Uncertainty-Aware Source-Free Adaptive Image Super-Resolution with Wavelet Augmentation Transformer

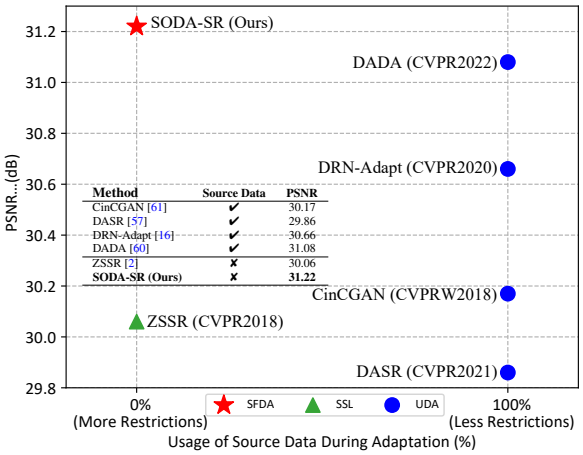
Yuang Ai[1,2] Xiaoqiang Zhou[1,3] Huaibo Huang[1,2*] Lei Zhang[4,5] Ran He[1,2,6]
[1]MAIS & CRIPAC, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]University of Science and Technology of China [4]OPPO Research Institute
[5]The Hong Kong Polytechnic University [6]ShanghaiTech University
shallowdream555@gmail.com, xq525@mail.ustc.edu.cn,
huaibo.huang@cripac.ia.ac.cn, cslzhang@comp.polyu.edu.hk, rhe@nlpr.ia.ac.cn

Abstract

Unsupervised Domain Adaptation (UDA) can effectively address domain gap issues in real-world image Super-Resolution (SR) by accessing both the source and target data. Considering privacy policies or transmission restrictions of source data in practical scenarios, we propose a SOurce-free Domain Adaptation framework for image SR (SODA-SR) to address this issue, i.e., adapt a source-trained model to a target domain with only unlabeled target data. SODA-SR leverages the source-trained model to generate refined pseudo-labels for teacher-student learning. To better utilize pseudo-labels, we propose a novel wavelet-based augmentation method, named Wavelet Augmentation Transformer (WAT), which can be flexibly incorporated with existing networks, to implicitly produce useful augmented data. WAT learns low-frequency information of varying levels across diverse samples, which is aggregated efficiently via deformable attention. Furthermore, an uncertainty-aware self-training mechanism is proposed to improve the accuracy of pseudo-labels, with inaccurate predictions being rectified by uncertainty estimation. To acquire better SR results and avoid overfitting pseudo-labels, several regularization losses are proposed to constrain target LR and SR images in the frequency domain. Experiments show that without accessing source data, SODA-SR outperforms state-of-the-art UDA methods in both synthetic→real and real→real adaptation settings, and is not constrained by specific network architectures.

1. Introduction

Single image super-resolution (SISR), which is a fundamental task in low-level vision, aims to reconstruct a high-resolution (HR) image from its low-resolution (LR) counterpart. In recent years, owing to the thriving advancements in deep learning, numerous deep learning-based approaches have been applied to SISR, culminating in significant breakthroughs in this task. Predominantly, these methods employ Convolutional Neural Networks (CNNs) [8, 31] or Vision Transformers (ViTs) [9, 37] as their architectural foundation. However, the majority are trained on synthetic datasets that generate LR images using simplistic and predetermined degradation kernels (*e.g.*, bicubic).



Method	Source Data	PSNR
CinCGAN [61]	✔	30.17
DASR [57]	✔	29.86
DRN-Adapt [16]	✔	30.66
DADA [60]	✔	31.08
ZSSR [2]	✘	30.06
SODA-SR (Ours)	✘	**31.22**

Figure 1. PSNR vs. the usage of source data on the DRealSR [56] dataset. The less source data a method uses, the more restrictions it faces. SFDA and SSL represent source-free domain adaption and self-supervised learning methods respectively.

However, the domain gap between the predetermined degradation and real-world degradation often leads to poor generalization capability of the SR model in real-world scenarios. To address this issue, several unsupervised domain adaptation (UDA) methods [16, 25, 26, 52, 57, 61] have been proposed to adapt the model from the source domain with synthetic image pairs to the target domain with unlabeled real images (*i.e.*, synthetic→real adaptation). Alternatively, some real-world datasets have been collected to

*Corresponding author. Project Page.

train and evaluate real-world SISR methods, such as RealSR [4] and DRealSR [56]. These datasets include LR and HR image pairs captured on the same scene taken by different cameras. In the meanwhile, there exists a significant gap between degradation kernels for images captured by different cameras, which can be regarded as a cross-device domain gap. [60] found that this kind of domain gap is harmful to the model's performance, and thus proposed a UDA method to adapt model from the source domain with paired real images to the target domain with unlabeled real images (*i.e.*, real→real adaptation).

Though these UDA methods have achieved promising results in synthetic→real and real→real adaptation tasks, they do have certain limitations. Firstly, all these methods utilize source data to retain source knowledge and relieve domain shift during adaptation, which is often inaccessible due to privacy policies or transmission restrictions in practical scenarios. Besides, most of these methods are designed for specific SR network architectures and cannot be easily transferred to other networks, thus lacking generalizability.

In this paper, we propose and attempt to address a new and practical issue, namely Source-Free domain adaptation for image Super-Resolution (SFSR). It aims to adapt a model pre-trained on labeled source data to a target domain with only unlabeled target data. Recently, several Source-Free Domain Adaptation (SFDA) methods have been proposed to address similar challenges in image classification [32, 36, 38, 40], semantic segmentation [12, 41, 42] and object detection [33, 34, 51]. These methods are designed specifically for classification tasks with focus on obtaining reliable pseudo-labels or generating samples similar to the source domain distribution. However, when it comes to pixel-wise regression tasks such as image SR, which do not involve the notion of classes, these techniques are not directly applicable.

To address this issue, we present a novel method named SODA-SR, which is the first SOurce-free Domain Adaptation framework for image SR. Motivated by [3, 47, 49] adding appropriate perturbations to the input (*e.g.*, noise, data augmentation) or feature space (*e.g.*, dropout [48], stochastic depth [22]) of the student model for better teacher-student learning in semi-supervised image classification, we adopt the teacher-student framework and apply the strategy of pseudo-labeling for optimization of the student model. However, existing perturbation methods designed for classification tasks cannot be directly used for SFSR, which may impact the performance of SR models. In light of this issue, we propose two distinct augmentation methods suitable for SR targeting the input and feature levels, respectively. Firstly, we flip and rotate the target LR image to generate seven geometrically augmented images for pseudo-label refinement. Furthermore, we propose a novel Wavelet Augmentation Transformer (WAT) to implicitly generate augmented data, which can be flexibly incorporated with existing networks. By performing a multi-level wavelet decomposition for latent features, WAT learns low-frequency information of varying levels across diverse samples. It performs a proposed Batch Augmentation Attention (BAA) at different levels to mix image features batch-wisely and efficiently fuses these features through deformable attention [72]. WAT enables the student model with the ability to learn and explore appropriate augmentation in the feature space, which facilitates the student model in acquiring robust features.

Beyond that, an uncertainty-aware self-training mechanism is proposed to improve the accuracy of pseudo-labels by transferring knowledge from the target data to the teacher model. Specifically, the teacher model is updated with an exponential moving average of the student model to produce pseudo-labels. For one LR input, the teacher model runs multiple times to obtain the mean and variance as uncertainty estimation, which are then used to rectify pseudo-labels. Finally, we introduce several regularization losses to constrain the frequency information between target LR and SR images. These loss functions effectively prevent the student model from overfitting pseudo-labels by mining frequency information in target LR images, which leads to better SR results. As shown in Fig. 1, SODA-SR successfully generalizes the pre-trained source model on the target domain and achieves better PSNR against existing UDA and self-supervised methods.

The main contributions can be summarized as follows:

- We propose a novel SODA-SR framework to address the SFSR problem. To the best of our knowledge, this is the first research on SFSR.
- We present a wavelet augmentation transformer (WAT) to implicitly synthesize augmented data. WAT learns cross-level low-frequency information of varying levels across diverse samples effectively and improves the robustness of the student model.
- An uncertainty-aware self-training mechanism is introduced to improve the accuracy of pseudo-labels. Inaccurate predictions are rectified by uncertainty estimation.
- Extensive experiments show that our source-free SODA-SR outperforms state-of-the-art UDA methods and is not constrained by specific network architectures.

2. Related Work

2.1. Single Image Super-Resolution

With the rapid and dramatic development of deep learning, more and more SISR models have been proposed and yielded state-of-the-art performance among diverse datasets. Dong *et al*. [8] presented an approach to learning the mapping function from LR images to HR images just using three convolutional layers. After that, a

mass of CNN-based architectures [10, 17, 18, 23, 24, 29–31, 39, 53, 59, 67–69] with more elaborate modules were proposed to improve the SISR performance. Recently, Transformer-based methods [6, 7, 35, 37, 55, 62, 64, 70, 71] were proposed for low-level vision tasks to utilize the great capability to model long-range dependency of Vision Transformer [9]. Some other works [1, 11, 28, 43, 44, 54] adopted diffusion models [20] to generate highly realistic images for image restoration.

2.2. Real-World Image Super-Resolution

Nowadays, in order to circumvent the limitations resulted from synthetic datasets, real-world SR has attracted more and more attention in the community. Some real-world SR datasets [4, 5, 56, 66] have been collected to train and evaluate real-world SR methods. Among them, DRealSR [56] stands as the singular real-world SR dataset encompassing multiple cameras and explicitly specifying the originating camera for each image.

In fact, there are relatively few UDA methods available for real-world SR. CinCGAN [61] adopted a cycle-in-cycle network structure based on GAN to map real-world LR images to a noise-free space. DRN-Adapt [16] utilized paired synthetic data and unpaired real-world data to achieve adaptation with a dual regression constraint. DASR [57] addressed the domain gap between training data and testing data with domain-gap aware training. [60] firstly explored the cross-device real-world SR and proposed an unsupervised mechanism to address this issue. However, all these methods need to access the source data and most of them are designed for specific SR network architectures.

2.3. Source-Free Domain Adaptation

Recently, plenty of methods have been proposed to tackle SFDA for image classification, semantic segmentation and object detection. SHOT [36] and SHOT++ [38] froze the classifier of the source model and matched the output target features to source feature distribution with information maximization and pseudo-label strategy. SFDA [41] generated fake samples with a BNS constraint and designed a dual attention distillation mechanism to transfer and retain the contextual information for semantic segmentation. LODS [33] presented a style enhancement method to overlook the target domain style in source-free object detection. However, all of these methods are specifically designed for classification tasks and cannot be directly applied to pixel-level regression tasks, such as SR. To the best of our knowledge, this paper is the first work for Source-Free domain adaptation for image Super-Resolution (SFSR).

3. Methodology

According to the settings commonly used in the UDA task, the source dataset $D_s = \left\{\left(x_s^i, y_s^i\right)\right\}_{i=1}^{n_s}$ with n_s pairs of labeled samples and the target dataset $D_t = \left\{x_t^i\right\}_{i=1}^{n_t}$ with n_t unlabeled samples are given. In our source-free settings, the source dataset D_s is only accessible during pre-training. Our goal is to adapt the pre-trained source model to the target domain without accessing the source data.

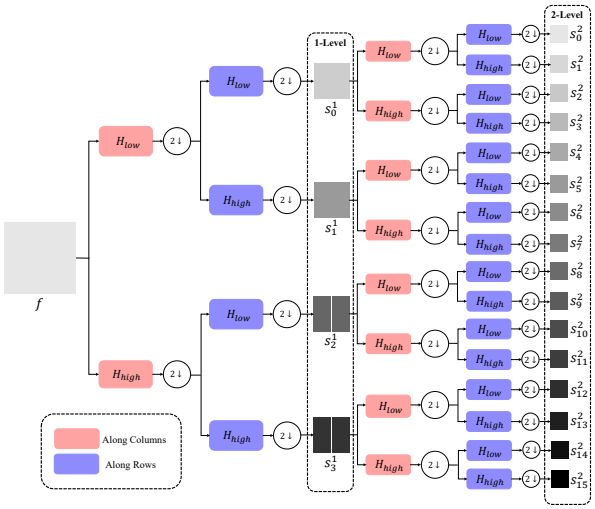


Figure 2. Illustration of 2-level haar wavelet packet transform (WPT). WPT employs low-pass filters H_{low} and high-pass filters H_{high} in a recursive manner to decompose the original features into multiple sub-bands at different frequency resolutions.

3.1. Overview

As shown in Fig. 3, SODA-SR is based on the teacher-student architecture. After pre-training, we can only access the well-trained teacher model f_ξ and unlabeled target data x_t, where ξ denotes the parameters of the teacher model. The student model f_θ has an additional wavelet augmentation transformer (WAT) built upon the teacher model. Let θ_w denote the parameters of WAT and θ_o denote the parameters of other modules in the student model, excluding WAT. θ_o is initialized as the pre-trained teacher model. The proposed WAT is based on wavelet-transform, more specifically wavelet packet transform (WPT). As illustrated in Fig. 2, WPT can decompose the feature map into such sub-bands that have the same spatial size. Given a feature map of pixel-size $H \times W$ with a ℓ-level WPT, we can get 4^ℓ sub-bands of pixel-size $\frac{H}{2^\ell} \times \frac{W}{2^\ell}$. WAT learns low-frequency information of varying levels using multi-level WPT.

SODA-SR consists of two distinct augmentation methods to facilitate teacher-student mutual learning. The first one is to rotate and flip one target LR image to generate seven geometrically augmented images. After feeding eight images into the teacher model to generate the SR images, the resulting SR images will be inverse transformed to their original geometry. The eight outputs will be averaged to produce the refined pseudo-label. The other is to use WAT to learn appropriate augmentation in the latent feature space. During training, there is a 50% probability that the

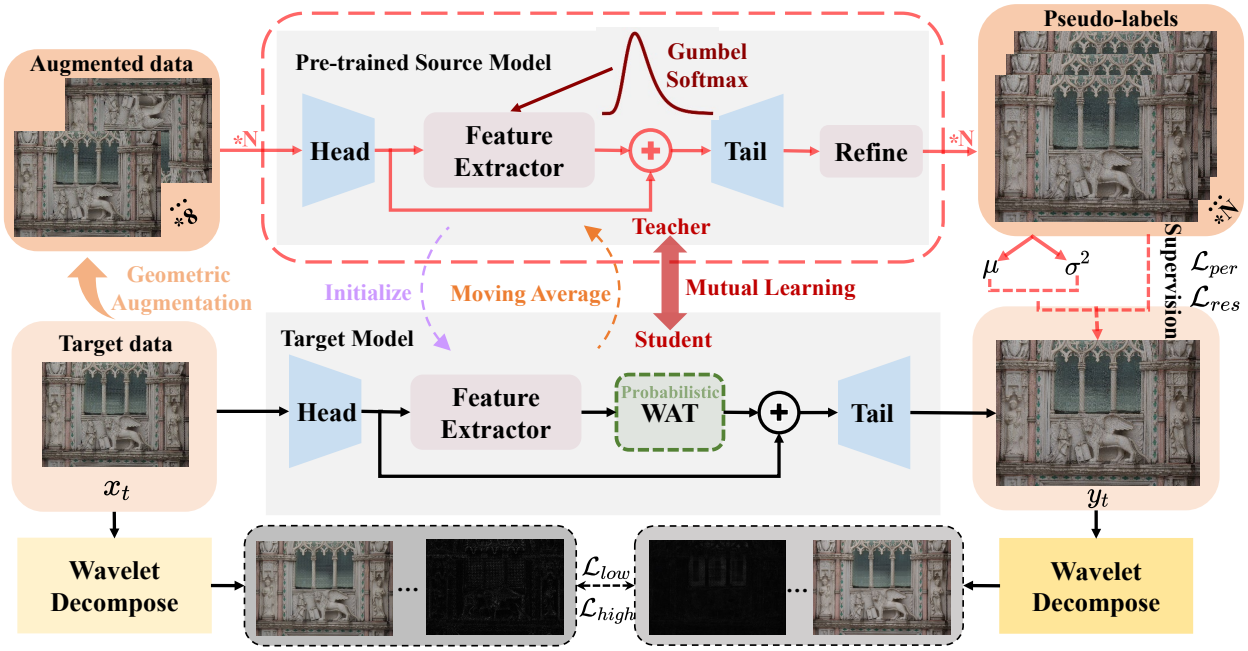


Figure 3. Architecture of the proposed SODA-SR framework. One target LR input image together with its seven geometrically augmented images (*i.e.*, rotate and flip the input) will be fed into the teacher model to generate the refined pseudo-label. The Softmax normalization function in the teacher model will be replaced by Gumbel-Softmax [27]. For one LR input image, the teacher model will run multiple times to generate N pseudo-labels and calculate their mean and variance for uncertainty estimation.

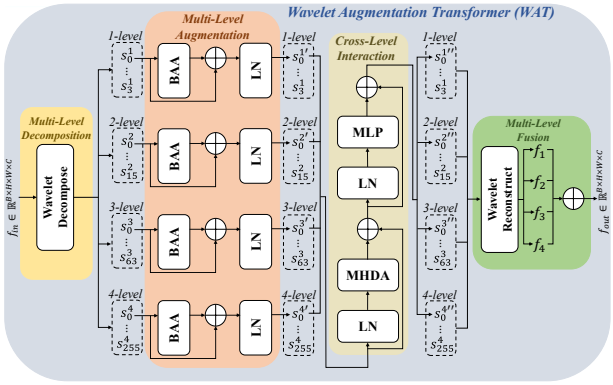


Figure 4. Wavelet Augmentation Transformer (WAT).

feature maps extracted by the feature extractor will be fed into the WAT or not. It's worth noting that WAT will not be utilized during inference, resulting in no additional computational cost.

3.2. Wavelet Augmentation Transformer

As shown in Fig. 4, WAT consists of four key modules, including multi-level decomposition, multi-level augmentation, cross-level interaction, and multi-level fusion. We will proceed to introduce each of these modules as follows.

Multi-Level Decomposition. Given a batch of input image feature $f_{in} \in \mathbb{R}^{B\times H\times W\times C}$, WAT firstly employs a set of ℓ-level WPT to decompose it into multi-level wavelet sub-bands, *i.e.*, $\{s^{(\ell)} \in \mathbb{R}^{B\times m^{(\ell)}\times h^{(\ell)}\times w^{(\ell)}\times C} | \ell \in \mathrm{P}\}$, where $m^{(\ell)} = 4^\ell, h^{(\ell)} = \frac{H}{2^\ell}, w^{(\ell)} = \frac{W}{2^\ell}$ are the number, height, and width of the sub-bands, respectively. $\mathrm{P} = \{1, 2, 3, 4\}$ is a set of wavelet levels. Then we flatten them on the spatial dimension and get $\{s^{(\ell)} \in \mathbb{R}^{B\times m^{(\ell)}\times n^{(\ell)}\times C} | \ell \in \mathrm{P}\}$, where $n^{(\ell)} = h^{(\ell)} \times w^{(\ell)}$. With the wavelet decomposition, the input image feature is transformed into the wavelet space, and the features are independent across different wavelet sub-bands. The image content is embedded in the low-frequency sub-band feature, while the detail and degradation information are embedded in the high-frequency sub-band feature.

Multi-Level Augmentation. After the feature is disentangled in the wavelet feature space, we conduct the feature augmentation across different samples in the input batch. The proposed Batch Augmentation Attention (BAA) performs self-attention in a batch-wise manner. It's worth noting that the self-attention in BAA is conducted across the batch dimension rather than the spatial or channel dimension, i.e., computing cross-covariance across samples to achieve a learnable feature-level augmentation implicitly. To preserve the valuable high-frequency information in multi-level wavelet sub-bands, only the four low-frequency sub-bands, *i.e.*, $\left\{s_0^{(\ell)} | \ell \in \mathrm{P}\right\}$ in Fig. 2, will be fed into BAA simultaneously. Given $s_0^{(\ell)} \in \mathbb{R}^{B\times n^{(\ell)}\times C}$ as the input of the BAA. Firstly, we transpose the first two dimen-

sions of $s_0^{(\ell)}$ so that $s_0^{(\ell)} \in \mathbb{R}^{n^{(\ell)} \times B \times C}$. Then it employs the standard self-attention mechanism for $s_0^{(\ell)}$. We have $Q^{(\ell)}, K^{(\ell)}, V^{(\ell)} \in \mathbb{R}^{n^{(\ell)} \times B \times C}$ and compute the output as

$$\mathrm{SA}(s_0^{(\ell)}) = \mathrm{softmax}(\frac{Q^{(\ell)}(K^{(\ell)})^T}{\sqrt{C}})V^{(\ell)} \in \mathbb{R}^{n^{(\ell)} \times B \times C}, \tag{1}$$

where $(\cdot)^T$ represents the transpose of the second and third dimensions. Finally, we transpose the first two dimensions of the output. Then the final output of BAA has the same shape as the original $s_0^{(\ell)} \in \mathbb{R}^{B \times n^{(\ell)} \times C}$.

BAA is akin to a feature-level Mixup [21, 63], enabling the interaction of information between distinct samples. The proposed multi-level augmentation module can learn low-frequency information of varying levels across diverse samples while also effectively preserving sensitive but valuable high-frequency information in image SR.

Cross-Level Interaction. Inspired from [72] that uses deformable attention to aggregate multi-scale feature maps in object detection, we employ the deformable attention to achieve cross-level information interaction.

As shown in Fig. 4, $\left\{ s_0^{(\ell)'} | \ell \in \mathrm{P} \right\}$ will be fed into the Multi-Head Deformable Attention (MHDA) module to facilitate information exchange across different levels, where $s_0^{(\ell)'} \in \mathbb{R}^{B \times n^{(\ell)} \times C}$. Firstly we concatenate them on the second dimension, denoted as $X' \in \mathbb{R}^{B \times N \times C}$, where $N = \sum_{\ell \in \mathrm{P}} n^{(\ell)}$. Let $x' \in \mathbb{R}^{N \times C}$ denote one sample in X'. For the i^{th} feature $x'_i \in \mathbb{R}^C$ in x', where $i \in \{1, \cdots, N\}$, K features are sampled in each level for each attention head. Let $p_i \in [0, 1]^2$ be the normalized coordinates that represent the spatial position of x'_i in the original feature map. The position of sampling features can be denoted as $p_{hlik} = p_i + \Delta p_{h\ell ik}$, where $\Delta p_{h\ell ik}$ denotes the sampling offset of the k^{th} sampling point in the h^{th} attention head of the ℓ^{th} level. Bilinear interpolation is used to sample the feature and the sampled feature is denoted as $x'_{h\ell ik}$ for simplicity. The output of MHDA can be formulated as

$$x''_i = \sum_{h=1}^{H} W_h^{(1)} [\sum_{\ell \in P} \sum_{k=1}^{K} A_{h\ell ik} W_h^{(2)} x'_{h\ell ik}], \tag{2}$$

where H denotes the number of attention heads, $W_h^{(1)} \in \mathbb{R}^{C \times \frac{C}{H}}$ and $W_h^{(2)} \in \mathbb{R}^{\frac{C}{H} \times C}$ are projection matrices. $A_{h\ell ik}$ is the attention weight, which is obtained by projecting x'_i through a FC layer and normalizing it with softmax.

Through MHDA, the information across different levels can effectively interact. The LayerNorm layer is incorporated prior to both MHDA and MLP, with the addition of a residual connection for each module, which can be formulated as

$$\begin{aligned} X'' &= \mathrm{MHDA}(\mathrm{LN}(X')) + X', \\ X'' &= \mathrm{MLP}(\mathrm{LN}(X'')) + X'', \end{aligned} \tag{3}$$

where $X'' \in \mathbb{R}^{B \times N \times C}$. Then we split it on the second dimension and recover the multi-level feature maps $\left\{ s_0^{(\ell)''} | \ell \in \mathrm{P} \right\}$, whose information has been effectively aggregated via MHDA across different levels.

Multi-Level Fusion. As shown in Fig. 4, we add up the four features after performing wavelet reconstruction on the wavelet sub-bands of different levels respectively. The output feature f_{out} combines information of different levels.

The proposed WAT performs BAA at different levels to mix image features batch-wisely and efficiently fuses these features through deformable attention. WAT can be regarded as a novel form of model noise [58], which stimulates the student model to learn harder from pseudo-labels, thereby acquiring robust features.

3.3. Uncertainty-aware Self-training Mechanism

When the domain gap between the source domain and the target domain is huge, the pseudo-labels generated from the teacher model may be unreliable. To further improve the accuracy of the pseudo-labels, we present an uncertainty-aware self-training mechanism.

Knowledge Transfer. As shown in Fig. 3, the parameters of the teacher model ξ are updated with an exponential moving average (EMA) of the parameters of the student model (excluding WAT) θ_o after each training step:

$$\xi = \eta \cdot \xi + (1 - \eta) \cdot \theta_o, \tag{4}$$

where $\eta \in [0, 1]$ is the decay rate, which is a hyper-parameter to control the update rate of the teacher model. This approach has been proven effective in semi-supervised learning [49] and self-supervised learning [14, 19]. In our SODA-SR, the target domain knowledge learned by the student model can be slowly and progressively transferred to the teacher model via EMA, thereby improving the accuracy of pseudo-labels and promoting mutual learning between the teacher model and the student model.

Pseudo-label Rectification. In order to alleviate the adverse effects of inaccurate pseudo-labels, we incorporate uncertainty estimation into the self-training process to rectify the pseudo-labels.

Specifically, we replace the Softmax normalization function in the teacher model with Gumbel-Softmax [27] to introduce stochasticity in the generation of pseudo-labels. Given 1D vector $v \in \mathbb{R}^n$, the output of Gumbel-Softmax is formulated as

$$v_i = \frac{\exp((log(v_i) + g_i)/\tau)}{\sum_{j=1}^{n} \exp((log(v_j) + g_j)/\tau)}, \tag{5}$$

where $g_1, \cdots, g_n$ are sampled from $\mathrm{Gumbel}(0, 1)$ distribution and τ is a temperature parameter. $g_1, \cdots, g_n$ introduce stochasticity to the teacher model, enabling it to produce diverse SR results. For one target LR input, we run the

teacher model multiple times to generate N pseudo-labels $y_p^1, \cdots, y_p^N$. Then we compute the mean and variance as the uncertainty estimation, which is formulated as

$$y_{mean} = \frac{1}{N}\sum_{n=1}^{N} y_p^n,\ \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(y_p^n - y_{mean})^2. \quad (6)$$

Compared with existing methods, the proposed simple yet effective uncertainty estimation approach does not require additional components, *e.g.*, Batch Normalization [50] or Dropout [13], which may affect the SR results. Then we compute the pixel-level confidence map cof as following

$$cof = \beta - \text{Sigmoid}(\frac{\sigma^2}{\alpha}), \quad (7)$$

where α and β are hyper-parameters that adjust the value range of cof and they are set empirically to 0.0004 and 1.5, respectively. The confidence map reflects the magnitude of pixel-wise uncertainty. During training, we calculate the pixel-wise weighted L1 loss between the output of the student model f_θ and the averaged pseudo-labels y_{mean} using the confidence map:

$$\mathcal{L}_{rec} = \|cof \odot f_\theta(x_t) - cof \odot y_{mean}\|_1. \quad (8)$$

The inaccurate pseudo-labels will be rectified by the confidence map. In addition to the L1 loss, we also utilize VGG-19 [45] to calculate the perceptual loss $\mathcal{L}_{per}$.

Regularization Losses. Furthermore, to prevent the student model from overfitting pseudo-labels, two regularization losses are proposed to constrain the frequency information between LR and SR images. As shown in Fig. 3, we conduct wavelet decomposition on LR and SR images at different levels, ensuring that the resulting wavelet subbands have the same resolution. We impose L1 loss in the low-frequency space while adversarial loss in the high-frequency space. The L1 loss is defined as

$$\mathcal{L}_{low} = \left\| \text{wavelet}_{\mathcal{L}}^{(l1)}(x_t) - \text{wavelet}_{\mathcal{L}}^{(l2)}(f_\theta(x_t)) \right\|_1, \quad (9)$$

where $\text{wavelet}_{\mathcal{L}}^{(l*)}(\cdot)$ represents the low-frequency subband of $l*$-level wavelet decomposition. The adversarial loss for the generator (*i.e.*, the student model f_θ) and the discriminator $\mathcal{D}$ is respectively defined as

$$\mathcal{L}_{high}^G = -\mathbb{E}_{x_t}[log(\mathcal{D}(\text{wavelet}_{\mathcal{H}}^{(l2)}(f_\theta(x_t))))], \quad (10)$$

$$\begin{aligned}\mathcal{L}_{high}^D = &-\mathbb{E}_{x_t}[log(\mathcal{D}(\text{wavelet}_{\mathcal{H}}^{(l1)}(x_t)))] \\ &-\mathbb{E}_{x_t}[log(1-\mathcal{D}(\text{wavelet}_{\mathcal{H}}^{(l2)}(f_\theta(x_t))))],\end{aligned} \quad (11)$$

where $\text{wavelet}_{\mathcal{H}}^{(l*)}(\cdot)$ represents the high-frequency subband of $l*$-level wavelet decomposition.

Full objective function. The full objective function for the student model f_θ is defined as

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_1\mathcal{L}_{per} + \lambda_2\mathcal{L}_{low} + \lambda_3\mathcal{L}_{high}^G, \quad (12)$$

where λ_1, λ_2 and λ_3 are loss weights to balance each item.

4. Experiments

4.1. Experimental Setup

Datasets and Metrics. We evaluate our method on the DRealSR [56] dataset. DRealSR is a large-scale real-world SR benchmark, which is collected by five DSLR cameras (*i.e.*, Panasonic, Sony, Olympus, Nikon, and Canon) in real-world scenarios. Following DADA [60], we choose images captured from three cameras (Panasonic, Sony and Olympus) for our experiments, which contain 197, 145 and 190 image pairs for training; 20, 17 and 19 image pairs for testing, respectively. The image SR performance is evaluated by calculating PSNR, SSIM, and LPIPS [65]. PSNR and SSIM are computed on the Y channel (transformed YCbCr space) and RGB space, respectively.

Implementation Details. Following DADA [60], we adopt CDC [56] as the backbone network to achieve a fair comparison. Following DADA [60], the training patch size is set to 48 $\times$ 48. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a fixed learning rate of 2×10^{-6}. The decay rate η in Eq. (4) is set to 0.999 to make the training process more stable. The temperature parameter τ in Eq. (5) is set to 0.1 to mitigate the impact on the performance of the teacher model while introducing stochasticity. The number of generated pseudo-labels N is set to 5. The levels of wavelet decomposition l_1 and l_2 in Eq. (9) are set to 1 and 3, respectively. For hyper-parameters in Eq. (12), the loss weights $\lambda_1, \lambda_2, \lambda_3$ are set to 0.01, 0.1, and 0.005, respectively. The proposed framework will converge after about 4000 iterations with a batch-size of 32.

4.2. Comparison with state-of-the-art methods

Due to the absence of the source-free method for real-world image SR, we compare our SODA-SR with the state-of-the-art UDA methods and self-supervised methods for real-world image SR. The competing UDA methods include CinCGAN [61], DASR [57], DRN-Adapt [16] and DADA [60] while self-supervised methods include ZSSR [2] and MZSR [46]. Our experiments include real→real adaptation and synthetic→real adaptation. In the real→real adaptation task, our source model is trained on the real-world image pairs from DRealSR. Alternatively, in the synthetic→real adaptation task, our source model is trained on synthetic image pairs (*i.e.*, LR images are bicubically downsampled from the real-world HR images). Our experiments are for scaling factor $\times4$.

Table 1 shows the quantitative results in six camera→camera adaptation tasks. "Source only" represents the model trained on the source data without domain adaptation. "Target only" represents the model trained on the labeled target data. In six real→real tasks, our method achieves the best performance on PSNR and SSIM, and the second best performance on LPIPS.

Method	SF	Panasonic → Sony			Sony → Panasonic			Olympus → Panasonic		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Real→Real										
Target Only	✗	32.71	0.855	0.296	32.33	0.845	0.318	32.33	0.845	0.318
Source Only	✗	31.32	0.841	0.314	30.72	0.820	0.372	30.49	0.820	0.363
CinCGAN [61]	✗	27.76	0.821	0.391	28.33	0.792	0.410	29.37	0.799	0.381
DASR [57]	✗	30.08	0.777	0.269	30.45	0.772	0.316	30.06	0.785	0.272
DRN-Adapt [16]	✗	31.85	0.845	0.321	30.96	0.821	0.380	30.80	0.822	0.356
DADA [60]	✗	32.13	0.849	0.327	31.25	0.825	0.363	31.27	0.824	0.348
SODA-SR (Ours)	✓	32.24	0.851	0.312	31.40	0.833	0.345	31.41	0.832	0.344
Synthetic → Real										
Source Only	✗	31.44	0.828	0.373	30.45	0.808	0.433	30.44	0.806	0.434
CinCGAN [61]	✗	27.59	0.788	0.405	27.19	0.743	0.414	28.38	0.739	0.422
DASR [57]	✗	29.95	0.764	0.298	29.79	0.749	0.339	30.02	0.777	0.293
DRN-Adapt [16]	✗	31.42	0.829	0.359	30.47	0.808	0.429	30.45	0.808	0.433
DADA [60]	✗	31.50	0.830	0.369	30.72	0.809	0.376	30.74	0.808	0.362
SODA-SR (Ours)	✓	31.61	0.831	0.354	30.80	0.810	0.372	30.82	0.809	0.361
Method	**SF**	**Panasonic → Olympus**			**Sony → Olympus**			**Olympus → Sony**		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Real→Real										
Target Only	✗	31.67	0.834	0.375	31.67	0.834	0.375	32.71	0.855	0.296
Source Only	✗	30.23	0.812	0.438	30.48	0.810	0.449	30.45	0.812	0.323
CinCGAN [61]	✗	28.85	0.791	0.461	30.17	0.814	0.443	30.05	0.823	0.365
DASR [57]	✗	29.32	0.768	0.306	29.86	0.762	0.372	30.29	0.787	0.270
DRN-Adapt [16]	✗	30.73	0.816	0.431	30.66	0.810	0.459	31.47	0.833	0.312
DADA [60]	✗	31.08	0.820	0.433	31.08	0.817	0.438	32.05	0.843	0.343
SODA-SR (Ours)	✓	31.16	0.828	0.386	31.22	0.828	0.403	32.13	0.850	0.311
Synthetic → Real										
Source Only	✗	30.10	0.798	0.480	30.09	0.798	0.473	31.43	0.828	0.371
CinCGAN [61]	✗	28.43	0.766	0.407	29.34	0.767	0.451	29.50	0.792	0.392
DASR [57]	✗	28.30	0.752	0.375	29.51	0.755	0.402	29.40	0.737	0.327
DRN-Adapt [16]	✗	30.11	0.799	0.475	30.11	0.799	0.473	31.45	0.829	0.362
DADA [60]	✗	30.40	0.800	0.403	30.62	0.803	0.411	31.52	0.829	0.355
SODA-SR(Ours)	✓	30.42	0.801	0.400	30.63	0.804	0.408	31.54	0.830	0.351

Table 1. Quantitative comparison with state-of-the-art UDA methods for × 4 SR. "**SF**" represents whether the method is under source-free setting. Except *Target Only*, the best and second best performance are in red and blue colors, respectively.

Panasonic Testset				**Sony Testset**				**Olympus Testset**			
Method	PSNR↑	SSIM↑	LPIPS↓	Method	PSNR↑	SSIM↑	LPIPS↓	Method	PSNR↑	SSIM↑	LPIPS↓
MZSR [46]	28.73	0.785	0.398	MZSR [46]	29.00	0.796	0.366	MZSR [46]	28.54	0.777	0.443
ZSSR [2]	30.42	0.805	0.417	ZSSR [2]	31.34	0.823	0.344	ZSSR [2]	30.06	0.794	0.455
Ours (S→P)	31.40	0.833	0.345	**Ours (O →S)**	32.13	0.850	0.311	**Ours (P→O)**	31.16	0.828	0.386
Ours (O→P)	31.41	0.832	0.344	**Ours (P→S)**	32.24	0.851	0.312	**Ours (S→O)**	31.22	0.828	0.403

Table 2. Quantitative comparison with typical self-supervised SR methods for ×4 SR. "P", "S" and "O" represent Panasonic, Sony and Olympus, respectively. For the test set of one camera, our method has two adaptation settings.

Although DASR achieves the best LPIPS, its PSNR and SSIM are inferior to ours, and it often produces SR images containing artifacts and noises. In six synthetic→real tasks, our method also performs better than other methods. Table 2 shows the quantitative results compared with self-supervised SR methods on three test sets. ZSSR and MZSR require only a single LR image to train a SR network specifically tailored to that LR image. Since our method preserves the domain-invariant knowledge in the pre-trained source model and reduces the cross-domain discrepancy by model adaptation, our method performs much better than the self-supervised SR methods.

Fig. 5 shows that our method can not only reason the correct structure of the buildings but also generate clear details, while other methods may generate deformed structure and blurry results. The results of our method are closest to those trained with target labels. To validate the effectiveness of the proposed method on other backbone network, we take experiments on using SwinIR [37] as the backbone. The results are reported in the Appendix for page limit.

4.3. Ablation Study

We conduct an ablation study to evaluate the respective roles of each part in our method. As shown in Table 3,

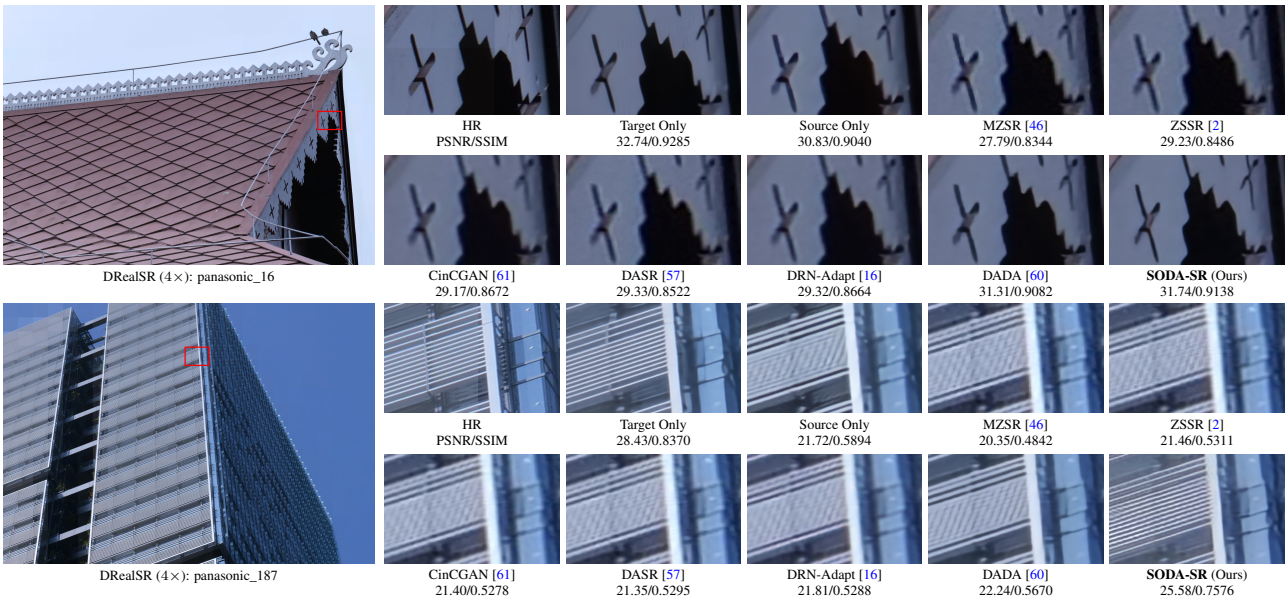


Figure 5. Visual comparison for × 4 SR on DRealSR dataset (Sony→Panasonic). Best viewed with zoom in.

Metrics	Source Only	w/o WAT	w/o EMA	w/o Reg. Loss	w/o UE	Ours
PSNR↑	30.49	30.64	31.14	31.31	31.33	**31.41**
SSIM↑	0.820	0.822	0.826	0.829	0.830	**0.832**
LPIPS↓	0.363	0.365	0.354	0.356	0.346	**0.344**

Table 3. Ablation results on DRealSR (Olympus→Panasonic). We evaluate the effectiveness of WAT, EMA strategy, regularization losses, and uncertainty estimation (UE).

Figure 6. Visual illustration of the uncertainty estimation.

the performance drops a lot when separately removing WAT and EMA. This demonstrates the significant role of WAT and EMA in facilitating teacher-student learning. The results also demonstrate that the regularization loss and uncertainty estimation (UE) can improve the quality of SR images. As shown in Fig. 6, pixels with higher error in the pseudo-label will be assigned lower confidence, which also proves the effectiveness of the proposed UE. Fig. 7 also demonstrates that WAT can empower the model to utilize a broader range of pixels, leading to enhanced SR results. A detailed analysis of these components and the loss function in Eq. (12) are provided in the Appendix.

5. Conclusion

In this paper, we propose a novel source-free domain adaptation framework named SODA-SR, which attempts to adapt a source-trained model to the target domain without accessing source data. By using our proposed wavelet augmentation transformer (WAT), the student model is capable of learning low-frequency information of varying levels across diverse samples, which is aggregated efficiently via deformable attention. Besides, an uncertainty-aware self-training mechanism is proposed to facilitate knowledge transfer and rectify pseudo-labels. Several regularization losses are proposed to avoid overfitting pseudo-labels. Extensive experiments under real→real and synthetic→real adaptation settings on DRealSR demonstrate the effectiveness of our method.

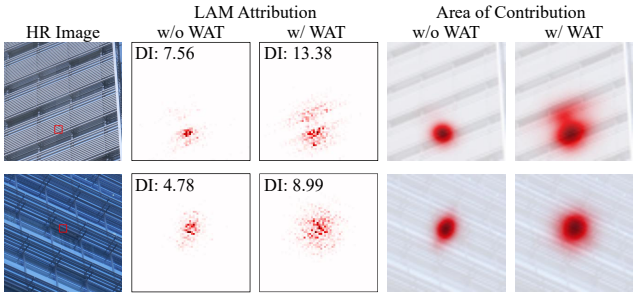


Figure 7. LAM [15] illustration on challenging cases. LAM attribution shows how important each pixel in the LR image is for reconstructing the marked patch. Diffusion Index (DI) reflects how many pixels are used, with a higher DI meaning more pixels are involved. The results indicate WAT enables the SR model to utilize a wider range of pixels for reconstruction.

Acknowledgements: This research is partially funded by Youth Innovation Promotion Association CAS (Grant No. 2022132), Beijing Nova Program (20230484276), National Natural Science Foundation of China (Grant No. U21B2045, U20A20223) and OPPO Research fund.

References

[1] Yuang Ai, Huaibo Huang, Xiaoqiang Zhou, Jiexiang Wang, and Ran He. Multimodal prompt perceiver: Empower adaptiveness, generalizability and fidelity for all-in-one image restoration. *arXiv preprint arXiv:2312.02918*, 2023. 3

[2] Michal Irani Assaf Shocher, Nadav Cohen. "zero-shot" super-resolution using deep internal learning. In *CVPR*, pages 3118–3126, 2018. 1, 6, 7, 8

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, pages 5049–5059, 2019. 2

[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 *ICCV*, pages 3086–3095, 2019. 2, 3

[5] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *CVPR*, pages 1652–1660, 2019. 3

[6] 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 *CVPR*, pages 12299–12310, 2021. 3

[7] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, pages 22367–22377, 2023. 3

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 38(2):295–307, 2015. 1, 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 3

[10] Boyan Duan, Chaoyou Fu, Yi Li, Xingguang Song, and Ran He. Cross-spectral face hallucination via disentangling independent factors. In *CVPR*, pages 7930–7938, 2020. 3

[11] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*, pages 9935–9946, 2023. 3

[12] Francois Fleuret et al. Uncertainty reduction for model adaptation in semantic segmentation. In *CVPR*, pages 9613–9623, 2021. 2

[13] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016. 6

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, pages 21271–21284, 2020. 5

[15] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *CVPR*, pages 9199–9208, 2021. 8

[16] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. Closed-loop matters: Dual regression networks for single image super-resolution. In *CVPR*, pages 5407–5416, 2020. 1, 3, 6, 7, 8

[17] Viet Khanh Ha, Jin-Chang Ren, Xin-Ying Xu, Sophia Zhao, Gang Xie, Valentin Masero, and Amir Hussain. Deep learning based single image super-resolution: A survey. *Machine Intelligence Research*, 16(4):413–426, 2019. 3

[18] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, pages 1664–1673, 2018. 3

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 5

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 3

[21] Zhi Hou, Baosheng Yu, and Dacheng Tao. Batchformer: Learning to explore sample relationships for robust representation learning. In *CVPR*, pages 7256–7266, 2022. 5

[22] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, pages 646–661, 2016. 2

[23] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *ICCV*, pages 1689–1697, 2017. 3

[24] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet domain generative adversarial network for multi-scale face hallucination. *IJCV*, 127(6):763–784, 2019. 3

[25] Huaibo Huang, Aijing Yu, and Ran He. Memory oriented transfer learning for semi-supervised image deraining. In *CVPR*, pages 7732–7741, 2021. 1

[26] Huaibo Huang, Mandi Luo, and Ran He. Memory uncertainty learning for real-world single image deraining. *TPAMI*, 45(3):3446–3460, 2022. 1

[27] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017. 4, 5

[28] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, pages 23593–23606, 2022. 3

[29] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016. 3

[30] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *TPAMI*, 41(11):2599–2613, 2018.

[31] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017. 1, 3

[32] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *CVPR*, pages 9641–9650, 2020. 2

[33] Shuaifeng Li, Mao Ye, Xiatian Zhu, Lihua Zhou, and Lin Xiong. Source-free object detection by learning to overlook domain style. In *CVPR*, pages 8014–8023, 2022. 2, 3

[34] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. In *AAAI*, pages 8474–8481, 2021. 2

[35] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *CVPR*, pages 18278–18289, 2023. 3

[36] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, pages 6028–6039, 2020. 2, 3

[37] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, pages 1833–1844, 2021. 1, 3, 7

[38] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *TPAMI*, 44 (11):8602–8617, 2022. 2, 3

[39] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, pages 136–144, 2017. 3

[40] Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *CVPR*, pages 7640–7650, 2023. 2

[41] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *CVPR*, pages 1215–1224, 2021. 2, 3

[42] Shao-Yuan Lo, Poojan Oza, Sumanth Chennupati, Alejandro Galindo, and Vishal M. Patel. Spatio-temporal pixel-level contrastive learning-based source-free domain adaptation for video semantic segmentation. In *CVPR*, pages 10534–10543, 2023. 2

[43] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*, pages 1–10, 2022. 3

[44] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *TPAMI*, 45(4):4713–4726, 2022. 3

[45] K. Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[46] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *CVPR*, pages 3516–3525, 2020. 6, 7, 8

[47] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, pages 596–608, 2020. 2

[48] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15 (1):1929–1958, 2014. 2

[49] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pages 1195–1204, 2017. 2, 5

[50] Mattias Teye, Hossein Azizpour, and Kevin Smith. Bayesian uncertainty estimation for batch normalized deep networks. In *ICML*, pages 4907–4916, 2018. 6

[51] Vibashan VS, Poojan Oza, and Vishal M. Patel. Instance relation graph guided source-free domain adaptive object detection. In *CVPR*, pages 3520–3530, 2023. 2

[52] Wei Wang, Haochen Zhang, Zehuan Yuan, and Changhu Wang. Unsupervised real-world super-resolution: A domain adaptation perspective. In *ICCV*, pages 4318–4327, 2021. 1

[53] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. Esrgan: Enhanced super-resolution generative adversarial networks. *arXiv preprint arXiv:1809.00219*, 2018. 3

[54] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *ICLR*, 2023. 3

[55] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17683–17693, 2022. 3

[56] 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 *ECCV*, pages 101–117, 2020. 1, 2, 3, 6

[57] Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, and Hengjie Song. Unsupervised real-world image super resolution via domain-distance aware training. In *CVPR*, pages 13385–13394, 2021. 1, 3, 6, 7, 8

[58] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020. 5

[59] Ling-Yi Xu and Zoran Gajic. Improved network for face recognition based on feature super resolution method. *Machine Intelligence Research*, 18(6):915–925, 2021. 3

[60] Xiaoqian Xu, Pengxu Wei, Weikai Chen, Yang Liu, Mingzhi Mao, Liang Lin, and Guanbin Li. Dual adversarial adaptation for cross-device real-world image super-resolution. In *CVPR*, pages 5667–5676, 2022. 1, 2, 3, 6, 7, 8

[61] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *CVPRW*, pages 701–710, 2018. 1, 3, 6, 7, 8

[62] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022. 3

[63] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 5

[64] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. In *ICLR*, 2023. 3

[65] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6

[66] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *CVPR*, pages 3762–3770, 2019. 3

[67] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018. 3

[68] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, pages 2472–2481, 2018.

[69] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *TPAMI*, 43(7):2480–2495, 2021. 3

[70] Xiaoqiang Zhou, Huaibo Huang, Ran He, Zilei Wang, Jie Hu, and Tieniu Tan. Msra-sr: Image super-resolution transformer with multi-scale shared representation acquisition. In *ICCV*, pages 12665–12676, 2023. 3

[71] Xiaoqiang Zhou, Huaibo Huang, Zilei Wang, and Ran He. Ristra: Recursive image super-resolution transformer with relativistic assessment. *TMM*, pages 1–12, 2024. 3

[72] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 2, 5