Scale-arbitrary Invertible Image Downscaling

JINBO XING†, The Chinese University of Hong Kong, Hong Kong SAR, China
WENBO HU†, The Chinese University of Hong Kong, Hong Kong SAR, China
TIEN-TSIN WONG†, The Chinese University of Hong Kong, Hong Kong SAR, China

Conventional social media platforms usually downscale the HR images to restrict their resolution to a specific size for saving transmission/storage cost, which leads to the super-resolution (SR) being highly ill-posed. Recent invertible image downscaling methods jointly model the downscaling/upsampling problems and achieve significant improvements. However, they only consider fixed integer scale factors that cannot downscale HR images with various resolutions to meet the resolution restriction of social media platforms. In this paper, we propose a Scale-Arbitrary Invertible Image Downscaling Network (AIDN), to natively downscale HR images with arbitrary scale factors. Meanwhile, the HR information is embedded in the downsampled low-resolution (LR) counterparts in a nearly imperceptible form such that our AIDN can also restore the original HR images solely from the LR images. The key to supporting arbitrary scale factors is our proposed Conditional Resampling Module (CRM) that conditions the downscaling/upsampling kernels and sampling locations on both scale factors and image content. Extensive experimental results demonstrate that our AIDN achieves top performance for invertible downscaling with both arbitrary integer and non-integer scale factors. Code will be released upon publication.

CCS Concepts: • Computing methodology → Computational photography.

Additional Key Words and Phrases: Invertible image rescaling, machine learning, image reconstruction

Reference Format:
Jinbo Xing, Wenbo Hu, and Tien-Tsin Wong. 2022. Scale-arbitrary Invertible Image Downscaling. arXiv preprint arXiv:2201.12576.

1 INTRODUCTION

With the rapid development of smartphone cameras, exploding amount of high-resolution (HR) images/videos are produced in our daily life. For the transmission/storage cost concern, most social media platforms (e.g., WhatsApp, Messenger, WeChat Moments and etc.) will downscale the HR images to restrict their resolution to a specific size when users distribute them over the platform (Figure 1 (a)). Therefore, image upsampling/super-resolution (SR) is indispensable when receivers want to explore the details of the distributed images. But the lost information during the downscaling process makes the SR problem highly ill-posed [Dong et al. 2015; Glasner et al. 2009; Yang et al. 2010]. To relieve this problem, recent works [Kim et al. 2018; Sun and Chen 2020; Xiao et al. 2020] regard the downscaling/upsampling as a dual problem and jointly optimize them. By doing so, the information from the HR images can be better preserved or embedded during the downscaling process, which leads to a higher restoration performance for the upsampling counterpart.

However, existing works only consider a fixed integer scale factor, e.g., ×2, ×3, or ×4. In real-world scenarios, the scale factor should be arbitrary, since the social media platform will downscale images with various resolutions to meet the resolution restriction (Figure 1 (a)). Although with certain pre-/post-processing operations (including the Bicubic interpolation [Mitchell and Netravali 1988]) on the input/output images from multiple trained scale-fixed invertible models [Sun and Chen 2020; Xiao et al. 2020], one can indirectly and virtually achieve the arbitrary scale factor, this also means numerous trained models have to be stored in place. Moreover, the experiment shows such an indirect solution cannot produce satisfactory results (Figure 2), and its performance fluctuates across the scale factors.

In this paper, we propose a universal encoder-decoder styled Scale-Arbitrary Invertible Image Downscaling Network (AIDN), to downscale/upscale images with arbitrary scale factors. The encoder network is designed for the downscaling problem while the decoder network is for the upsampling task, and the encoder-decoder is trained jointly to address their duality. As shown in Figure 1 (b), in the sending-end, the encoder of our AIDN can natively downscale HR images using arbitrary required scale factors to meet the resolution restriction of a particular social media platform, and embed the HR information to the LR image in a nearly imperceptible form;

†Equal contribution
†Corresponding author

Authors’ addresses: Jinbo Xing, The Chinese University of Hong Kong, Hong Kong SAR, China; Wenbo Hu, The Chinese University of Hong Kong, Hong Kong SAR, China; Tien-Tsin Wong, The Chinese University of Hong Kong, Hong Kong SAR, China; [jbxing,wbhu,ttwong]@cse.cuhk.edu.hk.

Fig. 1. Part (a) shows the conventional pipeline of distributing HR images over the social media platform, i.e., WhatsApp. In the sending-end, the client will downscale HR images to make the height/width of the images equal to or smaller than 1600 pixels if the resolution of them is too large. And in the receiving-end, image super-resolution (SR) will be employed if users want to explore the details of the distributed images. Part (b) shows the distributing pipeline with our proposed AIDN. In the sending-end, before feeding into the WhatsApp client, the HR images are first downscaled by our AIDN encoder to meet the resolution restriction of the WhatsApp, thus, WhatsApp would not downscale the images anymore. And in the receiving-end, the AIDN decoder can faithfully restore the original HR image whenever necessary.

The resolution restriction of WhatsApp is determined by experiments, and other social media platforms (e.g., Messenger, WeChat Moments and etc.) share similar rules.
while in the receiving-end, the decoder of our AIDN can consume the embedded information to faithfully restore the original HR image, solely from the LR counterpart, whenever receivers want to explore the details of the distributed images. The key to supporting arbitrary scale factors is our proposed Conditional Resampling Module (CRM). It can dynamically resample the feature map to target resolution by parameterizing the downsampling/upsampling kernels and sampling locations conditioned on both the target scale factor and the image content. The content-adaptive characteristic allows the resampling operation to adapt to the textural/structural content at both training and inference time, so that we can produce visually pleasant results. Our proposed CRM can be easily applied to multiple existing backbones in the encoder-decoder network to enable the scale-arbitrary invertible image downscaling.

We evaluated our AIDN on multiple public datasets, both quantitatively and qualitatively. Experiment results show our AIDN achieves top performance for both integer (e.g., ×2, ×3 and ×4) and arbitrary non-integer scale factors (e.g., ×1.60, ×2.75 and ×3.20). Moreover, the performance changes smoothly (i.e., more predictable), without fluctuation, across the scale factors, as shown in Figure 2. Also, the visualization of routing weights confirms our CRM can dynamically adjust resampling kernels for various scale factors and different image content (Section 4.5). Our contributions are summarized below.

- To the best of our knowledge, this is the first attempt to tackle the scale-arbitrary invertible image downscaling problem with a single encoder-decoder network.
- We propose a Conditional Resampling Module (CRM) to dynamically resample feature maps to the target resolution according to both the required scale factor and the image content. It can be easily applied to existing SR backbones to achieve scale-arbitrary invertible image downscaling.
- Extensive experiments demonstrate our AIDN achieves top performance for invertible downscaling with both arbitrary integer and non-integer scale factors. Moreover, the amount of parameters of our AIDN is significantly reduced (~1/10) compared to conventional SR networks, which indicates the high efficiency of our simple yet effective system design.

2 RELATED WORK

2.1 Image Rescaling

Image downscaling and upscaling/super-resolution (SR) have been studied with a long history. Downscaling aims at reducing the resolution of images. It can be performed by resampling together with interpolation, e.g., the Bilinear and Bicubic [Mitchell and Netravali 1988] interpolation, which enjoys high efficiency but usually incurs visual artifacts, such as aliasing, ringing, blurring, etc. To tackle these problems, several detail-preserved and perceptual-quality-oriented approaches [Kopf et al. 2013; Liu et al. 2017; Oeztireli and Gross 2015; Weber et al. 2016] are proposed. On the other hand, SR tries to restore the high-resolution (HR) image from its low-resolution (LR) counterpart. Thanks to deep learning, we have witnessed significant progress [Dai et al. 2019; Dong et al. 2014; Haris et al. 2018; Jo and Kim 2021; Kim et al. 2016; Liang et al. 2021; Niu et al. 2020; Wang et al. 2021; Zhang et al. 2018a,b; Zhong et al. 2018] on the SR problem. Recently, several works propose to learn a single network for scale-arbitrary SR, using meta-learning [Hu et al. 2019], local implicit image function [Chen et al. 2021], and scale-aware upsampling [Wang et al. 2021].

The downscaling and the upscaling act as a dual problem in nature, however, all the above works independently model them. Thus, they may be sub-optimal when working together, which is a common scenario when distributing HR images over social media platforms. Differently, we model the downscaling and upscaling problems as a universal encoder-decoder network and optimize them jointly. By doing so, the downscaling and the upscaling can mutually reinforce each other to produce more visually satisfying results for both downsampled and upscaled images.

2.2 Invertible Image Conversion

The goal of invertible image conversion is to build an invertible transformation between certain visual content and an embedding image, where the original content can be restored from the embedding image whenever necessary [Cheng et al. 2021]. As a pioneer, Xia et al. [Xia et al. 2018] propose an auto-encoder-style network to encode the color information into the generated grayscale image, from which the original color image can be decoded back. Furthermore, several work the invertible transformation between color and halftone images [Xia et al. 2021], binocular and monocular videos [Hu et al. 2020], short videos and key-frame [Zhu et al. 2020], multiview images and JPEG image [Wu et al. 2021], camera raw data and sRGB image [Xing et al. 2021]. And recently, Cheng et al. [Cheng et al. 2021] present a generic invertible neural network (INN) [Behrmann et al. 2019; Dinh et al. 2015, 2017; Kingma and Dhariwal 2018] based framework for multiple invertible image conversion problems.

Image downscaling/upsampling can also be formulated as invertible image conversion, Kim et al. [Kim et al. 2018] present a task-aware image downscaling (TAD) method to jointly optimize the downscaling and upsampling networks as a unified task. Li et al. [Li et al. 2018] propose to use a CNN to estimate compact-resolution images (CNN-CR), and then leverage a specified or learned SR method to restore the HR images; Recently, Sun et al. [Sun and Chen 2020]
present a learnable image downsampling method based on content-adaptive resampler (CAR) that can be jointly trained with existing SR networks, and IRN [Xiao et al. 2020] adopts the INN to model the invertible image rescaling task as a bijective mapping from HR image to LR image while capturing the distribution of lost information using a latent variable. Although these works demonstrate the effectiveness of invertible image downsampling, they can only downscale/upscale images with fixed integer scale factors, e.g., ×2 and ×4. Supporting arbitrary scale factors is crucial in real-world scenarios, as shown in Figure 1. Different from the above methods, our method can downscale images with arbitrary scale factors and faithfully recover the HR images.

3 METHODOLOGY

3.1 Overview

Overall, as shown in Figure 3, our Scale-Arbitrary Invertible Image Downscaling Network (AIDN) contains an encoder \( E_\theta \) and a decoder \( D_\phi \), where \( \theta \) and \( \phi \) denote their parameters, respectively. Given a high-resolution (HR) image to be shared through social media platforms, \( I_{HR} \), and the required arbitrary scale factor to meet the resolution restriction of social media platforms, \( s \in \{1, 4\} \), the encoder network \( E_\theta \) can downscale the \( I_{HR} \) to produce a low-resolution (LR) image, \( I_{LR} \), which looks like the reference LR image \( I_{LR} \) (say, the bicubic-downsampled [Mitchell and Netravali 1988] image, without loss of generality). The produced LR image \( I_{LR} \) has the same 8-bit precision as conventional images for compatibility with current platforms. And importantly, the decoder network \( D_\phi \) can accurately restore the HR image \( I_{HR} \) solely from the LR image \( I_{LR} \), whenever users want to explore the details of the distributed image.

We design a universal encoder-decoder network for downscaling/upsampling with arbitrary scale factors. To leverage the duality of the downsampling and the upsampling, we train the encoder and decoder jointly with the supervision on both \( I_{LR} \) and \( I_{HR} \). By doing so, the information from HR images can be embedded in the \( I_{LR} \) and the decoder network can consume it to restore the high-quality \( I_{HR} \).

3.2 Network Architecture

As shown in Figure 3, the encoder and decoder sub-networks in our AIDN share a similar structure, where the encoder network contains the feature extractor, the Conditional Resampling Module (CRM), and the quantization layer; and decoder network consists of the feature extractor and the CRM.

**Feature extractor.** Considering the duality of the downsampling and the upsampling, we adopt the same architecture for feature extractors in the encoder \( E_\theta \) and decoder \( D_\phi \) networks. As EDSR [Lim et al. 2017] shows powerful capability for feature extraction in the SR task, we adopt a similar network structure for the feature extractor. For the efficiency concern, we employ the EDSR-baseline structure as the backbone block, thus, the amount of parameters of our whole network is only 3.8M, as shown in Table 1. It consists of a convolutional layer as the head and a series of residual blocks (Conv+ReLU+Conv) [He et al. 2016] as the backbone structure to learn the residual components. Note that many other types of backbone blocks are also applicable to our framework, e.g., RDN [Zhang et al. 2018b] and RCAN [Zhang et al. 2018a]. To better extract scale-adaptive features for our goal, invertible image downsampling with arbitrary scale factors, we also equip the backbone block with the scale-aware feature adaption module [Wang et al. 2021] that takes the feature map from the previous layer and the required arbitrary scale factor as input. More details about the feature extractor can be found in the Supplementary Material.

**Conditional resampling module (CRM).** After extracting features from \( I_{HR} \) or \( I_{LR} \), we need to resample the feature map to the target resolution with arbitrary scale factors, \( s \in \{1, 4\} \). Previous invertible image downsampling methods [Kim et al. 2018; Sun and Chen 2020; Xiao et al. 2020] adopt either PixelShuffle [Shi et al. 2016] or Haar transformation as the resampling module, which, however, inherently only serves a fixed integer scale factor \( r \), i.e., \( r \in \{2, 3, 4, \ldots\} \) for the PixelShuffle and \( r \in \{2, 4, 8, \ldots\} \) for the Haar transformation.

Inspired by the conditional convolution [Chen et al. 2020; Tian et al. 2020; Yang et al. 2019; Zhang et al. 2020] that conditions convolutional kernels on instances for visual recognition, we propose a Conditional Resampling Module (CRM) to dynamically downsample/upsample feature maps to the target resolution by parameterizing the resampling kernels and sampling locations conditioned on both the required scale factor and the image content.
Wang et al. [Wang et al. 2021] also presents a scale-aware upsampling layer used for SR networks to support arbitrary scale factors. Unlike the scale-aware upsampling layer, our CRM can be used for both downscaling and upscaling, and is not only scale-aware but also content-adaptive. The content-adaptive characteristic allows our method better to deal with various textual/structural patterns at both training and inference time, so that our method can produce visually satisfying results, as to be demonstrated in Sec. 4.

More concretely, as shown in Figure 4, given the input feature map $F^\text{in}$ and the required arbitrary scale factor $s$, our CRM can dynamically resample $F^\text{in}$ according to scale factor and image content, for producing the output feature $F^\text{out}$.

Fig. 4. Conditional Resampling Module (CRM). Given the input feature $F^\text{in}$ and required arbitrary scale factor $s$, our CRM can dynamically resample $F^\text{in}$ according to scale factor and image content, for producing the output feature $F^\text{out}$.

3.3 Loss Function
Following Xiao et al. [Xiao et al. 2020], we adopt two loss terms to drive the network training:

$$
\mathcal{L}_G(\theta) = \mathbb{E}_{\hat{I}_{\text{HR}} \in I} \left\{ \| f(\hat{I}_{\text{HR}}) - \hat{I}_{\text{HR}} \|_2^2 \right\}
$$

(5)

$$
\mathcal{L}_I(\theta, \phi) = \mathbb{E}_{\hat{I}_{\text{HR}} \in I} \left\{ \| I_{\text{HR}} - \hat{I}_{\text{HR}} \|_1 \right\}
$$

(6)

where the guidance loss $\mathcal{L}_G(\theta)$ is to supervise the produced down-scaled image $\hat{I}_{\text{LR}}$ to be similar with the conventional LR image that is generated by the Bicubic interpolation $f(\cdot)$ from the original HR image $I_{\text{HR}}$; and the invertibility loss $\mathcal{L}_I(\theta, \phi)$ is to constrain the reconstructed HR image $\hat{I}_{\text{LR}}$ to match $I_{\text{HR}}$. Here $\mathbb{E}$ denotes the average operator over all images in training dataset $I$; $\theta$ and $\phi$ denote the parameters of encoder and decoder, respectively. Note that $\mathcal{L}_I(\theta, \phi)$ effectively imposes constraints over the parameters of both the encoder and decoder, since they are jointly trained. Altogether, we optimize the proposed AIDN by minimizing the total loss $\mathcal{L}(\theta, \phi)$:

$$
\mathcal{L}(\theta, \phi) = \lambda \mathcal{L}_G(\theta) + \mathcal{L}_I(\theta, \phi),
$$

(7)

where $\lambda$ is a weight for balancing the two terms and set to be 1.0 in our experiments.

4 EXPERIMENTS
4.1 Dataset and Settings
We employed the DIV2K dataset [Agustsson and Timofte 2017] to train our model, which contains 900 high-quality 2K resolution images. We followed the official training and validation set splits. Besides, we evaluated our AIDN on additional four benchmark datasets, i.e., Set5 [Bevilacqua et al. 2012], Set14 [Zeyde et al. 2010], B100 [Martin et al. 2001] and Urban100 [Huang et al. 2015].
Table 1. Quantitative results (PSNR) of the reconstructed HR images produced by multiple methods with various scale factors on the five benchmark datasets. Methods marked with *, †, and ‡ are the type I, II, and III methods explained in Sec. 4.2, respectively. The best and second-best results are marked in bold and underline, respectively.

| Method | Downscaling+Upscaling | Param. (M) | Set5 $\times 2$ | Set5 $\times 1.6$ | Set14 $\times 2$ | Set14 $\times 1.65$ | B100 $\times 2$ | Urban100 $\times 2$ | DIV2K $\times 2$ |
|--------|-----------------------|------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Bicubic + Bicubic | - | 33.66 | 31.06 | 30.24 | 31.83 | 29.56 | 32.95 | 26.88 | 27.05 | 31.01 | 32.46 |
| Bicubic + EDSR-$\times 3$ [Lim et al. 2017] * | 40.7 | 38.19 | 40.39 | 33.95 | 35.95 | 32.36 | 36.79 | 32.95 | 32.69 | 35.03 | 36.95 |
| Bicubic + ArbEDSR [Wang et al. 2021] † | 39.2 | 38.19 | 40.64 | 34.05 | 36.22 | 32.37 | 36.92 | 33.02 | 33.30 | - | - |
| TAD + TAU(-$\times 2$) [Kim et al. 2018] ‡ | - | 38.46 | - | 35.52 | - | 36.68 | - | 35.03 | - | 39.01 | - |
| CNN-CR + CNN-SR(-$\times 3$) [Li et al. 2018] ‡ | - | 38.88 | - | 35.40 | - | 33.92 | - | 33.68 | - | - | - |
| CAR + EDSR(-$\times 2$) [Sun and Chen 2020] ‡ | 51.1 | 38.94 | 40.09 | 35.61 | 36.45 | 33.83 | 36.41 | 35.24 | 35.28 | 38.26 | 34.29 |
| IRN-$\times 2$ [Xiao et al. 2020] ‡ | 1.7 | 43.99 | 43.42 | 40.79 | 39.24 | 41.32 | 39.63 | 39.92 | 35.28 | 44.32 | 42.00 |
| AIDN (Ours) | 3.8 | 44.13 | 48.81 | 40.81 | 44.25 | 40.72 | 52.11 | 40.28 | 39.27 | 44.12 | 47.49 |

During training, we adopted scale factors varying from 1.0 to 4.0 with a stride of 0.1, i.e., $S = \{1.1, 1.2, ..., 4\}$. To address the varying difficulties of different scale factors, we randomly sampled $s$ with probability of $\frac{s^2}{\sum s^2}$ from $S$. We leave other training details in the Supplementary Material, due to the space constraint.

4.2 Evaluation on Reconstructed HR Images

First, we evaluated the quality of the reconstructed HR images $I_{HR}$ for our AIDN and compared it with three types of state-of-the-art solutions, including:

I. downscaling by the Bicubic interpolation and upscaling by the scale-fixed SR methods, i.e., Bicubic+EDSR-$\times 2$/3/4× [Lim et al. 2017], and further bicubic-downscaling the output to the target resolution with required non-integer scale factors;

II. downscaling by the Bicubic interpolation and upscaling by the scale-arbitrary SR method, i.e., Bicubic+ArbEDSR [Wang et al. 2021]; and

III. scaled-fixed invertible image downscaling models, i.e., TAD+TAU [Kim et al. 2018], CNN-CR+CNN-SR [Li et al. 2018], CAR+EDSR [Sun and Chen 2020] and IRN [Xiao et al. 2020].

To make the type III methods support non-integer scale factors $s$, we first bicubicly upscale $I_{HR}$ by $\frac{s}{2}$ or $\frac{1}{2}$ times for $1.0 < s < 2.0$ and $2.0 < s < 4.0$, respectively; and then feed it to the corresponding $\times 2$/$\times 4$ invertible downscaling models; finally bicubicly downscale the output image to the target resolution with scale factor $\frac{s}{2}$ or $\frac{1}{2}$. Note that, this solution is not memory- and computation-efficient since it increases the resolution of images before feeding into the network and decreases the resolution after getting the network's output. Differently, our AIDN can natively support arbitrary scale factors, without any pre- or post-processing.

Quantitative results. Table 1 shows the PSNR of reconstructed HR images $I_{HR}$ produced by multiple methods with various scale factors on five benchmark datasets. Our AIDN significantly outperforms the type I and II methods, which confirms that jointly modeling the downscaling and upscaling process is beneficial. For the type III methods that jointly optimize the downscaling and upscaling.
our AIDN still outperforms them by a large margin for non-integer scale factors on all the five benchmark datasets, while achieving comparable results for integer scale factors. Note that, results of the type III methods for different scale factors are produced by multiple trained models, e.g., IRN-×2 and IRN-×4, while results of our AIDN are generated by a universal model. More importantly, as shown in Figure 2, the performance of our AIDN changes much more smoothly across scale factors, compared with the type III methods. This demonstrates the effectiveness of our CRM for invertible image downscaling with arbitrary scale factors. Besides, the number of parameters of our AIDN is very small compared with other methods, which indicates the efficiency.

**Qualitative results.** To further qualitatively evaluate the reconstructed HR images \(I_{HR}\) for non-integer scale factors, we compared our AIDN with the conventional Bicubic interpolation (BI+BI) and state-of-the-art scale-fixed invertible downscaling method, IRN [Xiao et al. 2020], in Figure 5. Although the results of IRN were produced by multiple models, we can see results of our AIDN still have better perceptual quality and fewer artifacts, e.g., our method faithfully recovers both the letters and the flat region in the ‘ppt3’ case, while BI+BI suffers from blurriness for the letters and IRN incurs color distortion in the flat region; and in the ‘img092’ case, our method produces more accurate structure patterns than both BI+BI and IRN. The high fidelity of our results for various scale factors demonstrates the success of information embedding and the CRM.

### 4.3 Evaluation on Downscaled LR Images

Then, we evaluated the quality of downscaled LR images \(I_{LR}\) by measuring the SSIM between them and bicubic-downscaled images. And we compared the results of IRN and our AIDN in Table 2; we can see both of them have extremely high SSIM values. Besides, the

**Table 2.** SSIM between the bicubic-downscaled images and the results produced by IRN and our AIDN on the B100 [Martin et al. 2001] dataset.

| Method | \(\times 1.6\) | \(\times 2.1\) | \(\times 2.6\) | \(\times 3.1\) | \(\times 3.6\) | \(\times 3.9\) |
|--------|----------------|----------------|----------------|----------------|----------------|----------------|
| IRN    | 0.9963         | 0.9949         | 0.9942         | 0.9936         | 0.9932         | 0.9930         |
| Ours   | 0.9951         | 0.9937         | 0.9924         | 0.9915         | 0.9909         | 0.9907         |

Fig. 6. Visual comparisons of the downscaled image \(I_{LR}\) for the ‘ppt3’ image in the Set14 [Zeyde et al. 2010] with the scale factor of \(\times 3.9\).

qualitative comparison in Figure 6 shows the perceptual quality of our results is even better than IRN, i.e., IRN incurs color distortion in the flat region while the result of our AIDN is free of the distortion. It indicates our downscaled images are almost the same as the conventional bicubic-downscaled ones.

### 4.4 Ablation Study

To verify the effectiveness of some key designs in our AIDN, we conducted ablation experiments on the Set5 [Bevilacqua et al. 2012] dataset by considering the following methods:

- **Bicubic:** down-/up-scaling with the Bicubic interpolation;
we can see the AIDN

Table 3. PSNR/SSIM results of the reconstructed HR images produced by the Bicubic interpolation, the variants of our method, and our full method on the

| Method                | ×2  | ×2.3 | ×2.6 | ×3  | ×3.3 | ×3.6 | ×4  |
|-----------------------|-----|------|------|-----|------|------|-----|
| Bicubic               | 33.66/0.9299 | 32.43/0.9121 | 31.46/0.8944 | 30.39/0.8692 | 29.62/0.8491 | 29.03/0.8333 | 28.42/0.8104 |
| AIDN<sub>w/o CRM</sub>| 45.73/0.9921 | 40.25/0.9766 | 39.33/0.9708 | 38.62/0.9642 | 36.86/0.9515 | 36.20/0.9456 | 35.53/0.9381 |
| AIDN<sub>w/o content</sub> | 44.06/0.9866 | 41.81/0.9794 | 40.25/0.9727 | 38.62/0.9639 | 37.68/0.9573 | 36.85/0.9513 | 35.93/0.9429 |
| AIDN (full method)    | 44.13/0.9868 | 41.86/0.9797 | 40.31/0.9729 | 38.70/0.9640 | 37.79/0.9579 | 36.98/0.9516 | 36.06/0.9436 |

Fig. 7. Visualization of the routing weights produced by the CRM in the decoder with varying scale factors (a), and image contents (b). The blue and red dots indicate two sampling locations for fixing the image content while varying the scale factor. The yellow line is the sampling locations for fixing the scale factor to be ×4 while changing the image content.

- AIDN<sub>w/o CRM</sub>: the scale-fixed variant of our AIDN by removing the scale-aware feature adaptation module in the feature extractor and replacing the CRM with the PixelShuffle;
- AIDN<sub>w/o content</sub>: the variant of our AIDN that is not content-adaptive by making the CRM only conditioned on scale factors; and
- AIDN: our full method.

As shown in Table 3, the PSNR/SSIM values of the Bicubic method are much lower than others, which confirms jointly modeling the downscaling and the upsampling can significantly improve the performance. Comparing the results of AIDN<sub>w/o CRM</sub> and AIDN<sub>w/o content</sub>, we can see the AIDN<sub>w/o content</sub> performs better for the non-integer scale factors while the AIDN<sub>w/o CRM</sub> performs slightly better for the integer factors. Note that, the AIDN<sub>w/o CRM</sub> is designed for fixed integer scale factors, so we adopt a similar solution as the type III methods explained in Sec. 4.2 to support non-integer scale factors, which means multiple trained models have to be stored in place. Most importantly, our full method outperforms the AIDN<sub>w/o content</sub> for both integer and non-integer scale factors. It demonstrates applicability of our framework and the effectiveness of our content-adaptive design for the conditional resampling module (CRM).

4.5 Discussion

Visualization of routing weights. To verify the effectiveness of the CRM for dynamically producing resampling kernels for different scale factors and image contents, we visualized the routing weights produced by the CRM in the decoder network, with various scale factors and image contents. We first fixed the image content and changed the scale factors to see the resulting routing weights in the decoder. As shown in Figure 7 (a), for the same location on the image, eight experts are activated differently when the scale factor varies, and the tendency to change is not the same for different locations e.g., the two marked locations with blue and red dots. Then, we fixed the scale factor to be ×4 and uniformly sample the locations on a line in the image (Figure 7 (b)) to observe the resulting routing weights. We can see the eight routing weights are almost the same within the flat region, while changing sharply along with the edge in the image. It demonstrates our CRM can produce adaptive resampling kernels according to both the scale factor and the image content.

HR information embedding. To further explore the embedding mechanism of HR information in the LR image, we visualized the difference maps between our embedded LR images and bicubic-downscaled ones under various scale factors in Figure 8. For convenience, we resized difference maps into the same resolution. We can observe that the differences are trivial when the scale factor is near to one, while the differences are concentrated on the edges when the scale factor is faraway from one, and the differences become more significant with the scale factors increasing. This is because there is very little HR information need to be embedded when the scale factor is near to one, while more HR information is demanded to be embedded when the scale factor increases.

Limitation. Since our AIDN encodes the HR information into the downscaled LR image, the quality of the restored HR image highly depends on the embedded information. Therefore, the downscaled LR image by our AIDN may not be resistant to general image manipulations, such as JPEG compression and editing. But many studies [Du et al. 2021; Hu et al. 2020; Liu et al. 2021; Wu et al. 2021; Xing et al. 2021] have shown that incorporating the compression during training can significantly boost the robustness to compression, which is the potential direction of our future work.

5 CONCLUSION

We present a scale-arbitrary invertible image downscaling network (AIDN) to natively downscale high-resolution (HR) images with arbitrary scale factors, such that the downscaled images can meet the resolution restriction of conventional social media platforms. Meanwhile, the HR information is embedded in the downscaled low-resolution (LR) counterparts; thus, our AIDN can also restore the original HR images with high-quality solely from the LR images whenever users want to explore the details of the distributed images. Our technical contribution is the proposed Conditional Resampling
Module (CRM) that can dynamically resample feature maps to the target resolution according to both the scale factor and the image content. Extensive ablation experiments and the visualization of the routing weights verified our design intent for the CRM. Both quantitative and qualitative results demonstrate our AIDN achieves top performance for invertible image downsampling with both arbitrary integer and non-integer scale factors.

REFERENCES

Eirikur Agustsson and Radu Timofte. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW).

Johannes Ballé, Valero Laraparra, and Eero P Simoncelli. 2016. End-to-end optimization of nonlinear transform codes for perceptual quality. In Picture Coding Symposium (PCS).

Johannes Ballé, Valero Laraparra, and Eero P Simoncelli. 2017. End-to-end optimized image compression. In International Conference on Learning Representations (ICLR).

Jens Behrman, Will Graft Mohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. 2019. Invertible residual networks. In International Conference on Machine Learning (ICML).

Yoshua Bengio, Nicholas Leonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neural networks. In Proceedings of the 30th International Conference on Machine Learning (ICML). 134–142.

Yuwen Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. 2019. Second-order routing weights verified our design intent for the CRM. Both quantitative and qualitative results demonstrate our AIDN achieves top performance for invertible image downsampling with both arbitrary integer and non-integer scale factors.
Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018a. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision (ECCV)*.

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. 2018b. Residual dense network for image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yikang Zhang, Jian Zhang, Qiang Wang, and Zhao Zhong. 2020. Dynet: Dynamic convolution for accelerating convolutional neural networks. *arXiv preprint arXiv:2004.10694* (2020).

Zhisheng Zhong, Tiancheng Shen, Yibo Yang, Zhouchen Lin, and Chao Zhang. 2018. Joint sub-bands learning with clique structures for wavelet domain super-resolution. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Qianshu Zhu, Chu Han, Guoqiang Han, Tien-Tsin Wong, and Shengfeng He. 2020. Video Snapshot: Single Image Motion Expansion via Invertible Motion Embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2020).