Towards Bidirectional Arbitrary Image Rescaling: Joint Optimization and Cycle Idempotence

Zhihong Pan[1], Baopu Li[1], Dongliang He[2]
Mingde Yao[2,3], Wenhao Wu[2], Tianwei Lin[2], Xin Li[2], Errui Ding[2]
[1]Baidu Research (USA), Sunnyvale, CA 94089, USA
[2]Department of Computer Vision Technology (VIS), Baidu Inc., Beijing, China
[3]University of Science and Technology of China

Abstract

Deep learning based single image super-resolution models have been widely studied and superb results are achieved in upscaling low-resolution images with fixed scale factor and downscaling degradation kernel. To improve real world applicability of such models, there are growing interests to develop models optimized for arbitrary upscaling factors. Our proposed method is the first to treat arbitrary rescaling, both upscaling and downscaling, as one unified process. Using joint optimization of both directions, the proposed model is able to learn upscaling and downscaling simultaneously and achieve bidirectional arbitrary image rescaling. It improves the performance of current arbitrary upscaling models by a large margin while at the same time learns to maintain visual perception quality in downscaled images. The proposed model is further shown to be robust in cycle idempotence test, free of severe degradations in reconstruction accuracy when the downscaling-to-upscaling cycle is applied repetitively. This robustness is beneficial for image rescaling in the wild when this cycle could be applied to one image for multiple times. It also performs well on tests with arbitrary large scales and asymmetric scales, even when the model is not trained with such tasks. Extensive experiments are conducted to demonstrate the superior performance of our model.

1. Introduction

In real world applications, it is common to rescale an image with arbitrary scale factors, either scaling up or down, for various purposes like display, storage or transmission. While recent deep learning based image super-resolution (SR) method have advanced the performance of image upscaling significantly, they are mostly optimized for fixed scale factors and known downscaling degradation kernels. Lately, there are growing interests in SR models that support arbitrary scale factors and great successes have been achieved, including arbitrary upscaling for scale factors in certain range [11], or learning a continuous image representation to resize it at any larger resolution [6], or asymmetric arbitrary upscaling where the vertical and horizontal scale factors could be different [26]. Like standard SR models, these methods are all optimized for the unidirectional upscaling process. In contrast, another line of image rescaling models [14, 25, 29] are developed to optimize the downscaling process together with the inverse upscaling and are able to improve accuracy on the upscaling task significantly comparing to unidirectional SR models of the same scale factors. Currently these bidirectional rescaling models are limited to a specific integer scale as far as we know.

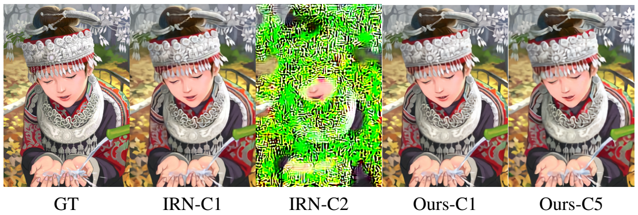


Figure 1. Visual examples of quality degradation from multiple downscaling-to-upscaling cycles: IRN [29] and ours.

Here we propose a joint optimization process that is able to learn arbitrary downscaling and arbitrary upscaling simultaneously. By modeling both downscaling and upscaling as equivalent subpixel splitting and merging processes and learning through a downscaling-to-upscaling cycle, the proposed method is able to achieve the best arbitrary upscaling accuracy while maintaining high perception-quality in downscaled outputs. An LIIF [6]-like subpixel weight function (SVF) and a novel subpixel weight function (SWF) are introduced for subpixel splitting and merging respectively. Using ground-truth (GT) image as supervision for high-resolution (HR) reconstruction, plus a weak supervision in low-resolution (LR), the proposed model, jointly optimized for both upscaling and downscaling, is able to greatly advance performances in arbitrary image rescaling, including very large or asymmetric scales.

In addition, so far as we know, current models are only evaluated for a single application of the downscaling-to-upscaling cycle and the effects of multiple cycles have never been studied. Ideally, application of additional cycle should not introduce any further changes beyond the initial cycle. This ideal downscaling-to-upscaling process, by definition, is an idempotent operation. In other words, for a function f of variable x, f is idempotent if $\forall x, f(f(x)) = f(x)$. While an ideal idempotent rescaling cycle may not be feasible, it is desirable to have minimum additional degradation when more than one downscaling-to-upscaling cycle is applied. Here a proxy objective is studied for optimizing both reconstruction accuracy and idempotence, and a cycle idempotence test is introduced to assess the output quality from multiple cycles in comparison to the original GT. As shown in Fig. 1, IRN [29] has high quality result for the first cycle (C1), but severe artifacts appear pervasively when the output from C1 is used as input for C2. In comparison, results from ours have similar high quality at C1, and little visible artifact even at C5.

In summary, the main contributions of our work include:

- First model to consider bidirectional arbitrary image downscaling and upscaling as a joint process and set SOTA performance in arbitrary image rescaling.
- A newly proposed cycle idempotence test which demonstrates our method's superior performance in model robustness after repetitive downscaling-to-upscaling cycles.
- Achieving SOTA in tests of arbitrary asymmetric scales and large out-of-distribution scales too.

2. Related Work

Arbitrary Scale Super-Resolution. Deep learning based image super-resolution have been studied extensively for the last few years [8,15,20,31,32], and these methods commonly train one model for one fixed scale factor. Lim *et al.* [20] was the first to propose a multi-scale SR model, which shares one feature learning backbone for different scales but still needs scale specific processing modules to handle the last step for multiple scales. Later, Li *et al.* [19] proposed a multi-scale residual network that learns multi-scale spatial features using convolution layers with different kernel sizes. However, these methods are still limited to a fixed set of integer scale factors. Inspired by weight prediction techniques in meta-learning [18], Hu *et al.* [11] proposed a single Meta-SR model to solve SR at arbitrary scale factors by predicting weights of convolutional layers for arbitrary scale factors, not limited to a fixed set of integer ones. The newest ArbSR [26] proposed a plug-in module to further optimize existing SR models for arbitrary asymmetric SR where scale factors along horizontal and vertical directions could be different. These arbitrary SR works are often limited to a fixed maximum scale factor to maintain high performance. Most recently, Chen *et al.* [6] proposed to learn pixel representation features to replace pixel value features in previous methods. With a learned local implicit image function (LIIF), this model is able to predict pixel values at arbitrary large scales. Our work extends the idea of LIIF to be applicable for arbitrary downscaling and upscaling at the same time.

Bidirectional Image Rescaling. As pointed out above, most super-resolution models rely on LR-HR pairs where each LR image is downscaled from the corresponding HR using frequency-based kernels like Bicubic [22]. These models are trained for upscaling reconstruction only without taking the image downscaling method into joint consideration. To take advantage of the potential mutually beneficial reinforcement between downscaling and the inverse upscaling, Kim *et al.* [14] proposed an auto-encoder framework to jointly train image downscaling and upscaling together. Similarly, Sun *et al.* [25] proposed a new content adaptive-resampler based image downscaling method, which can be jointly trained with any existing differentiable upscaling (SR) models. More recently, Xiao *et al.* [29] proposed an invertible rescaling net (IRN) that has set the state-of-the-art (SOTA) 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 z from the normal distribution without need of the case specific $\hat{z}$. Current methods for bidirectional image rescaling are limited to a fixed integer scale factor like $\times 4$. As a contrast, we propose a bidirectional arbitrary rescaling approach in this work.

Idempotent Image Processing. For image processing, there are numerous examples of idempotent filters such as median filter [23], cascaded median filters [9] and basic morphological operations like opening and closing [10]. For many image processing application, it is beneficial to have idempotent filters or processes for various reasons. In the case of image JPEG compression, an image could be compressed multiple times as it is not known for sure if an image in the wild is already compressed. To reflect the importance of repetitive image compression, there is a specific key feature of multi-generation robustness in the standardization process of JPEG XS [7]. Lately, it has been discovered that there is model instability issue in successive deep image compression that results in severe visual artifacts [16]. Here we will specifically study cycle idempotence of image SR and rescaling models after repetitive applications of the downscaling-to-upscaling process.

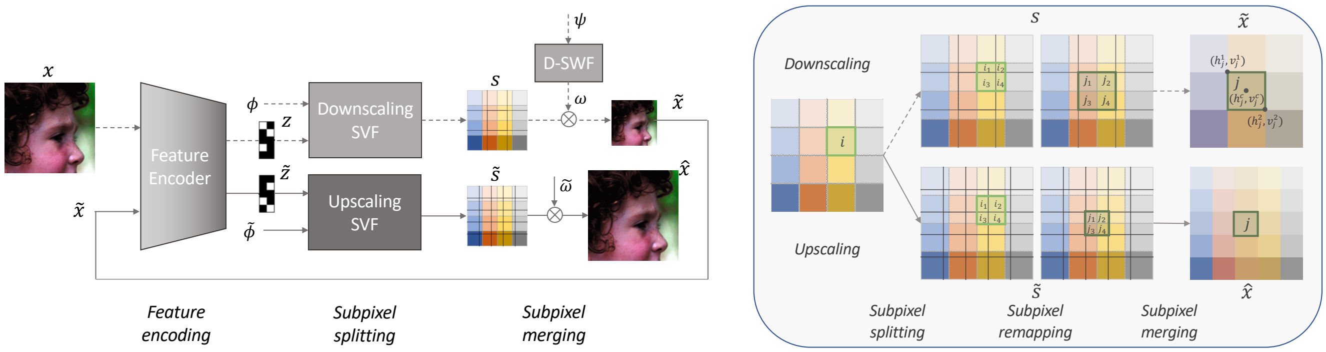


Figure 2. Proposed framework for a bidirectional arbitrary image rescaling network (BAIRNet) with magnified illustration of the subpixel splitting and merging process.

Weak Supervision in Low Level Vision. Weak supervision in a branch of supervised learning where supervision signals, like labels for image classification, are from either imprecise or noisy sources. While it has been widely studied in high level tasks like object detection [4] and semantic segmentation [13], its applications in low level vision tasks like image restoration and reconstruction, however, are not fully addressed. For image SR where LR-HR pairs are needed, LR images are often synthesized from HR using bicubic interpolation and they could be imprecise comparing to real world LR images. There have been efforts to co-collect true LR and HR images to build real-world SR dataset like RealSR [5] and DRealSR [28]. However, these pairs are subject to registration imprecision and/or local motion blur as LR and HR are taken sequentially using different lenses. For newest bidirectional rescaling models [14, 25, 29] that learn downscaling and upscaling jointly, although accuracy in inverse upscaling is the primary goal, a weakly supervised learning for downscaling is still needed. While the LR image from bicubic interpolation has been previously used as the downscaling reference, new forms of weak supervision are investigated in our study.

3. Proposed Method

3.1. Bidirectional Arbitrary Image Rescaling

The proposed framework for a bidirectional arbitrary image rescaling network (BAIRNet) is illustrated in Fig. 2. It is a bidirectional process, including downscaling to convert the GT image x to an LR image $\tilde{x}$, and upscaling to restore a HR image $\hat{x}$ from $\tilde{x}$. As illustrated on the left, each of the two directions consists of the same three steps: feature encoding, subpixel splitting and subpixel merging. These steps are denoted as

$$\begin{aligned}\tilde{x}&=f_m(s,\omega)=f_m(f_s^D(z,\phi),\omega)=f_m(f_s^D(f_e(x),\phi),\omega)\\ \hat{x}&=f_m(\tilde{s},\tilde{\omega})=f_m(f_s^U(\tilde{z},\tilde{\phi}),\tilde{\omega})=f_m(f_s^U(f_e(\tilde{x}),\tilde{\phi}),\tilde{\omega})\end{aligned} \quad (1)$$

for downscaling and upscaling respectively. For two of the three steps, the same feature encoding f_e and subpixel merging f_m are shared for both downscaling and upscaling. For subpixel splitting though, different subpixel value function (SVF) are trained for downscaling and upscaling, denoted as f_s^D and f_s^U respectively. For variables, z and $\tilde{z}$ are feature vectors, s and $\tilde{s}$ are subpixel values, ϕ and $\tilde{\phi}$ are subpixel coordinates, and ω and $\tilde{\omega}$ are subpixel weights used in merging. Each pair has the same definition and ˜ is used to differentiate upscaling from downscaling.

To illustrate the subpixel splitting and merging process, a magnified illustration is included in Fig. 2 on the right. One pixel i of the input is split to one or more subpixels first, and after remapping, a new set of subpixels is merged to one pixel j in the rescaled image. A subpixel k in the intermediate step is defined as one rectangle in the image space that reside wholly inside one original pixel as well as inside one rescaled pixel, and its boundaries are aligned with the boundaries of input and/or rescaled pixels. Here we use p_i, r_j and s_k to represent values of pixels i, j and k respectively. In the illustrated downscaling example, pixel i is split to 4 subpixels, denoted as $\mathbb{P}_i = \{i_1, i_2, i_3, i_4\}$. To merge subpixels to rescaled pixels, a remapping process is needed to associate groups of subpixels corresponding to rescaled pixels. In the example in Fig. 2, output pixel j is merged from 4 subpixels, denoted as $\mathbb{R}_j = \{j_1, j_2, j_3, j_4\}$. Lastly, as illustrated in the downscaled pixel j, its central, top-left and bottom-right coordinates are denoted as (h_j^c, v_j^c), (h_j^1, v_j^1) and (h_j^2, v_j^2) respectively.

For input pixel i and subpixel $k, k \in \mathbb{P}_i$, defining ϕ_k^i as relative coordinates of k in reference to i: $\phi_k^i = (h_k^1 - h_i^c, v_k^1 - v_i^c, h_k^2 - h_i^c, v_k^2 - v_i^c)$, the process to predict subpixel values s_k during subpixel-splitting is denoted as

$$s_k = f_s(z_i, \phi_k^i) \quad (2)$$

where f_s is the SVF and z_i is the feature vector of pixel i. This process is the same for both downscaling and upscaling but separate SVFs are trained and denoted differently in Eq. 1 for distinction.

The value of pixel j at subpixel-merging is computed as

$$r_j = \sum_{k \in \mathbb{R}_j} \omega_k^j s_k / \sum_{k \in \mathbb{R}_j} \omega_k^j \quad (3)$$

where ω_k^j is the weight of subpixel k during merging of pixel j. For the subpixel merging weights in upscaling, $\tilde{\omega}_k^j$ is simply defined as the area of subpixel k. As the majority of upscaled pixel j consists of just one subpixel k, and the others have either 2 or 4, the area based weights are sufficient to represent the significance of each subpixel. While in the case of downscaling, each pixel j may include a large number of subpixels and the impact of each subpixel should dependent on both its size and location. Here we propose a subpixel weight function (SWF) module to learn the subpixel weights for merging during the end-to-end training, denoted as $\omega_k^j = f_w(\psi_k^j)$. Similar to ϕ_k^i, ψ_k^j is defined as $(h_k^1 - h_j^c, v_k^1 - v_j^c, h_k^2 - h_j^c, v_k^2 - v_j^c)$.

While this framework has some resemblance with two prior works, that is, IRN [29] and LIIF [6], there are some substantial differences between our proposed and the previous two. First, IRN is limited to one fixed integer scale per trained model. Although it is also trained to optimize both downscaling and upscaling together, it is based on an invertible network which uses forward and backward inferences for downscaling and upscaling respectively. In contrast, ours is utilizing the same three-step process for both directions, and only one model is needed to handle arbitrary scales. IRN also samples auxiliary latent variables randomly during the backward upscaling process which brings uncertainty and causes severe artifacts in cycle idempotence tests. Comparing to LIIF, our model consolidates the downscaling and upscaling process to utilize similar implicit functions for both arbitrary downscaling and upscaling. As a result, it can be trained for bidirectional arbitrary rescaling and leads to great improvements in performance. Lastly, asymmetric scales are not studied in LIIF.

3.2. Idempotent Image Rescaling

The rescaling cycle defined in Eq. 1 can be simplified as $\hat{x} = f(x)$. Without considering constraints in LR, the primary goal to optimize this cycle is to minimize its reconstruction loss, but it is also desirable to learn an idempotent one. These two objectives could be defined separately as

$$\begin{aligned} f &= \arg\min\nolimits_{f_\zeta} \mathcal{L}(x, f_\zeta(x)) \\ f &= \arg\min\nolimits_{f_\eta} \mathcal{L}(f_\eta(x), f_\eta(f_\eta(x))) \end{aligned} \tag{4}$$

As these two objectives may conflict, an empirical proxy objective is proposed to learn a compromise between the two. In practice, the model is trained to minimize reconstruction error after n cycles, described as

$$f = \arg\min\nolimits_{f_\theta} \mathcal{L}(x, f_\theta^n(x)), n \in [1, N] \tag{5}$$

where f_θ^n means f_θ is applied n times. When N is set as 1, this proxy objective is equivalent to the primary task of 1-cycle reconstruction. In our experiments, different values of N are investigated to compare the trade-off between two objectives: reconstruction accuracy and cycle idempotence.

3.3. Weak Supervision in LR

Considering the multi-cycle optimization as in Eq. 5 and the need to generate visually coherent LR images, the overall loss for training our model is defined as

$$\mathcal{L} = \lambda_1 \mathcal{L}_{rec}(x, \hat{x}^n) + \lambda_2 \mathcal{L}_{ref}(x, \tilde{x}^n) \tag{6}$$

where $\mathcal{L}_{rec}$ is the reconstruction loss for HR, $\mathcal{L}_{ref}$ is the reference loss in LR, and $\tilde{x}^n$ and $\hat{x}^n$ are the LR and HR outputs after n-cycles respectively. Although it is possible to train it fully self-supervised by setting $\lambda_2 = 0$, it will lead to visually non-meaningful $\tilde{x}$ due to random initialization. In previous methods [14, 25, 29], a L_2 reference loss $L_2(\tilde{x}, \bar{x})$, where $\bar{x}$ is the LR reference downsampled from x using Bicubic [22] method, is used as an imprecise supervision. In contrast to previous methods, various strategies, like reducing λ_2 to 0 at later stage of training, or calculating $\mathcal{L}_{ref}$ from the mean values of each color channel instead of per-pixel, are explored in our study to demonstrate the advantage of weak supervision in LR.

4. Experiments

4.1. Data and Settings

For fair comparison with previous works like LIIF and IRN, the same 800 HR images from DIV2K [1] are used for training. For quantitative evaluation, we use HR images of five commonly used benchmark datasets, including Set5 [3], Set14 [30], BSD100 [21], Urban100 [12] and Manga109 [12], plus 100 HR images from the DIV2K validation set. Following previous practices like LIIF, we take the peak noise-signal ratio (PSNR) and SSIM [27] on the luminance channel for the 5 benchmark sets, but use the same metrics in RGB color space for DIV2K validation set.

For the 200×200 input HR patches in one training batch, each is assigned with a random downscaling scale sampled from a uniform distribution of $\mathcal{U}(1, 4)$. For individual modules, we use RDN [32] minus its upsampling module as the feature encoder, which generates a feature map with the same size as the input image. For both downscaling and upscaling SVF, a 5-layer MLP with ReLU activation and hidden dimensions of 256 is used. For the downscaling SWF, a 5-layer MLP with hidden dimensions of 16 is used. With a batch size of 8, all models are trained using Adam [17] optimizer. In order to conduct ablation studies efficiently, a pretrained model is generated after 500 epochs, 300 iteration each, from an initial learning rate of 10^{-4}. The learning rate decays by half after every 100 epochs. For this stage, $\mathcal{L}_{rec}$ is set as a pixel-level L1 loss and $\mathcal{L}_{ref}$ is set as L2, and no SWF module is included. The pretrained one is further trained for 500 epochs get the base model BAIRNet, with downscaling SWF included, and $\mathcal{L}_{ref}$ is set as L2 for the mean pixel value per color channel. λ_1 and λ_2 are set as 1 unless specified otherwise. BAIRNet is further fine-tuned for 200 epochs using the proxy objective as defined

Table 1. Quantitative comparison of SOTA SR and rescaling methods with the best two results highlighted in red and blue respectively (methods in **bold** require multiple models and additional interpolations to achieve arbitrary scales).

Method	Scale	Param.	Set5	Set14	BSD100	Urban100	Manga109	DIV2K
Bicubic	×1.5	-	36.75/0.9611	32.86/0.9268	32.16/0.9133	29.49/0.9095	34.79/0.9707	33.95/0.9416
RCAN [31]	×1.5	15.4M	40.97/0.9767	37.05/0.9578	35.59/0.9516	35.93/0.9660	42.33/0.9889	38.47/0.9701
Meta-SR [11]	×1.5	15.5M	41.47/0.9785	37.52/0.9601	35.86/0.9543	36.91/0.9696	43.17/0.9904	38.88/0.9718
LIIF [6]	×1.5	22.3M	41.23/0.9774	37.37/0.9591	35.76/0.9536	36.70/0.9684	42.84/0.9894	38.82/0.9717
ArbSR [26]	×1.5	16.6M	41.47/0.9786	37.51/0.9603	35.86/0.9547	36.92/0.9697	43.12/0.9904	38.84/0.9719
CAR [25]	×1.5	51.1M	40.50/0.9763	37.08/0.9596	35.72/0.9535	34.70/0.9635	40.90/0.9881	37.93/0.9683
IRN [29]	×1.5	1.66M	43.55/0.9891	39.52/0.9795	39.28/0.9833	36.52/0.9811	42.64/0.9936	40.18/0.9838
BAIRNet†	×1.5	22.4M	47.13/0.9849	43.12/0.9760	46.63/0.9959	44.01/0.9946	45.49/0.9948	44.99/0.9920
Bicubic	×2.5	-	31.76/0.8983	28.52/0.8196	28.13/0.7853	25.43/0.7837	28.56/0.8954	29.40/0.8505
RCAN [31]	×2.5	15.6M	36.05/0.9436	31.69/0.8815	30.47/0.8508	30.42/0.8990	36.59/0.9634	32.72/0.9079
Meta-SR [11]	×2.5	15.5M	36.18/0.9441	31.90/0.8814	30.47/0.8508	30.57/0.9003	36.55/0.9639	32.77/0.9086
LIIF [6]	×2.5	22.3M	35.98/0.9434	31.64/0.8813	30.45/0.8510	30.42/0.8992	36.39/0.9630	32.78/0.9091
ArbSR [26]	×2.5	16.6M	36.21/0.9448	31.99/0.8830	30.51/0.8536	30.68/0.9027	36.67/0.9646	32.77/0.9093
CAR [25]	×2.5	52.8M	37.33/0.9548	33.78/0.9169	32.53/0.9020	32.19/0.9301	37.63/0.9717	34.32/0.9310
IRN [29]	×2.5	4.35M	39.78/0.9742	36.39/0.9553	35.56/0.9542	33.99/0.9589	39.33/0.9836	36.60/0.9607
BAIRNet†	×2.5	22.4M	40.11/0.9664	36.62/0.9469	36.29/0.9563	36.62/0.9679	40.26/0.9830	37.46/0.9627
Bicubic	×3.5	-	29.30/0.8374	26.52/0.7362	26.50/0.7003	23.70/0.6935	25.83/0.8203	27.38/0.7802
RCAN [31]	×3.5	15.6M	33.47/0.9138	29.24/0.8141	28.42/0.7731	27.61/0.8348	32.74/0.9328	30.13/0.8511
Meta-SR [11]	×3.5	15.5M	33.59/0.9146	29.60/0.8140	28.42/0.7728	27.71/0.8356	32.75/0.9337	30.18/0.8524
LIIF [6]	×3.5	22.3M	33.41/0.9133	29.20/0.8131	28.39/0.7714	27.60/0.8334	32.60/0.9324	30.16/0.8517
ArbSR [26]	×3.5	16.6M	33.63/0.9149	29.58/0.8147	28.41/0.7744	27.69/0.8360	32.84/0.9339	30.14/0.8518
CAR [25]	×3.5	52.8M	34.98/0.9303	31.38/0.8643	30.14/0.8326	29.97/0.8871	35.00/0.9507	31.88/0.8865
IRN [29]	×3.5	4.35M	37.12/0.9546	33.65/0.9196	32.54/0.9047	31.84/0.9277	36.86/0.9690	33.84/0.9281
BAIRNet†	×3.5	22.4M	36.85/0.9472	32.97/0.9074	32.36/0.8986	32.71/0.9338	36.98/0.9671	33.87/0.9266

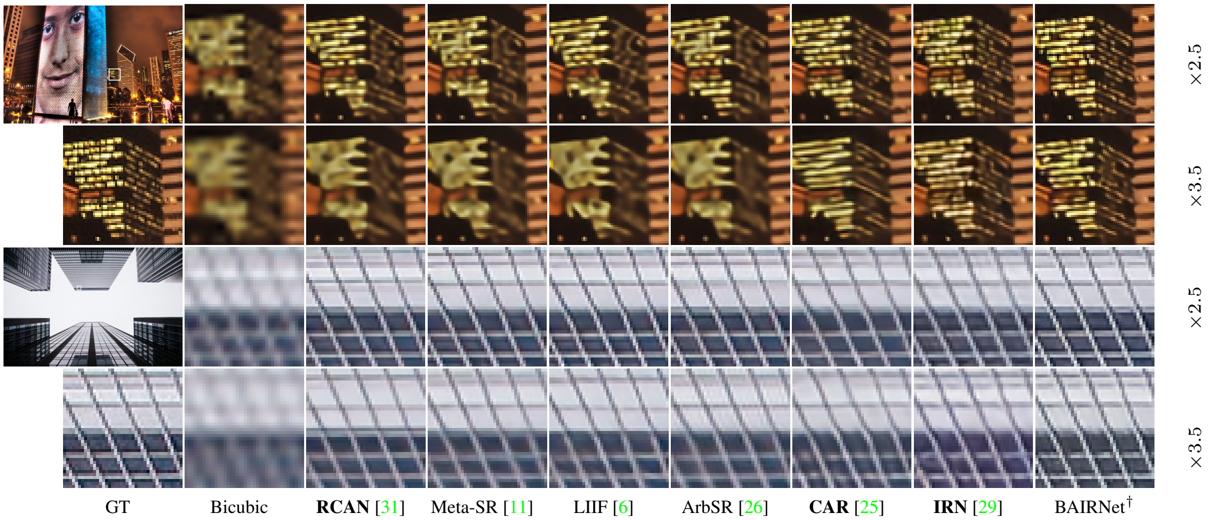


Figure 3. Visual examples of arbitrary rescaling from Urban100 and DIV2K at two scales: ×2.5 and ×3.5 (Best viewed for online version).

in Eq. 5, where N is set as 3 and the final model is denoted as BAIRNet† with † used for distinction.

4.2. Arbitrary Rescaling Performance

To assess the performance of our proposed method for arbitrary rescaling, we compare rescaled HR images using a set of arbitrary scales. For each fixed scale, the resolution of LR images are kept the same for all methods for fair comparison. For models trained for integer scales only, like RCAN and IRN, evaluation on arbitrary scales is implemented as upscaling LR using the closest oversampled integer scale (use ×3 for any scales between 2 and 3) before resampling using bicubic interpolation to target size. For bidirectional CAR [25] and IRN, HR inputs are also pre-upsampled accordingly. As listed in Table 1, PSNR and SSIM results from three scales (×1.5/2.5/3.5) are compared. It shows that our BAIRNet† outperforms others by a comfortable margin for ×1.5 and ×2.5, and it is the best in ×3.5 tests for the 3 large test sets out of 6 while trailing slightly behind IRN for the other 3. Visually as shown in Fig. 3, bidirectional methods like IRN and ours are the best overall. Between the two, IRN is more blurry

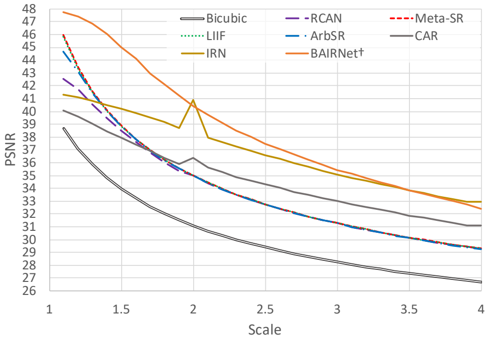


Figure 4. Cross-scale performance comparison for arbitrary rescaling testing ($\times 1.1 - \times 4$) on DIV2K validation set.

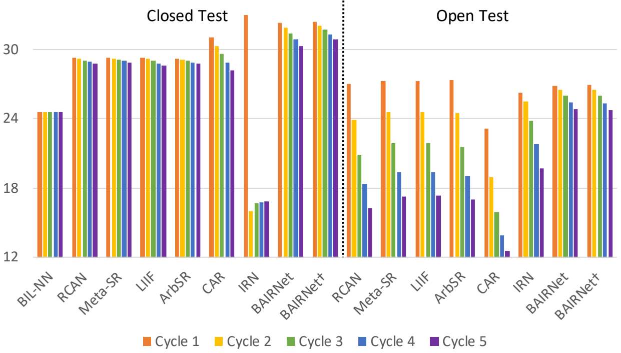


Figure 5. PSNR results from closed and open idempotence tests for 1-5 cycles on DIV2K validation set ($\times 4$).

in $\times 2.5$ and its color is off in the second example of $\times 3.5$. Results of continuous scales between $\times 1.1$ and $\times 4$ (sampled every 0.1) are also illustrated in Fig. 4 to compare our model with others. For arbitrary upscaling models like Meta-SR, LIIF and ArbSR, they are essentially equivalent to RCAN and with each other for scales above $\times 2$. Bidirectional models CAR and IRN improves performances in larger scales greatly but their performances suffer at small arbitrary scales. BAIRNet† is clearly the best overall, at the top for all scales except trailing slightly behind IRN for scales above $\times 3.5$, plus the spike at $\times 2$, where no extra interpolation needed is needed for IRN.

4.3. Cycle Idempotence

A cycle idempotence test is defined as the assessment of $\mathcal{L}(x, f^n(x))$ for different number of cycles, where f^n means the rescaling cycle f is applied n times. Here we use the PSNR value in place of $\mathcal{L}$ for test assessment. For the first set of test, defined as closed test, the downscaling function is fixed as the one best matches its upscaling one. So for RCAN, Meta-SR, ArbSR and LIIF, matlab_imresize [24] is used for its equivalence with the Matlab one. For other bidirectionally trained models, their own corresponding downscaling process are applied respectively. For the open test, which means the downscaling is set freely, cv2.resize with INTER_AREA interpolation is picked for its wide application, and it is used for all methods for fair comparison.

To avoid extra interpolation for RCAN and IRN, a $\times 4$ scale is chosen here for testing on the DIV2K validation set and the results are compared in Fig. 5. For the closed test, BIL-NN (bilinear for downsampling and nearest-neighbour for upsampling) is included as a perfectly idempotent reference. IRN has the best performance at cycle 1 with its bidirectional learning and invertible network structure. However, there is a drastic drop from cycle 2 and hereafter, probably caused by the random latent variable sampling during the upscaling process. For our method, both BAIRNet and BAIRNet† are included to show the improvement in idempotence of BAIRNet†, which is fine-tuned from BAIRNet using the proxy objective of multiple-cycle losses. At cycle 1, both are closely behind IRN, and their additional quality losses from multiple cycles are gradual. After 5 cycles, both are still significantly better than upscaling-only models at cycle 1, and BAIRNet† is clearly better than BAIRNet. This shows the advantage of our method in robustness to repetitive rescaling cycles in closed settings, and the effectiveness of multi-cycle losses. For open tests on the right, while all models are subject to significant performance losses at cycle 1 comparing to closed tests, our models have a much slower degradation for multiple cycles. Due to page limitation, more visual examples of cycle idempotence tests are included as supplementary materials.

Table 2. PSNR improvements over base BAIRNet after fine-tuning using N-cycle losses (Eq. 5), testing 5 cycles each for 3 scales.

Cycle	$N=1$			$N=3$			$N=5$		
	$\times 4$	$\times 3$	$\times 2$	$\times 4$	$\times 3$	$\times 2$	$\times 4$	$\times 3$	$\times 2$
1	0.09	0.09	0.09	0.08	0.10	0.13	0.05	0.08	0.13
2	0.14	0.16	0.21	0.16	0.21	0.33	0.15	0.21	0.34
3	0.21	0.24	0.32	0.27	0.36	0.56	0.28	0.39	0.60
4	0.29	0.34	0.42	0.40	0.54	0.78	0.44	0.60	0.86
5	0.38	0.43	0.49	0.54	0.73	1.00	0.61	0.82	1.11

To study the effectiveness of various N in Eq. 5, the base BAIRNet model is trained for another 200 epochs using $N = 1, 3, 5$ respectively. As shown in Table 2, the improvements in PSNR for three fine-tuned models are compared for 1-5 cycles at 3 different scales. There are consistent improvements across scales and cycles even when $N = 1$, indicating the base BAIRNet is not fully trained. For $N = 3$, it is shown to improve PSNR more significantly after multi-cycles, especially for smaller scales, while only trailing by 0.01 at 1-cycle for $\times 4$. For $N = 5$, the corresponding gain at multi-cycles is larger, but there is trade-off of accuracy at 1-cycle. Overall, it is demonstrated that the proposed proxy objective is effective at increasing model robustness in cycle idempotence while maintaining high performance at the primary goal of 1-cycle reconstruction accuracy.

Table 3. PSNR results achieved for both symmetric and asymmetric scale factors (methods in **bold** require extra interpolations).

	Set5				Set14				BSD100				Urban100				Manga109			
	$\frac{\times 2}{\times 3}$	$\frac{\times 1.6}{\times 3.2}$	$\frac{\times 3.6}{\times 1.2}$	$\times 2.5$	$\frac{\times 2}{\times 3}$	$\frac{\times 1.6}{\times 3.2}$	$\frac{\times 3.6}{\times 1.2}$	$\times 2.5$	$\frac{\times 2}{\times 3}$	$\frac{\times 1.6}{\times 3.2}$	$\frac{\times 3.6}{\times 1.2}$	$\times 2.5$	$\frac{\times 2}{\times 3}$	$\frac{\times 1.6}{\times 3.2}$	$\frac{\times 3.6}{\times 1.2}$	$\times 2.5$	$\frac{\times 2}{\times 3}$	$\frac{\times 1.6}{\times 3.2}$	$\frac{\times 3.6}{\times 1.2}$	$\times 2.5$
ArbSR	35.90	35.90	35.85	36.21	31.89	31.96	31.59	31.99	30.58	30.87	30.24	30.51	30.59	30.60	29.74	30.68	36.17	35.88	35.30	36.67
IRN	38.66	38.55	38.80	39.78	35.49	35.31	35.02	36.39	34.87	34.69	34.19	35.56	32.75	32.44	32.09	33.99	37.42	37.08	37.12	39.33
BAIRNet	39.40	39.05	38.34	40.03	36.00	35.71	34.36	36.53	35.33	35.14	33.47	36.24	34.14	33.34	31.52	36.51	39.01	38.30	36.95	40.11
BAIRNet‡	40.06	40.42	40.11	40.16	36.66	37.00	36.33	36.68	36.28	36.96	36.64	36.17	36.32	36.74	35.82	36.43	40.01	40.27	39.42	40.14

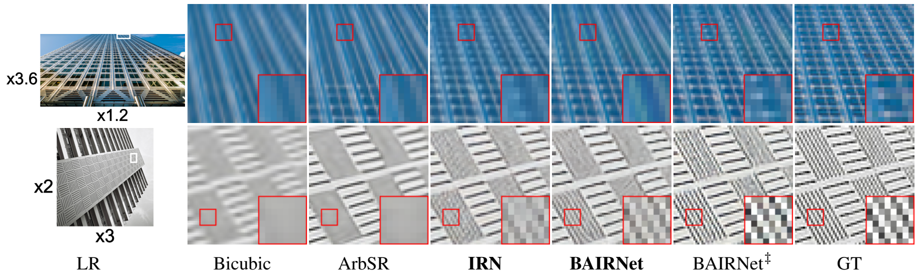


Figure 6. Visual examples of arbitrary asymmetric rescaling from Urban100 test set (Best viewed for online version).

Table 4. PSNR results for large out-of-distribution scales.

	×6	×12	×18	×24	×30
Bicubic	24.82	22.27	21.00	20.19	19.59
LIIF	27.02	23.95	22.40	21.40	20.71
ArbSR	26.61	23.07	21.45	20.49	19.81
BAIRNet	29.29	25.55	23.84	22.75	21.97

4.4. Out-of-distribution Assessment

While our model is trained with symmetric scale factors randomly distributed between $\times 1 - \times 4$ and mainly tested using such in-distribution settings, there is no such limitation in the capacity of the proposed method. For assessment, as shown in Table 3 and 4 respectively, our model is compared with ArbSR for asymmetric scales and with LIIF for large scales. In both cases, BAIRNet is used as is for upscaling by simply changing the output resolution and using corresponding $\tilde{\phi}$ as in Eq. 1. These out-of-distribution tests further demonstrate robustness of our proposed method.

For asymmetric scales $\frac{s_v}{s_h}$ where s_v is the vertical scale and s_h is for horizontal, comparisons of 5 benchmark test sets are shown in Table 3. Results from ArbSR are included as a SOTA baseline for unidirectional models where only upscaling is learned. For bidirectional IRN, although additional interpolations are needed for both downscaling and upscaling as only $\times 2$ and $\times 4$ models are available, its performance is far more superior comparing to the ArbSR baseline. For our base BAIRNet, pre-interpolation is only needed for the downscaling stage, where the input GT image is resampled using bicubic interpolation with a s_m/s_v vertical scale and a s_m/s_h horizontal scale. Here $s_m = \sqrt{s_h s_v}$, which will convert the asymmetric scale to symmetric scale of s_m for downscaling, while keeping the number of input pixels approximately the same as GT for fair comparison. It is shown in Table 3 that BAIRNet is better than IRN with the exception of $\frac{\times 3.6}{\times 1.2}$, similar to the observation from Fig. 4 that IRN is slightly better than BAIRNet for scales larger than $\times 3.5$. Unlike IRN that is limited to training data with symmetric scales, BAIRNet could be further trained using data with asymmetric scales and the fine-tuned model is denoted as BAIRNet‡. It no longer needs the initial symmetric conversion step and shows further significant improvements in asymmetric tests, while subjects to slight degradation in symmetric test for only 2 out 5 test sets. Visual examples in Fig. 6 clearly show that BAIRNet‡ is able to reproduce more fine details at random asymmetric scales.

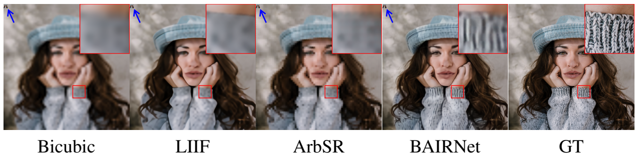
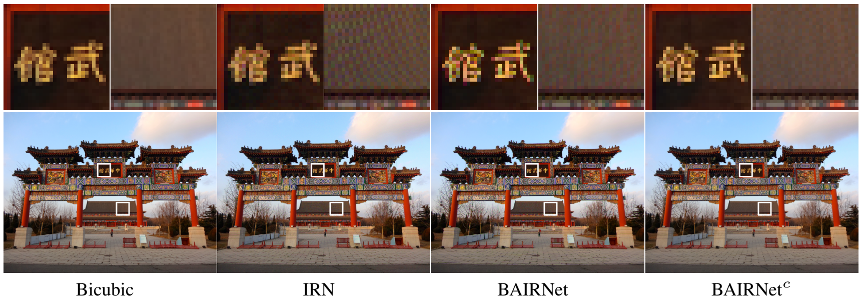


Figure 7. Examples of large scale factors ($\times 30$) with ↑ pointing to corresponding LR inputs (Best viewed when zooming in).

For tests of large scales as shown in Table 4, although all models are trained from images with scales up to $\times 4$, LIIF is much more robust for large scales up to $\times 30$ comparing to ArbSR. For testing BAIRNet at scale s, the GT image is pre-downscaled by a scale of $s/4$ using bicubic so the downscaling step in BAIRNet is capped at $\times 4$. This is a reasonable choice as this reduces the number of pixels used for the time-consuming step of downscaling feature encoding. It is shown that BAIRNet is consistently much better than LIIF quantitatively. For visual examples in Fig. 7, BAIRNet is clearly able to recover sharper details comparing to the others. All LR inputs, either from bicubic resizing as in LIIF and ArbSR or downscaling by BAIRNet, have the same low resolution and are included in Fig. 7 for reference.



Bicubic IRN BAIRNet BAIRNet^c

Figure 8. Visual examples of downscaled LR images with false color artifacts in magnified view.

4.5. Ablation Study

To assess the effectiveness of modules like SWF and different LR supervision strategies, the pretrained model is trained for an additional 200 epoch using different settings as shown in Table 5. For settings of $\mathcal{L}_{ref}$, L_2 means L_2 loss at pixel level, L_2^c means L_2 for C_b and C_r channels only, and L_2^m refers to using mean pixel value per color channel. Note that for L_2^c, λ_2 is set as 2. It is shown clearly in Table 5 that the SWF module in downscaling step is beneficial across different scales. For $\mathcal{L}_{ref}$, it is obvious that weaker supervision leads to improved restoration accuracy overall, and the weak L_2^m is effectively the same as no supervision in LR at all. For $\mathcal{L}_{rec}$, as images of various scales are included in each batch of model training, it is expected that those with lower scales have smaller reconstruction losses naturally. Intuitively, a simple scale-normalization, $L_s = L_1/s$ where s is the rescaling factor, is used in the training of BAIRNet. For comparison, the model is also trained without such loss normalization, and as shown in the table, it is equivalent with the primary model using L_s for larger scales, but suffers from performance loss at smaller scales.

Table 5. PSNR results for different model and training settings.

	D-SWF	✗	✓	✓	✓	✓	✓
	$\mathcal{L}_{ref}$	L_2	L_2	L_2^c	L_2^m	✗	L_2^m
	$\mathcal{L}_{rec}$	L_s	L_s	L_s	L_s	L_s	L_1
DIV2K	×4	31.19	31.42	31.78	32.12	32.13	32.14
	×3	34.24	34.42	34.66	35.13	35.12	35.10
	×2	38.89	39.42	39.32	40.11	40.15	39.93

For the specific L_2^c, it is designed and tested as a measure to suppress false-color artifacts noticed in downsampled LR from bidirectional models like IRN and ours. As shown in Fig. 8, the LR images from bicubic downsampling and the learned bidirectional models like IRN and BAIRNet are hardly differentiable in normal display resolution. However, when magnified, there are noticeable moiré-like and false color artifacts from both models. Comparing to BAIRNet trained with L_2^m in LR, BAIRNet^c trained with the newly designed L_2^c loss is able to successfully suppress false color artifacts as shown in Fig. 8, while sacrificing the overall restoration quality slightly as listed in Table 5.

4.6. Limitations

The main limitation of our method is its lower speed and large memory consumption for downscaling. Though it is slightly faster than LIIF in upscaling without using the feature unfolding option, it slows down both inference and training for downscaling as the same feature encoder is applied to a larger number of pixels in the HR inputs. For very large scale factors though, as demonstrated in Section 4.4, it is not necessary to use the full resolution as input and a pre-downscaling could be applied to reduce number of input pixels and increase efficiency in downscaling feature encoding. One potential future improvement is to optimize the feature encoding module in downscaling for higher efficiency while maintaining accuracy. Lastly, only the default RDN backbone is used in our model. A RCAN backbone is expected to further improve performance, as demonstrated in ArbSR, but not tested here as it is slower than RDN.

5. Conclusion

Current deep learning based image SR and arbitrary SR models are all subject to one or multiple limiting factors in related to downscaling degradation kernel and scale factors. Modeling arbitrary downscaling and upscaling as one unified subpixel splitting and merging process, a bidirectional arbitrary image rescaling network (BAIRNet) is shown to improve upscaling accuracy significantly by jointly optimizing arbitrary upscaling and downscaling. Cycle idempotence tests are also used to test robustness of various models when the downscaling-to-upscaling cycle is applied multiple times, including closed test where the downscaling is limited to model assumptions or training settings, and open test where the downscaling is not limited. For closed and open tests overall, BAIRNet is the best with great performance at cycle 1 and no sudden drop in accuracy for following cycles. Additionally, a proxy objective that minimize multi-cycle losses is demonstrated to further improve model robustness in cycle idempotence. It is also shown that, even when BAIRNet is only trained for random symmetric scales between $\times 1-4$, it achieves impressive results for rescaling at asymmetric or large scales, outperforming SOTA methods LIIF and ArbSR with substantial margins.

References

[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 4

[2] Lynton Ardizzone, Jakob Kruse, Sebastian Wirkert, Daniel Rahner, Eric W Pellegrini, Ralf S Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Köthe. Analyzing inverse problems with invertible neural networks. *arXiv preprint arXiv:1808.04730*, 2018. 2

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, pages 135.1–135.10. BMVA Press, 2012. 4

[4] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with convex clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1081–1089, 2015. 3

[5] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3086–3095, 2019. 3

[6] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8628–8638, 2021. 1, 2, 4, 5

[7] Antonin Descampe, Joachim Keinert, Thomas Richter, Siegfried Fößel, and Gaël Rouvroy. JPEG XS, a new standard for visually lossless low-latency lightweight image compression. In *Applications of Digital Image Processing XL*, volume 10396, page 103960M. International Society for Optics and Photonics, 2017. 2

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 2

[9] Petri Haavisto, Moncef Gabbouj, and Yrjö Neuvo. Median based idempotent filters. *Journal of Circuits, Systems, and Computers*, 1(02):125–148, 1991. 2

[10] Robert M Haralick, Stanley R Sternberg, and Xinhua Zhuang. Image analysis using mathematical morphology. *IEEE transactions on pattern analysis and machine intelligence*, PAMI-9:532–550, 1987. 2

[11] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-SR: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1575–1584, 2019. 1, 2, 5

[12] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015. 4

[13] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 3

[14] Heewon Kim, Myungsub Choi, Bee Lim, and Kyoung Mu Lee. Task-aware image downscaling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 399–414, 2018. 1, 2, 3, 4

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016. 2

[16] Jun-Hyuk Kim, Soobeom Jang, Jun-Ho Choi, and Jong-Seok Lee. Instability of successive deep image compression. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 247–255, 2020. 2

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[18] Christiane Lemke, Marcin Budka, and Bogdan Gabrys. Metalearning: a survey of trends and technologies. *Artificial intelligence review*, 44(1):117–130, 2015. 2

[19] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 517–532, 2018. 2

[20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2

[21] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001. 4

[22] Don P Mitchell and Arun N Netravali. Reconstruction filters in computer-graphics. *ACM Siggraph Computer Graphics*, 22(4):221–228, 1988. 2, 4

[23] Thomas Nodes and Neal Gallagher. Median filters: Some modifications and their properties. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 30(5):739–746, 1982. 2

[24] Aleksandr Petiushko. Python implementation of matlab imresize function. `https://github.com/fatheral/matlab_imresize`, 2018. 6

[25] Wanjie Sun and Zhenzhong Chen. Learned image downscaling for upscaling using content adaptive resampler. *IEEE Transactions on Image Processing*, 29:4027–4040, 2020. 1, 2, 3, 4, 5

[26] Longguang Wang, Yingqian Wang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning a single network for scale-arbitrary super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4801–4810, 2021. 1, 2, 5

[27] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[28] Pengxu Wei, Ziwei Xie, Hannan Lu, ZongYuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Proceedings of the European Conference on Computer Vision*, 2020. 3

[29] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In *European Conference on Computer Vision*, pages 126–144. Springer, 2020. 1, 2, 3, 4, 5

[30] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010. 4

[31] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 2, 5

[32] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. 2, 4