Continuous Optical Zooming: A Benchmark for Arbitrary-Scale Image Super-Resolution in Real World

Huiyuan Fu[1] Fei Peng[1] Xianwei Li[1] Yejun Li[1] Xin Wang[2] Huadong Ma[1]
[1]Beijing University of Posts and Telecommunications, China
[2]Stony Brook University
{fhy, pf0607, lixianwei, liyejun2415, mhd}@bupt.edu.cn
x.wang@stonybrook.edu

Abstract

Most current arbitrary-scale image super-resolution (SR) methods has commonly relied on simulated data generated by simple synthetic degradation models (e.g., bicubic downsampling) at continuous various scales, thereby falling short in capturing the complex degradation of real-world images. This limitation hinders the visual quality of these methods when applied to real-world images. To address this issue, we propose the Continuous Optical Zooming dataset (COZ), by constructing an automatic imaging system to collect images at fine-grained various focal lengths within a specific range and providing strict image pair alignment. The COZ dataset serves as a benchmark to provide real-world data for training and testing arbitrary-scale SR models. To enhance the model's robustness against real-world image degradation, we propose a Local Mix Implicit network (LMI) based on the MLP-mixer architecture and meta-learning, which directly learns the local texture information by simultaneously mixing features and coordinates of multiple independent points. The extensive experiments demonstrate the superior performance of the arbitrary-scale SR models trained on the COZ dataset compared to models trained on simulated data. Our LMI model exhibits the superior effectiveness compared to other models. This study is of great significance in developing more efficient algorithms and improving the performance of arbitrary-scale image SR methods in practical applications. Our dataset and codes are available at https://github.com/pf0607/COZ.

1. Introduction

In the field of computer vision, Super-Resolution (SR) has been a prominent area of research [7, 9, 11–13, 15, 19, 20, 22, 25, 27, 34, 35]. It aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) image. Recently, significant progress has been made in arbitrary-scale image SR, primarily based on learning the continuous representation of images. These methods typically require training with continuous fine-grained scale variation LR-HR image pairs within a specific range(i.e., ×1.0-×4.0). [5, 8, 10, 16, 18, 29, 31]

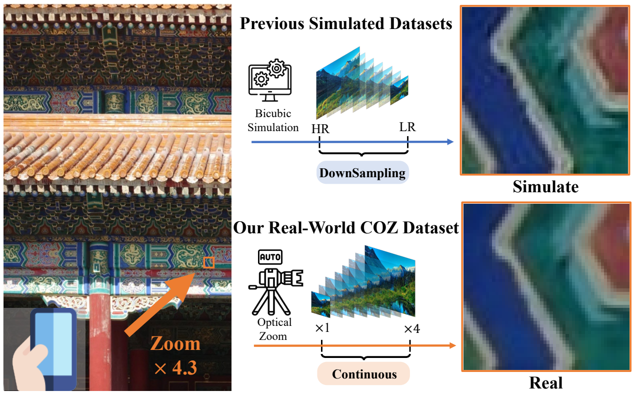


Figure 1. The result exhibits that SR model trained on simulated dataset struggles to address the real-world arbitrary-scale problem, displaying noticeable blurriness and artifacts. In contrast, our approach is more natural and performs comparably to the real-world continuous optical zooming effect.

However, there still remain problems when we apply arbitrary-scale image SR methods to the real-world application. As shown in Fig. 1, one problem is that most current methods are trained and evaluated on several widely used SR datasets including DIV2K [1], Urban100 [17], Manga109 [24], Set5 [3], Set14 [32], and BSD300 [23]. Typically, these datasets apply simple synthetic degradation models (e.g., bicubic downsampling) to obtain data at different resolutions. However, despite that satisfactory results are obtained on simulated data, image degradation in the real world is more complex, resulting in poor visual results on real-world images. Another problem is several real-world image SR datasets have been proposed, including RealSR [4], City100 [6], SR-RAW [33], and DRealSR [30] recently. However, these datasets are limited as they only capture image pairs at fixed magnification scales (e.g., ×2, ×3, ×4), lacking continuous representation of images.

Given the set of issues, we summarize them as the intricate real-world arbitrary-scale image SR problem. Current methods fail to learn a continuous representation of real-world images, resulting in SR outcomes that lack visual naturalness. As shown in Fig. 1, to solve this problem and improve the performance of current arbitrary-scale image SR methods so that the quality can be like optical zoom, we introduce a new dataset - the Continuous Optical Zoom dataset (COZ), as the first real-world dataset for arbitrary-scale image SR. We design and develop a continuous optical zooming imaging system, where optical lens are wirelessly controlled to rotate incrementally and uniformly within a specific focal length. We capture multiple pairs of continuous images from low to high magnification scales of the same scene. Using a two-stage image pair alignment algorithm based on SIFT matching points, we obtain accurately aligned real-world LR-HR image pairs. This dataset provides rich real-world image pairs at various magnification scales for training arbitrary-scale SR models, enabling the learning of continuous image degradation in real-world scenarios. Comparative experimental results demonstrate that models trained on our real-world image data outperform those trained on simulated data when applied to real images.

To enhance the model's robustness against real-world complex image degradation, we propose an arbitrary-scale image SR method based on MLP-mixer [28] architecture and meta-learning [14], named Local Mix Implicit network (LMI). In real world, texture information is manifested in space as multiple coordinates along with their corresponding RGB values. Our method utilizes meta-learning to simultaneously learn multiple local coordinate information and generate mix weights, which are applied to features associated with different coordinates to perform the effective mixing. This is fundamentally different from the previous methods that only consider one coordinate and its feature information at a time, which is susceptible to the interference of complex degradation. Experimental results demonstrate that our approach is effective in learning the continuous representation of real images and requires fewer parameters.

The primary contributions of this work are as follows:

- To our knowledge, this is the first work to address the difficult real-world arbitrary-scale image SR problem. Additionally, we build the first dataset for this task. It can be served as a benchmark for training and testing arbitrary-scale image SR models in real world.
- We propose the Local Mix Implicit network, which simultaneously considers multiple independent point coordinates and features, learning spatial texture information in a mix manner to enhance the robustness against real-world image degradation.
- We conduct extensive experiments to validate the effectiveness of our dataset and the Local Mix Implicit network by comparing our results with those produced by state-of-the-art methods.

Dataset	Conference	Real-World	Arbitrary-Scale	Zoom
DIV2K [1]	CVPRW 2017	✕	✓	-
RealSR [4]	ICCV 2019	✓	✕	Manual
City100 [6]	CVPR 2019	✓	✕	Manual
SR-RAW [33]	CVPR 2019	✓	✕	Manual
DRealSR [30]	ECCV 2020	✓	✕	Manual
Ours (COZ)	-	✓	✓	Automatic

Table 1. Comparison with previous image super-resolution datasets.

2. Continuous Optical Zooming Dataset

We propose a benchmark dataset named Continuous Optical Zooming dataset (COZ), for arbitrary-scale SR methods to learn real-world continuous image representation. We build an automatic continuous optical zooming imaging system to collect data. This system employs a remote control transmission device to incrementally and uniformly rotate the lens within a pre-defined focal length range, capturing images after each rotation. This process facilitates the acquisition of multiple images with fine-grained focal length variations within a specific focal range of the same scene. Subsequently, we apply an improved two-stage Scale-Invariant Feature Transform (SIFT) algorithm [21] to achieve the accurate alignment of images at different resolutions. The comparison between our COZ dataset and other SR datasets is presented in Tab. 1.

2.1. Basic Equipment

We collect data using a Canon EOS R10 camera, which boasts a resolution of 5328×4000 pixels. The camera is equipped with an optical zoom lens spanning a focal length range from 18mm to 150mm. Let's denote the focal length, object distance, and image distance as f, u, and v respectively, and the camera operates under the assumption that $u \gg f$ and v. Considering that the image distance v dictates the actual size of the image, let's contemplate capturing the same object using two distinct focal lengths, f_1 and f_2, along with corresponding object distances v_1 and v_2. The magnification ratio, denoted as M, can be expressed as follows:

$$M \approx \frac{v_1}{v_2} \approx \frac{f_1}{f_2} \tag{1}$$

A small focal length tends to induce distortion issues at the image edges, we opt not to commence image capture directly from an 18mm focal length. Instead, we select a focal length range of 35mm to 140mm for acquiring continuous optical zoom images during the training data collection, encompassing magnification scales from ×1.0 to ×4.0, as calculated using Eq. (1). For the testing data, we choose a focal length range of 25mm to 150mm to capture images with magnification scales ranging from ×1.0 to ×6.0.

2.2. Automatic Continuous Zooming System

Traditional optical lenses necessitate manual rotation for achieving zoom functionality. Frequent physical interaction with the lens can induce angular deviations in the camera,

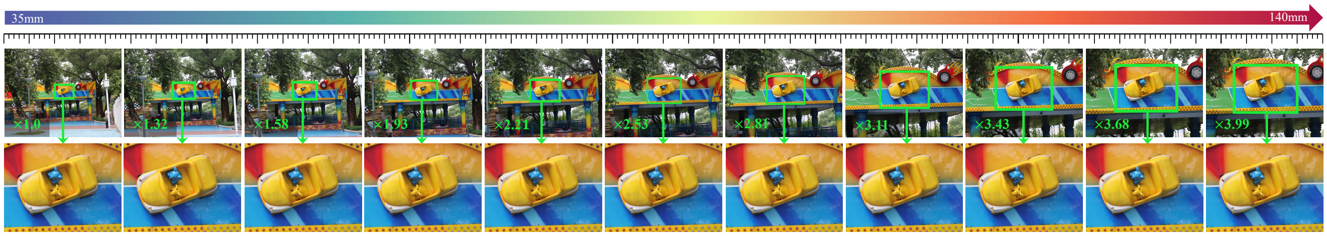


Figure 2. The example sequence of our COZ dataset. The top row shows a sample of 11 images from around 60 images captured within the focal length range of 35mm to 140mm. The second row shows the aligned results after cropping the central regions from these images.

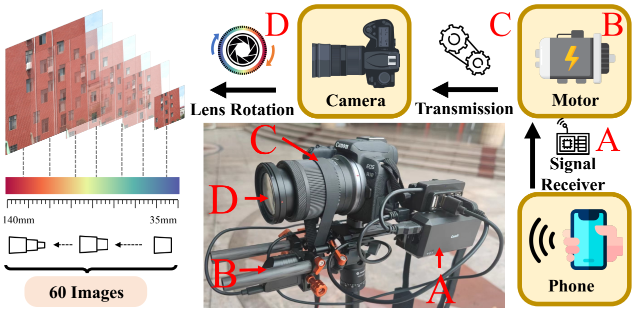
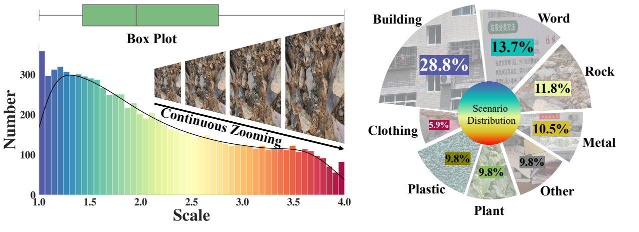


Figure 3. The automatic continuous zooming imaging system we build to collect data. A is the controller, B is the motor, C is the transmission belt, and D is the optical lens.



(a) (b)

Figure 4. Statistics of COZ dataset. (a) is the distribution statistics of images at different magnification scales ranging from ×1.0 to ×4.0 in training dataset. (b) is the scene diversity statistics.

leading to the accumulation of additional errors. We have developed a fully automatic continuous zoom imaging system, as illustrated in Fig. 3. First, we replace the manual lens zooming step with a tightly coupled connection between the lens zoom ring and a transmission belt (C). A precision motor (B), positioned below, rotates the belt, thereby advancing the lens element (D) forward to alter the focal length. The lens zooming is orchestrated by a precision motor, ensuring the maximum stability and accuracy throughout the entire process. The controller receives commands from a smartphone to automatically complete the shooting process. It first directs the motor to rotate the lens within a specific focal length range to record the total travel distance. Then, it divides this total distance into multiple evenly spaced segments, prompting the motor to sequentially move each segment's distance and capture photos.

2.3. Image Pair Alignment

The luminance and resolution variations resulting from lens zoom during the image capture can challenge widely used image alignment algorithms like ORB [26], SURF [2], and SIFT [21]. To tackle this, we propose a two-stage SIFT algorithm. Initially, we adjust the luminance in the first stage. SIFT matching points are gathered from both LR and HR images. Maintaining a consistent quantity of SIFT matching points regardless of resolution enables more precise luminance adjustment. We calculate the RGB standard deviation and mean values of SIFT matching points in LR and HR images as σ_H, σ_L, μ_H, and μ_L, respectively. Using the LR image (I_L), the luminance adjustment formula is applied as:

$$I_L = \frac{\sigma_H}{\sigma_L} I_L + \mu_H - \frac{\sigma_H}{\sigma_L} \mu_L \tag{2}$$

Post luminance adjustment in LR images, SIFT is reapplied to crop corresponding regions for image pairs. The alignment results exhibit precision due to minimal errors introduced during the capture and the adjustment of luminance differences. In Fig. 2, examples of captured images and alignment results from COZ dataset are presented. This figure illustrates a scene with a focal length range of 35mm-140mm, where we uniformly sample 11 images from a total of 60 captured images and their corresponding aligned pairs. The magnification scales relative to the lowest-resolution image are indicated for each image.

2.4. COZ Dataset Detail

The training set of our COZ dataset contains 153 scenes, comprising a total of 9,019 images. The testing dataset includes 37 scenes. As current arbitrary-scale image SR methods typically evaluate at specific fixed magnification scales, we specifically select images with the magnification scales closest to certain scales (×2.0, ×2.5, ×3.0, ×3.5, ×4.0, ×5.0, ×5.5, ×6.0) for the testing data. Our focus during the scene acquisition is to ensure the diversity by capturing objects with rich textures in both indoor and outdoor living scenes. We exclude scenes featuring moving objects. A small fraction of the data has a maximum magnification scale lower than ×4.0 (minimum being ×3.6) due to capturing limitations (e.g., friction loss). As we uniformly rotate the lens during the capture, the magnification scale variation corresponding to the focal length variation is not truly uniform, leading to a lesser number of HR images compared to LR images. The distribution of magnification scales for all the images in the training set is illustrated in Fig. 4.

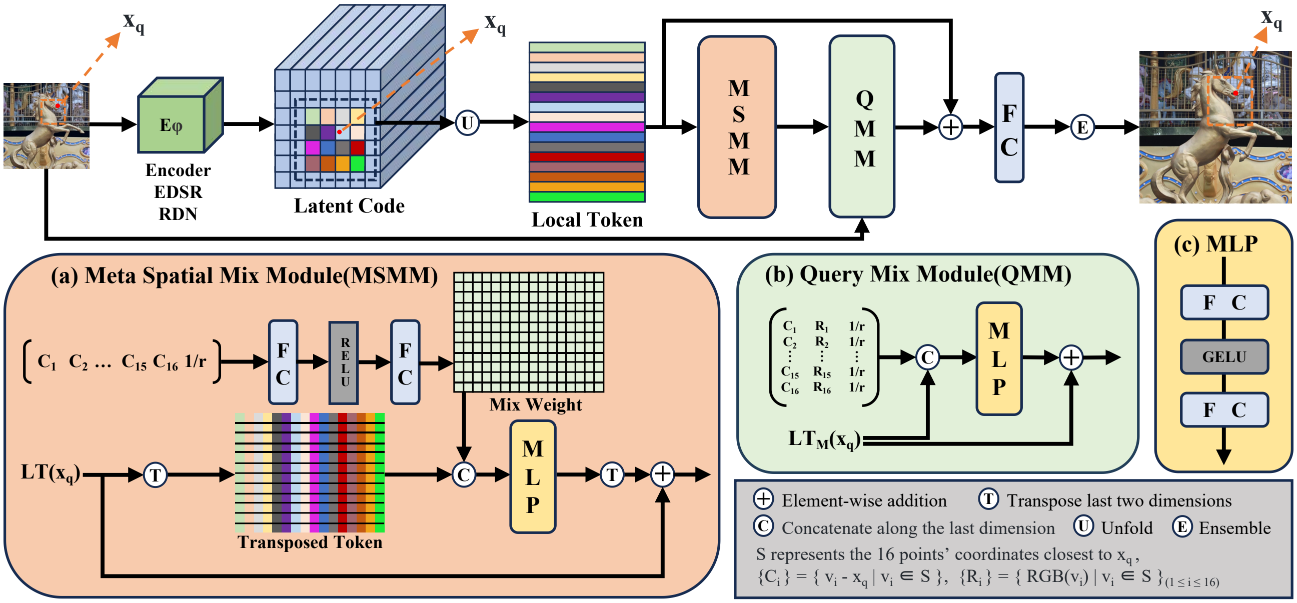

x_q
Eφ
Encoder
EDSR
RDN
x_q
Latent Code
U
Local Token
MSMM
QMM
FC
E
x_q
(a) Meta Spatial Mix Module(MSMM)
C₁ C₂ ... C₁₅ C₁₆ 1/r
FC
RELU
FC
Mix Weight
LT(x_q)
T
Transposed Token
C
MLP
T
(b) Query Mix Module(QMM)
C₁ R₁ 1/r
C₂ R₂ 1/r
C₁₅ R₁₅ 1/r
C₁₆ R₁₆ 1/r
C
MLP
LT_M(x_q)
(c) MLP
F C
GELU
F C
Element-wise addition
T Transpose last two dimensions
C Concatenate along the last dimension
U Unfold
E Ensemble
S represents the 16 points' coordinates closest to x_q ,
{C_i} = { v_i - x_q | v_i ∈ S }, {R_i} = { RGB(v_i) | v_i ∈ S }(1 ≤ i ≤ 16)

Figure 5. The proposed Local Mix Implicit Network framework.

3. Method

Recent arbitrary-scale image SR methods [5, 8, 10, 18, 29, 31] generally embrace an approach centered around constructing implicit functions for learning continuous image representation. Denoting a continuous image as I and coordinates within it as x. LR images are processed via commonly used encoders such as EDSR [20] and RDN [35] to extract latent codes Z, which are subsequently utilized in constructing a decoding implicit function f. The expression for SR prediction typically follows this form:

$$I(x) = f(Z, x) \tag{3}$$

For a specific query point x_q, assuming V^* is the coordinate nearest to x_q, and Z^* is the latent code corresponding to V^*, the RGB prediction formula for x_q can be articulated as:

$$I(x_q) = f(Z^*, V^* - x_q) \tag{4}$$

These approaches typically focus on individual coordinates and their corresponding latent codes in isolation. When applied to simulated data generated through simple linear synthetic degradation model, they demonstrate proficiency, as the encoder adeptly encodes local area information into the latent code. However, real-world image degradation is notably more complex, insufficient reference information such as one coordinate and latent code can easily lead to an unstable result.

In the case of constructing the texture information in the real world, texture is spatially manifested through multiple coordinates, each with its corresponding RGB values. Hence, considering multiple coordinates and their corresponding features within local regions simultaneously provides a direct means of capturing texture information.

3.1. Local Mix Implicit Network

This study introduces the Local Mix Implicit Network (LMI), an advanced model structure depicted in Fig. 5. Based on the mlp-mixer [28] architecture, LMI is crafted to adeptly learn complex texture information by simultaneously mixing multiple coordinates and their corresponding latent codes. Commencing with the extraction of numerous latent codes from local regions, each is treated as a token with its coordinates preserved. These tokens collectively form the foundational spatial information. LMI encompasses two stages of mix modules, illustrated in Fig. 5 (a) and (b).

The Meta Spatial Mix Module (MSMM), built on top of a meta-learning [14] network, transforms multiple coordinate information to mix weights for guiding the mixing of latent codes, facilitating the capture of spatial texture details. The Query Mix Module (QMM) concentrates on internal mixing within latent codes, embedding the original RGB value and coordinate into corresponding tokens as queries. In the final step, the results predicted from each token are ensemble to enhance the overall robustness.

3.2. Local Token Unfolding

To acquire sufficient spatial information for texture capture, we extract the latent codes of the 4×4 region closest to the query point x_q, denoting them as $\{Z_i^*\}$ with $1 \le i \le 16$. We maintain the independence of latent codes, treating each as an autonomous token. The tokens go through an unsqueezing operation and are concatenated along the extended dimension. Let Λ represent the concatenation operation. We denote these local tokens as $LT(x_q)$ and define them as follows:

$$LT(x_q) = \Lambda\{\text{unsqueeze}(Z_i^*)\} \quad (1 \le i \le 16) \tag{5}$$

Moreover, to appropriately learn local region information, we utilize relative coordinates. The coordinates of each token are defined as $\{V_i^*\}$ with $1 \leq i \leq 16$, and the relative coordinates of each token concerning the query coordinate x_q are defined as $\{C_i\}$ with $1 \leq i \leq 16$. C_i is defined as:

$$C_i = V_i - x_q \quad (1 \leq i \leq 16) \tag{6}$$

3.3. Meta Spatial Mix Module

In order to extract spatial texture information from multiple local tokens, we introduce the mixing operation between tokens. We employ a Multi-Layer Perception(MLP) for the token mixing and interaction, as illustrated in Fig. 5 (c). We transpose $LT(x_q)$, pass them through the MLP for mixing, and then transpose the result back. Let MLP_s represent the MLP used for spatial mixing, the mixed local tokens $LTM(x_q)$ is defined as:

$$LTM(x_q) = (MLP_s(LT(x_q)^T))^T \tag{7}$$

However, if we directly perform the mixing operation between tokens, while enhancing the information of each token, the local spatial relationship between tokens will be overlooked. To address this issue, we employ a straightforward approach which concatenates each relative coordinate C_i with the transposed tokens and subsequently performs the mixing. We repeat and expand coordinates to match the shape of the transposed tokens and concatenate them with the transposed tokens. Let E be the expand operation, Eq. (7) can be improved as:

$$LTM(x_q) = (MLP_s(\Lambda(LT(x_q)^T, \{C_i^E\})))^T \tag{8}$$

The mix network simultaneously learns coordinate information and mixes tokens, reducing the efficiency of the network. In Fig. 5 (a), we employ a meta-learning approach using an independent network to learn the spatial coordinate information and construct the spatial mix weight with the same shape as all tokens. We denote the mix weight as W and compute it through several fully connected layers. The weight calculation network is represented as ω, and we introduce a scale factor r to enhance the accuracy of spatial information learning. The expression for the weight is as follows:

$$W = \omega(\{C_i\}, 1/r) \tag{9}$$

The mix weight W is then concatenated with $LT(x_q)$ and input into the mix network MLP_s, enabling the network to focus on the mixing between tokens and acquire sufficient local spatial texture information. The final expression for $LTM(x_q)$ is defined as:

$$LTM(x_q) = (MLP_s(\Lambda(LT(x_q)^T, W)))^T \tag{10}$$

3.4. Query Mix Module

Following the spatial mixing, each token acquires the local texture information, enhancing its capability to provide the improved guidance for predicting the RGB value of x_q. In this stage, we incorporate the coordinate information C_i for decoding. Given that SR involves a task transitioning from one image to another, the original RGB information in the image exhibits a strong correlation with the predicted RGB values. Since each token corresponds directly to an image coordinate, we introduce a modified form of the original image's "residual connection" and embed the RGB values of the corresponding coordinates V_i^* from the input image to complement the token information. We denote the MLP used for query mixing as MLP_q, RGB values for coordinates V_i^* as R_i^*, and the query mixed tokens as $LTQ(x_q)$. Its expression is formulated as follows:

$$LTQ(x_q) = MLP_q(LTM(x_q), \{C_i, R_i^*\})_{(1 \leq i \leq 16)} \tag{11}$$

3.5. Ensemble

After two stages of mixing, the query mixed tokens $LTQ(x_q)$ are input into a fully connected layer for output. Due to the assimilation of spatial texture information through token mixing, each token comprehends valuable guidance for accurately predicting the value of x_q. In a manner reminiscent of LIIF's local ensemble approach [10], we compute the RGB value at coordinate x_q by directly ensembling outputs of each tokens with the weight calculated with respect to the area of the rectangle between x_q and V_i^*.

4. Experiments

Datasets. The COZ dataset serves as a reference for the supervised training. LR-HR image pairs are selected from a set of continuous images captured in the same scene. The widths of the LR and HR images are represented as W_L and W_H, respectively. The scale factor, denoted as s, is computed as W_H/W_L.

Implementation Details. We adhere to the experimental configurations established in previous studies [10, 16, 18]. Our method involves utilizing L1 loss and the Adam optimizer, with the encoder being either EDSR-baseline [20] or RDN [35]. The size of the input image patch are fixed at 48×48. Specifically, we sample 48^2 pixels from both high-resolution (HR) images and their corresponding coordinates. Despite the diminished data volume compared to the simulated dataset [1], we train all models for 300 epochs. The learning rate is initially set at 1e-4, with a decay to 0.5 at the 200-th epoch. The batch size is established at 16. In the case of other arbitrary-scale super-resolution (SR) methods [5, 8, 10, 16, 18, 29, 31], we maintain the original experimental configurations. The evaluation metric for all models is mainly Peak Signal-to-Noise Ratio (PSNR). We train all the models on a RTX3090 GPU and test them on a RTX A40 GPU.

Methods	Params	EDSR-baseline [20]								RDN [35]							
		In-scale					Out-of-scale			In-scale					Out-of-scale		
		×2	×2.5	×3	×3.5	×4	×5	×5.5	×6	×2	×2.5	×3	×3.5	×4	×5	×5.5	×6
MetaSR [16]	445.1K	28.70	27.43	26.55	25.62	25.17	24.31	23.93	23.25	28.80	27.55	26.65	25.80	25.22	24.39	24.09	23.31
LIIF [10]	346.9K	28.72	27.57	26.61	25.76	25.16	24.32	24.01	23.23	28.80	27.56	26.69	25.83	25.23	24.39	24.13	23.28
LTE [18]	493.8K	28.67	27.49	26.55	25.71	25.15	24.37	24.05	23.26	28.72	27.57	26.64	25.74	25.17	24.40	24.10	23.28
LINF [31]	794.9K	28.72	27.48	26.53	25.66	25.10	24.29	23.99	23.21	28.73	27.55	26.60	25.73	25.15	24.32	24.03	23.28
SRNO [29]	705.2K	28.73	27.54	26.59	25.70	25.15	24.31	24.05	23.25	28.74	27.60	26.67	25.73	25.19	24.40	24.09	23.28
LIT [8]	5.3M	28.74	27.56	26.58	25.71	25.16	24.35	24.00	23.19	28.80	27.63	26.66	25.79	25.19	24.36	24.03	23.25
LMI (ours)	**87.9K**	**28.86**	**27.63**	**26.66**	**25.78**	**25.22**	**24.39**	**24.08**	**23.29**	**28.86**	**27.68**	**26.74**	**25.86**	**25.30**	**24.48**	**24.14**	**23.37**

Table 2. Quantitative comparison with state-of-the-art methods for **arbitrary-scale image SR** on the **COZ** testing set (PSNR (dB)).

Selection	Methods	Scale				
		×2	×3	×4	×5	×6
Fixed	MetaSR [16]	28.76	26.53	25.07	24.21	23.11
	LIIF [10]	28.75	26.51	25.07	24.23	23.08
	LTE [18]	28.65	26.50	25.07	24.23	23.13
	LMI (Ours)	28.77	26.57	25.16	24.32	23.25
Random	MetaSR [16]	28.70	26.55	25.17	24.31	23.25
	LIIF [10]	28.72	26.61	25.16	24.32	23.23
	LTE [18]	28.67	26.55	25.15	24.37	23.26
	LMI (Ours)	28.86	26.66	25.22	24.39	23.29

Table 3. Quantitative comparison (PSNR (dB)) on our COZ testing set by methods trained with different HR image selection methods.

Degradation	Methods	Scale		
		×2	×3	×4
BD	MetaSR [16]	28.21 / 0.764	26.14 / 0.719	24.76 / 0.699
	LIIF [10]	28.24 / 0.765	26.17 / 0.720	24.79 / 0.699
	LTE [18]	28.21 / 0.764	26.14 / 0.719	24.76 / 0.699
	LMI (Ours)	28.35 / 0.778	26.23 / 0.732	24.83 / 0.710
Real	MetaSR [16]	28.70 / 0.809	26.55 / 0.762	25.17 / 0.736
	LIIF [10]	28.72 / 0.809	26.61 / 0.762	25.16 / 0.736
	LTE [18]	28.67 / 0.812	26.55 / 0.765	25.15 / 0.738
	LMI (Ours)	28.86 / 0.812	26.66 / 0.762	25.22 / 0.736

Table 4. Quantitative comparison (PSNR (dB)/SSIM) on our COZ testing set with methods trained on different datasets generated by bicubic downsampling (BD) and real degredation.

4.1. Quantitative Experiments

Randomly Selecting Training Strategy. The COZ dataset offers the flexibility in choosing the LR-HR image pairs. We introduce a more adaptable training strategy by randomly selecting two images of different resolutions from a set to create LR-HR pairs, as opposed to opting for the fixed highest-resolution image as the HR image. Three representative state-of-the-art (SOTA) models, namely MetaSR [16], LIIF [10], LTE [18], and our proposed LMI, each employing EDSR-baseline [20] as the encoder, are utilized. These models are trained using various HR image selection methods, resulting in a total of 8 models. Subsequently, all 8 models undergo testing on the COZ testing set, and the results are presented in Tab. 3. The models that select the fixed highest-resolution image as the HR image exhibit the superior performance at the low scale (×2) but experience a pronounced decline at higher scales (×3, ×4, ×5, and ×6).

COZ dataset vs. Simulated dataset. To evaluate the efficacy of the COZ dataset, we train three representative SOTA methods along with our proposed method on both simulated dataset and our real dataset. We designate the highest-resolution data from each scene within the COZ dataset as the ground truth (GT) and employ bicubic downsampling (BD) to generate simulated data. Structural Similarity (SSIM) evaluation metric is additionally incorporated. The results are presented in Tab. 4. Notably, the models trained on the COZ real-world data exhibit substantial improvements on both PSNR and SSIM. This indicates the effectiveness of COZ to capture the continuous real-world image degradation.

Models trained on COZ dataset. To showcase the effectiveness of our LMI model, we undertake a comparative study on the COZ dataset, assessing its performance alongside six SOTA models: MetaSR [16], LIIF [10], LTE [18], LINF [31], LIT [8], and SRNO [29]. Each model is trained using two encoders, namely EDSR-baseline [20] and RDN [35].

Comprehensive testing is conducted on the COZ testing set. Following the conventional protocol for arbitrary-scale SR experiments, the COZ dataset encompasses two types of testing scales: within the training scales (×2, ×2.5, ×3, ×3.5, ×4) and outside the training scales (×5, ×5.5, ×6). The experimental results are detailed in Tab. 2. Our LMI method demonstrates noteworthy improvements with fewer parameters compared to other methods, establishing its suitability for real-world arbitrary-scale image SR.

4.2. Qualitative Experiments

COZ testing set. In Fig. 6, we present visualizations of the testing outcomes generated by models trained on both BD-simulated and real (COZ) datasets. It is apparent that models trained on real-world data yield clearer and more natural results, in contrast to models trained on simulated data, which exhibit pronounced blurriness and artifacts. It is crucial to emphasize that even for simple objects, such as balloons, models trained on simulated data introduce noticeable noise-like blurring and lack a sense of "realism". Consequently, training with real-world captured continuous optical zoom images facilitates the development of SR models that produce more realistic results.

Generalizability. To demonstrate the generalization capabilities of models trained on our real dataset, we captured images outside of our dataset using the Sony RX100M4 digital camera, Huawei Mate 40 Pro, and iPhone XS smartphone cameras. Considering the perspective of optical zoom, we selected various magnification scales for the images based on visual perception. Subsequently, we applied bicubic interpolation, as well as the SOTA SRNO [29] method trained on

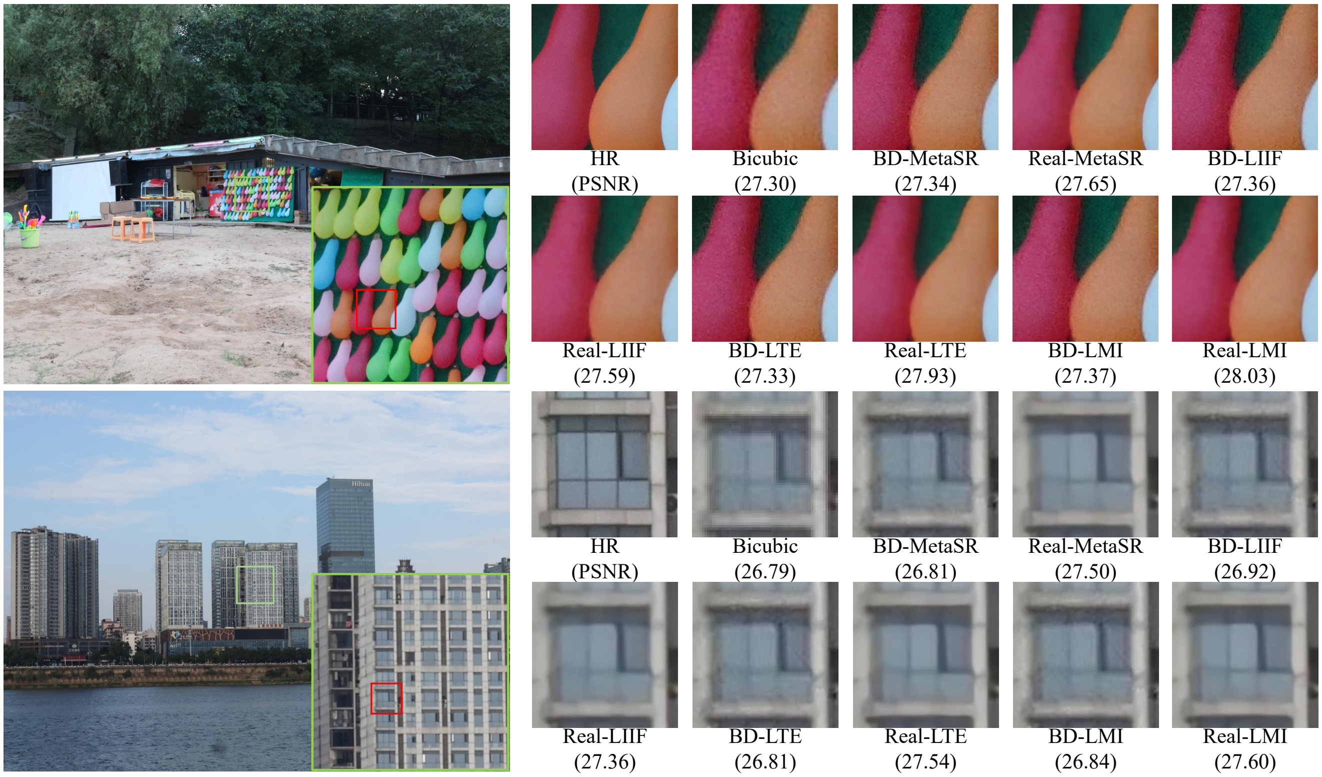


Figure 6. Visual SR results on our COZ testing set by different methods (trained on different datasets). The results above are for a × 4scale, while the results below are for a ×3 scale. Methods include MetaSR [16], LIIF [10], LTE [18] and **LMI (ours)**. It can be easily observed that even simple objects like balloons, when models are trained on simulated data, produce noticeable artifacts.

both BD simulated data and real data (COZ), and our LMI method trained on real data. The visual results are depicted in Fig. 7. Notably, models trained on real data exhibit consistent high performance across different devices, while those trained on simulated data display discernible blurriness and artifacts. This indicates COZ dataset's practical significance in improving digital zoom quality on various devices.

4.3. Ablation Experiments

MSMM. To validate the guiding function of meta-learning for multi-coordinate feature mixing, we conduct experiments by eliminating the meta-learning component and directly employing the multi-coordinate information embedded features from Eq. (8) to guide the mixing (referred to as LMI-a). Additionally, to demonstrate the effectiveness of multi-coordinate mixing, we perform experiments solely involving multi-feature mixing without coordinates (referred to as LMI-b) as outlined in Eq. (7), and we directly remove the MSMM module (referred to as LMI-c). The experimental results, as presented in Tab. 5, show a gradual decrease compared to LMI. This suggests that meta-learning-guided multi-coordinate feature mixing is well-suited for learning the continuous image degradation in real-world scenarios.

QMM. We exclude the embedding of LR's RGB values in the QMM module (referred to as LMI-d). When comparing LMI-d with LMI in Tab. 5, a performance decrease is observed, suggesting that the inclusion of RGB information from the LR image serves as an effective addition for real-world SR.

Methods	Scale				
	×2	×3	×4	×5	×6
LMI	28.86	26.66	25.22	24.39	23.29
LMI-a	28.77	26.58	25.15	24.34	23.23
LMI-b	28.74	26.55	25.13	24.32	23.20
LMI-c	28.67	26.51	25.11	24.31	23.18
LMI-d	28.81	26.64	25.22	24.37	23.27

Table 5. Quantitative ablation study of LMI. Evaluated on the COZ testing set (PSNR (dB)). "-a/b/c" refers to replacing meta-learning with direct coordinate embedding, removing meta-learning of MSMM, and removing MSMM, respectively. "-d" refers to removing the RGB embedding of QMM.

4.4. User Study

We conduct two user studies, comparing the visual quality of the LMI method with other methods, and comparing models trained on simulated data generated by bicubic downsampling with models trained on real-world data. For each study, we randomly select 30 real life images captured by Sony RX100M4, Huawei Mate 40 Pro, and iPhone XS, and ask 20 participants to rate each image. The SR scale is ×4 and

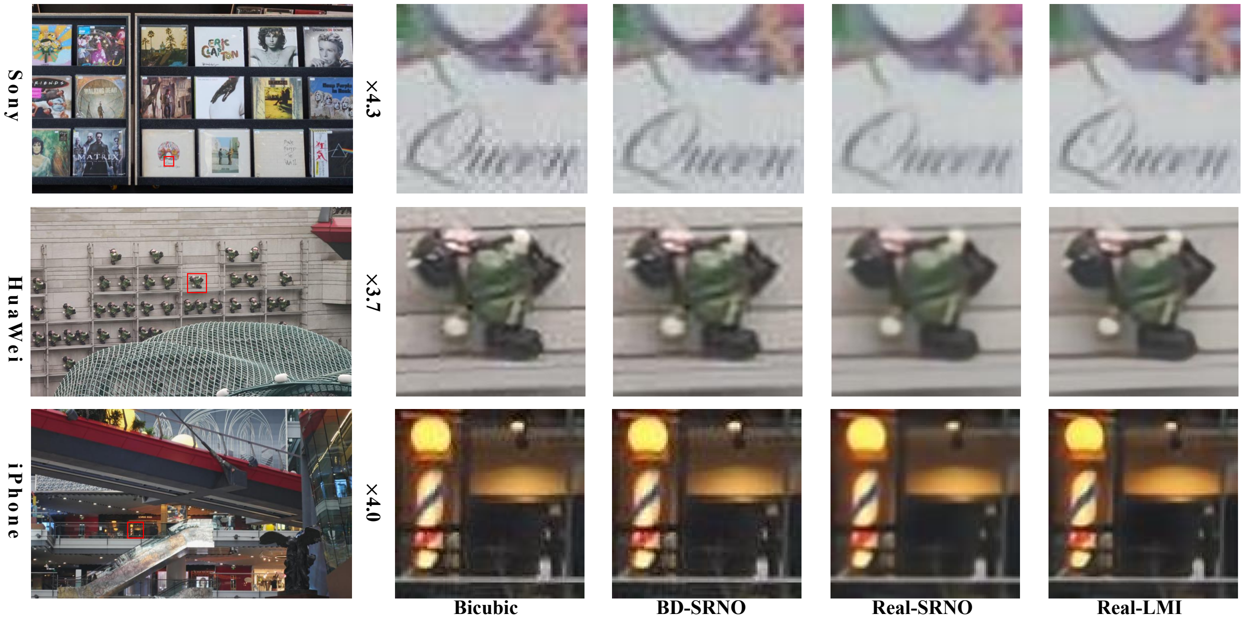
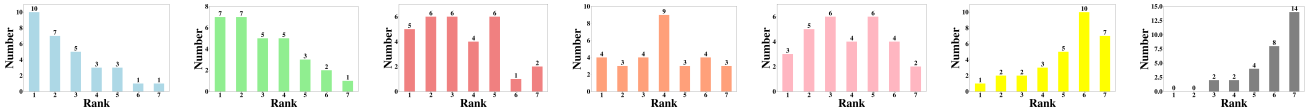


Figure 7. Visual SR results of different methods(trained on different datasets) on images captured by Sony RX100M4 digital camera, Huawei Mate 40 Pro and iPhone XS smartphone cameras. Methods include bicubic interpolation, SRNO [29] and **LMI (ours)**.



(a) LMI (b) LIT [8] (c) SRNO [29] (d) LTE [18] (e) LINF [31] (f) LIIF [10] (g) MetaSR [16]

Figure 8. The first user study result. In each histogram, the x-axis denotes the ranking index (1-7, "1" represents the highest), and the y-axis denotes the number of images in each ranking index.

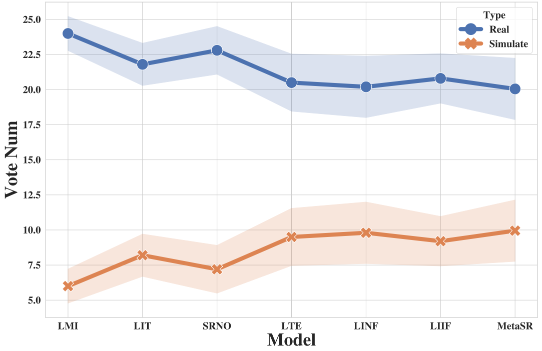


Figure 9. The second user study result. The region included for each data point, representing the variability or uncertainty associated with the votes for each model under different data types.

participants are asked to assess image quality, considering non-realistic artifacts and blurred edges.

For the first user study, participants are asked to rank 7 images generated by six previous methods [8, 10, 16, 18, 29, 31] and our LMI method each time. The average ranking for each method on each image is calculated and used to generate the final rankings. The results are presented in Fig. 8. Comparing the histograms reveals that our method produces superior results in terms of human subjective evaluations compared to the other methods. For the second user study, participants are asked to vote 2 images generated by methods trained on simulated data and our COZ real data each time. The results are presented in Fig. 9, where all the methods trained on real data receive more votings compared with simulated data, revealing that our COZ data could improve the visual results of arbitrary-scale image SR methods.

5. Conclusion

We introduce COZ, the first real-world dataset for arbitrary-scale image SR. Captured using our automatic continuous zooming imaging system, COZ offers accurately aligned continuous resolution change image pairs. Leveraging MLP-mixer and meta-learning, we propose the LMI model, which simultaneously considers multiple independent coordinates and corresponding features, learning spatial texture information in a mixed manner. Extensive experiments and user studies validate the effectiveness of our dataset and method, and the results surpass those from the SOTA approaches.

ACKNOWLEDGMENTS

This work is supported in part by the NSFC under No.62272059, the National Key R&D Program of China under No.2023YFF0904800, the Beijing Nova Program under No.20230484406, the Innovation Research Group Project of NSFC (61921003), and the 111 Project (B18008).

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. 1, 2, 5

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer, 2006. 3

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 1

[4] 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/CVF International Conference on Computer Vision*, pages 3086–3095, 2019. 1, 2

[5] Jiezhang Cao, Qin Wang, Yongqin Xian, Yawei Li, Bingbing Ni, Zhiming Pi, Kai Zhang, Yulun Zhang, Radu Timofte, and Luc Van Gool. Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1796–1807, 2023. 1, 4, 5

[6] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1652–1660, 2019. 1, 2

[7] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 1

[8] Hao-Wei Chen, Yu-Syuan Xu, Min-Fong Hong, Yi-Min Tsai, Hsien-Kai Kuo, and Chun-Yi Lee. Cascaded local implicit transformer for arbitrary-scale super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18257–18267, 2023. 1, 4, 5, 6, 8

[9] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023. 1

[10] 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, 4, 5, 6, 7, 8

[11] Haram Choi, Jeongmin Lee, and Jihoon Yang. N-gram in swin transformers for efficient lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2071–2081, 2023. 1

[12] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18470–18479, 2022.

[13] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burstormer: Burst image restoration and enhancement transformer. *arXiv preprint arXiv:2304.01194*, 2023. 1

[14] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 2, 4

[15] Simon Grosche, Andy Regensky, Jürgen Seiler, and André Kaup. Image super-resolution using t-tetromino pixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9989–9998, 2023. 1

[16] 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, 5, 6, 7, 8

[17] 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. 1

[18] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1929–1938, 2022. 1, 4, 5, 6, 7, 8

[19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1

[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. 1, 4, 5, 6

[21] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2, 3

[22] Salma Abdel Magid, Yulun Zhang, Donglai Wei, Won-Dong Jang, Zudi Lin, Yun Fu, and Hanspeter Pfister. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4288–4297, 2021. 1

[23] 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*, pages 416–423. IEEE, 2001. 1

[24] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76:21811–21838, 2017. 1

[25] Seung Ho Park, Young Su Moon, and Nam Ik Cho. Perception-oriented single image super-resolution using optimal objective estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1725–1735, 2023. 1

[26] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 3

[27] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 1

[28] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 2, 4

[29] Min Wei and Xuesong Zhang. Super-resolution neural operator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18247–18256, 2023. 1, 4, 5, 6, 8

[30] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 101–117. Springer, 2020. 1, 2

[31] Jie-En Yao, Li-Yuan Tsao, Yi-Chen Lo, Roy Tseng, Chia-Che Chang, and Chun-Yi Lee. Local implicit normalizing flow for arbitrary-scale image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1776–1785, 2023. 1, 4, 5, 6, 8

[32] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012. 1

[33] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2019. 1, 2

[34] 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. 1

[35] 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. 1, 4, 5, 6