current related SOTA methods in both HR and LR outputs, with reduced model size and faster inference comparing to BAIRNet. In addition, better performance is also achieved for asymmetric arbitrary rescaling tests.
References


[18] Zhihong Pan, Baopu Li, Dongliang He, Mingde Yao, Wen-hao Wu, Tianwei Lin, Xin Li, and Errui Ding. Towards bidirectional arbitrary image rescaling: Joint optimization and cycle idempotence. pages 17389–17398, 2022.


[26] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In Interna-