

Learning Mutual Modulation for Self-Supervised Cross-Modal Super-Resolution

Xiaoyu Dong^{1,2}, Naoto Yokoya^{1,2} (✉), Longguang Wang³, and Tatsumi Uezato⁴

¹ The University of Tokyo, Tokyo, Japan

² RIKEN AIP, Tokyo, Japan

³ National University of Defense Technology, Changsha, China

⁴ Hitachi, Ltd, Tokyo, Japan

dong@ms.k.u-tokyo.ac.jp, yokoya@k.u-tokyo.ac.jp

<https://github.com/palmdong/MMSR>

Abstract. Self-supervised cross-modal super-resolution (SR) can overcome the difficulty of acquiring paired training data, but is challenging because only low-resolution (LR) source and high-resolution (HR) guide images from different modalities are available. Existing methods utilize pseudo or weak supervision in LR space and thus deliver results that are blurry or not faithful to the source modality. To address this issue, we present a mutual modulation SR (MMSR) model, which tackles the task by a mutual modulation strategy, including a source-to-guide modulation and a guide-to-source modulation. In these modulations, we develop cross-domain adaptive filters to fully exploit cross-modal spatial dependency and help induce the source to emulate the resolution of the guide and induce the guide to mimic the modality characteristics of the source. Moreover, we adopt a cycle consistency constraint to train MMSR in a fully self-supervised manner. Experiments on various tasks demonstrate the state-of-the-art performance of our MMSR.

Keywords: Mutual Modulation, Self-Supervised Super-Resolution, Cross-Modal, Multi-Modal, Remote Sensing

1 Introduction

Multi-modal data, e.g., visible RGB, depth, and thermal, can reflect diverse physical properties of scenes and objects and are widely applied in practice [1,50,62,3]. While high-resolution (HR) visible data is easy to acquire, non-visible modalities are usually low-resolution (LR) due to sensor limitations [11,2]. This hinders practical applications and introduces the necessity of cross-modal super-resolution (SR).

Cross-modal SR aims at increasing the resolution of an LR modality (source) by using as guidance the structural cues from an HR modality (guide). This is difficult due to the spatial discrepancy between different modalities [11,8,56].

In recent years, deep CNNs have been widely studied to explore source-guide spatial dependency for cross-modal SR, and numerous networks have been

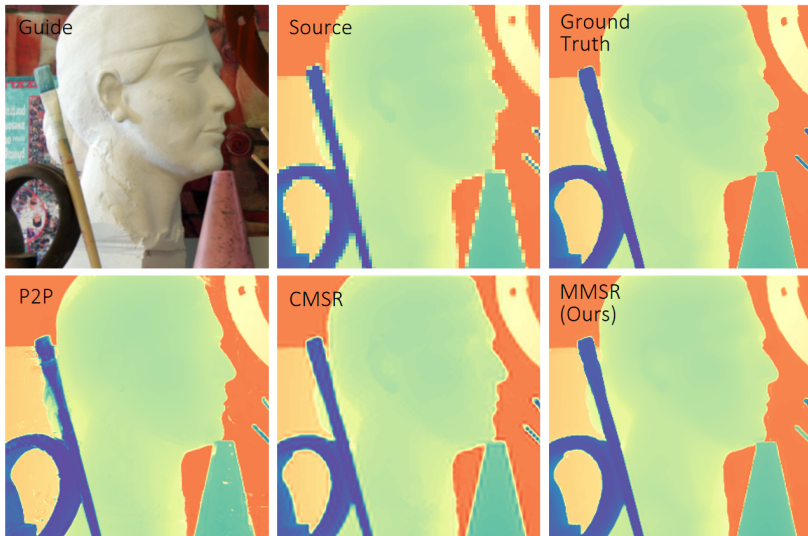


Fig. 1. $\times 4$ depth SR results from CMSR [47], P2P [36], and our MMSR. Our MMSR achieves results that are HR and faithful to the source modality

developed [17,2,28,11,8,19,51]. However, these methods rely on HR source images (i.e., ground truth) for supervised learning and suffer limited generalization performance, since large-scale training data with paired ground truth is hard to collect [36,16,47].

To address this issue, several efforts [47,36] have been made to learn cross-modal SR in a self-supervised¹ manner. Such methods do not require external training data and perform online learning on each combination of LR source and HR guide, thus providing a strong generalization capability. However, they face two technical challenges: **First**, training SR models using two images that have modality difference and cannot form a direct LR-to-HR mapping. **Second**, achieving high spatial resolution as well as faithful image modality without accessing the supervision from HR source. To find a solution, CMSR [47] further downsampled the LR source to generate pseudo paired data in LR space. P2P [36] formulated the task as a modality transformation problem of the HR guide, and employed the LR source as weak supervision. While succeeding in training the models, these methods cannot overcome the second challenge and deliver results that are blurry or not faithful to the source modality (Fig. 1). Overall, robust self-supervised cross-modal SR remains an open problem.

In this paper, we tackle self-supervised cross-modal SR by modulating the source and the guide with a cycle consistency constraint (Fig. 2). Specifically, we introduce a mutual modulation strategy, which includes a source-to-guide mod-

¹ In this paper, self-supervised learning [13] refers to learning from data without paired ground truth in source modality, i.e., only an LR source and an HR guide.

ulation (Fig. 3) to induce the source to emulate the resolution of the guide, and a guide-to-source modulation (Fig. 4) to bring the guide closer to the source with respect to the characteristics of the imaging modality. During the modulations, we develop cross-domain adaptive filters to fully exploit the spatial dependency between the source and the guide and drive our mutual modulation. Moreover, we adopt a cycle consistency loss between the downsampled SR result and the original LR source input to train our mutual modulation SR (MMSR) model in a fully self-supervised manner. It is demonstrated that our MMSR achieves state-of-the-art performance and produces results with both fine spatial detail and faithful image modality (Fig. 1).

Contributions: **(1)** We address an open problem in cross-modal SR, and develop a robust self-supervised MMSR model. **(2)** We propose a mutual modulation strategy, and show correlation-based filtering provides an effective inductive bias for deep self-supervised cross-modal SR models. **(3)** We validate our MMSR on depth, digital elevation model (DEM), and thermal SR, which involve benchmark data, noisy data, and real-world remote sensing data, demonstrating its robustness, generalizability, and applicability. **(4)** We compare our MMSR with state-of-the-art supervised and self-supervised methods, comprehensively demonstrating both its quantitative and qualitative superiority.

2 Related Work

In this section, we first review several mainstream works in cross-modal SR. Then we discuss techniques that are related to our work, including modulation networks, image filtering, and cycle-consistent learning.

2.1 Cross-Modal SR

Cross-modal SR has evolved from filtering-based [66,15,33,34], optimization-based [7,41,10], and dictionary-based methods [26,23] to learning-based methods [2,36,11] over the past decades. We focus on learning-based methods and review several supervised and self-supervised methods.

Supervised methods, as in other SR tasks [57,63,67,69,60,55], have made great progress. Early pioneers [17,27,28,2] have cast the task in a learning-based manner. Recent works [12,11,40,8,19,54] have studied the spatial dependency between the source and guide images. Representative work includes the weighted analysis sparse representation model [12,11] and the spatially variant linear representation model [40,8]. While obtaining promising performance, these methods suffer limited generalization performance in real-world scenes since large-scale paired training data is hard to acquire [36,16,47].

To address this issue, self-supervised methods without external training have been studied [47,36]. Such methods perform online learning on each combination of LR source and HR guide, and so can be adapted to any given scenario. Existing methods conduct the task by forming pseudo supervision in LR space [47]

or interpret the task as cross-modal transformation in a weakly supervised manner [36]. While successfully training the models, their delivered results, caused by the non-ideal supervisions, are blurry or not faithful to the source modality.

2.2 Modulation Networks

Modulation networks are emerging in different research fields [65,32,25]. In image restoration, researchers have developed modulation networks to control restoration capability and flexibility [14,61,59]. Wang et al. [58] presents a degradation-aware modulation block to handle different degradations in blind SR. Xu et al. [64] designs a temporal modulation network to achieve arbitrary frame interpolation in space-time video SR. In speech separation, Lee et al. [24] introduces a cross-modal affinity transformation to overcome the frame discontinuity between audio and visual modalities. In image retrieval, Lee et al. [25] introduces content-style modulation to handle the task by exploiting text feedback.

We propose a mutual modulation strategy to tackle self-supervised cross-modal SR. Our strategy enables our model to achieve results with both high spatial resolution and faithful image modality, and outperform even state-of-the-art supervised methods.

2.3 Image Filtering

Many vision applications involve image filtering to suppress and/or extract content of interests in images [15]. Simple linear filters have been extensively used in image tasks such as sharpening, stitching [42], and matting [52]. In the area of image restoration, local [43] and non-local [5,6] filtering have been studied. Liu et al. [31] first incorporated non-local operations in a neural network for denoising and SR. Later researchers used non-local layers to exploit the self-similarity prior for restoration quality improvement [70,39,38,49].

Differently, we aim at handling multi-modal images that have local spatial dependency [8,40], discrepancy, and resolution gap. Therefore, we learn filters confined to pixel neighborhoods across features from the source and guide modality domains to exploit the local spatial dependency of different modalities and drive our mutual modulation. Experiments in Section 4.3 show that our filters can eliminate the spatial discrepancy and resolution gap of multi-modal images, providing an effective inductive bias for cross-modal SR models.

2.4 Cycle-Consistent Learning

Given a data pair A and B , cycle-consistent learning aims to train deep models by establishing a closed cycle with a forward A -to- B mapping and a backward B -to- A mapping. This idea has been investigated in vision tasks such as visual tracking [18,53], dense semantic alignment [71,72], and image translation [73,68]. In image restoration, researchers imposed cycle consistency constraint to image dehazing [48] and unpaired SR [37].

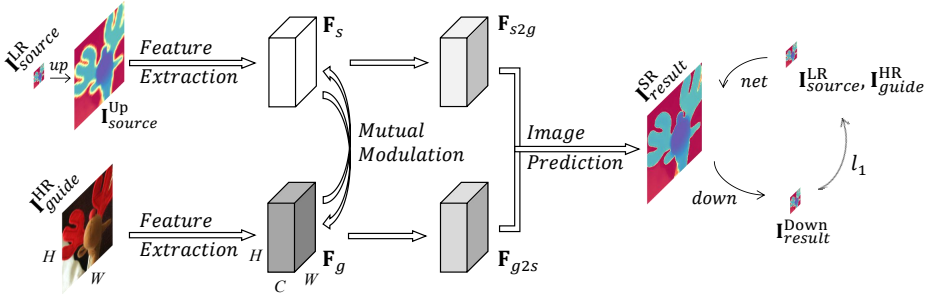


Fig. 2. An illustration of our MMSR model. During mutual modulation, the source is induced to emulate the resolution of the guide, while the guide is induced to mimic the characteristics of the source. A cycle consistency constraint is adopted to conduct training in a fully self-supervised manner. 'up', 'net', 'down', and ' l_1 ' represent upsampling, our network, downsampling, and loss term, respectively

We introduce cycle-consistent learning to cross-modal SR, and adopt a cycle consistency loss to encourage the downsampled SR source and the original LR source input to be consistent with each other. This allows our model to be trained in a fully self-supervised manner.

3 Method

As illustrated in Fig. 2, our MMSR starts from an LR source \mathbf{I}_{source}^{LR} and the HR guide \mathbf{I}_{guide}^{HR} . It then modulates the source feature \mathbf{F}_s extracted from the bilinearly upsampled source \mathbf{I}_{source}^{Up} , which still presents low resolution and lacks of spatial detail, and the guide feature \mathbf{F}_g extracted from \mathbf{I}_{guide}^{HR} , which contains HR structural cues important to the source and also discrepancy patterns. Finally, it predicts the SR source \mathbf{I}_{result}^{SR} from the fusion of the modulated \mathbf{F}_{s2g} and \mathbf{F}_{g2s} , and constrains itself by casting \mathbf{I}_{result}^{SR} back to \mathbf{I}_{source}^{LR} .

3.1 Mutual Modulation

In our mutual modulation, \mathbf{F}_s and \mathbf{F}_g are optimized by taking each other as reference. Cross-domain adaptive filters are developed as basic operators to drive the modulation.

In the **Source-to-Guide Modulation** (Fig. 3), \mathbf{F}_s is modulated to emulate the high resolution of \mathbf{F}_g . Specifically, to each pixel in \mathbf{F}_s (denoted as $\mathbf{s}_{(i,j)}$), we learn a filter $f_{(i,j)}^{s2g}(\cdot)$ confined to its neighbor pixels in an $n \times n$ neighborhood (denoted as $\mathbf{N}_{\mathbf{s}_{(i,j)}}$) and target its counterpart pixel in \mathbf{F}_g (denoted as $\mathbf{g}_{(i,j)}$)².

² $\mathbf{s}_{(i,j)}$ or $\mathbf{g}_{(i,j)}$ denotes the pixel at the i -th row and j -th column in \mathbf{F}_s or \mathbf{F}_g and is a vector of size $C \times 1$. $\mathbf{N}_{\mathbf{s}_{(i,j)}}$ contains $n \times n$ pixels and is a tensor of shape $C \times n \times n$.

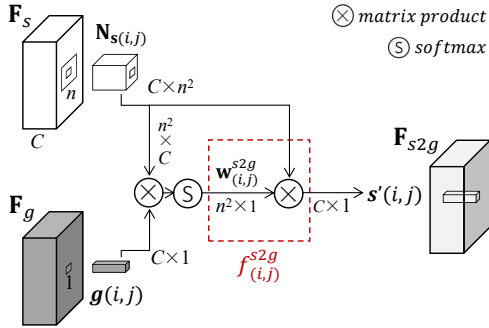


Fig. 3. An illustration of the source-to-guide modulation. Filters update the pixels in \mathbf{F}_s by targeting the counterparts in \mathbf{F}_g to induce \mathbf{F}_s to become HR

The filter weight is calculated as³:

$$\mathbf{w}_{(i,j)}^{s2g} = \text{softmax}\left(\left(\mathbf{N}_{s(i,j)}\right)^T \mathbf{g}_{(i,j)}\right), \quad (1)$$

and evaluates the correlation value between $\mathbf{g}_{(i,j)}$ and each pixel in $\mathbf{N}_{s(i,j)}$ [31,4]. Thus our filters allow fully exploitation of the local dependency between the source and guide modalities. In Section 4.3, we experimentally show that such adaptive filters with learning cross-domain correlations can deliver product that is spatially approaching a given target from a different domain, and are effective modulators to \mathbf{F}_s and \mathbf{F}_g in a case in which neither an HR source feature nor a guide feature without spatial discrepancy is available. Here, a filtering operation is expressed as⁴:

$$\mathbf{s}'_{(i,j)} = f_{(i,j)}^{s2g}(\mathbf{N}_{s(i,j)}) = \mathbf{N}_{s(i,j)} \mathbf{w}_{(i,j)}^{s2g}, \quad (2)$$

where the resulting $\mathbf{s}'_{(i,j)}$ is the update of $\mathbf{s}_{(i,j)}$ and is induced to spatially emulate $\mathbf{g}_{(i,j)}$, which is in HR domain. The whole source-to-guide modulation is conducted by updating all the pixels in \mathbf{F}_s by targeting the counterpart guide pixels, resulting in \mathbf{F}_{s2g} , which inherits the HR property of the guide.

The **Guide-to-Source Modulation** (Fig. 4) suppresses the discrepancy in \mathbf{F}_g to make its characteristics more like those of \mathbf{F}_s . To guide pixel $\mathbf{g}_{(p,q)}$, we learn a filter $f_{(p,q)}^{g2s}(\cdot)$ specific to the neighbor pixels in an $m \times m$ neighborhood $\mathbf{M}_{\mathbf{g}_{(p,q)}}$ and target the source pixel $\mathbf{s}_{(p,q)}$. The filter weight measuring the cross-domain correlation between $\mathbf{M}_{\mathbf{g}_{(p,q)}}$ and $\mathbf{s}_{(p,q)}$ is calculated as:

$$\mathbf{w}_{(p,q)}^{g2s} = \text{softmax}\left(\left(\mathbf{M}_{\mathbf{g}_{(p,q)}}\right)^T \mathbf{s}_{(p,q)}\right). \quad (3)$$

³ $\mathbf{N}_{s(i,j)}$ is first reshaped to a $C \times n^2$ matrix. Matrix multiplication is then taken between the transpose of the matrix and $\mathbf{g}_{(i,j)}$, which results in a $n^2 \times 1$ vector.

Filter weight $\mathbf{w}_{(i,j)}^{s2g}$ is obtained by taking a softmax normalization to the resulting vector and is also of size $n^2 \times 1$.

⁴ $\mathbf{w}_{(i,j)}^{s2g}$ weights the reshaped $\mathbf{N}_{s(i,j)}$ by taking matrix multiplication to result in a $C \times 1$ vector, i.e., $\mathbf{s}'_{(i,j)}$.

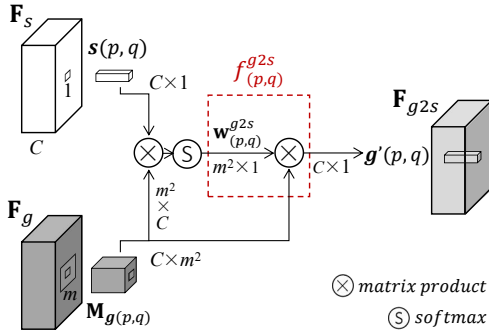


Fig. 4. An illustration of the guide-to-source modulation. Filters update the pixels in \mathbf{F}_g by targeting \mathbf{F}_s to bring \mathbf{F}_g closer to the source on modality characteristics

$\mathbf{g}_{(p,q)}$ is updated as:

$$\mathbf{g}'_{(p,q)} = f^{g^{2s}}(\mathbf{M}_{\mathbf{g}_{(p,q)}}) = \mathbf{M}_{\mathbf{g}_{(p,q)}} \mathbf{w}^{g^{2s}}. \quad (4)$$

Updating the guide pixels by considering the correlation to the pixels from the source modality domain allows our model to recognize which patterns in the guide are highly relevant to the source. Thus the guide-to-source modulation can adaptively suppress the discrepancy patterns in \mathbf{F}_g , resulting in $\mathbf{F}_{g^{2s}}$, which has modality characteristics that are close to the source and the structural cues necessary to super-resolve the source.

Ablation studies in Section 4.3 demonstrate that both the mutual modulation and the cross-domain adaptive filtering play critical roles in developing a model that can yield results with high spatial resolution and faithful image modality.

3.2 Cycle-Consistent Self-Supervised Learning

One technical challenge in self-supervised cross-modal SR is using source and guide images that cannot form direct LR-to-HR mapping to train SR models.

We argue the SR result should stay fully in the source modality domain, and therefore train our model with a cycle consistency constraint in which the start is the LR source along with the HR guide, while the end is still the LR source, as illustrated in Fig. 2. In the forward mapping, our network works as a regularizer that optimizes both the source and the guide to make a prediction induced to reach the guide in terms of spatial resolution and be faithful to the source in terms of image modality. In the backward mapping, we incentivize the consistency between the downsampled prediction and the original LR source input by minimizing l_1 norm:

$$\mathbb{C} = \left\| f_{down}(f_{net}(\mathbf{I}_{source}^{LR}, \mathbf{I}_{guide}^{HR})) - \mathbf{I}_{source}^{LR} \right\|_1, \quad (5)$$

where f_{net} denotes our network and f_{down} denotes average pooling downsampling. Our mutual modulation strategy enables our MMSR to successfully avoid

a trivial solution of Equation (5), i.e., an identity function for f_{net} . Experimental support is provided in Section 4.3.

Unlike other self-supervised methods that utilize pseudo or weak supervision in LR space, our model starts from the source modality, seeks an optimal prediction in HR space, and then constrains itself by getting back the start. In this way, both high resolution and faithful modality can be achieved and the whole process is fully self-supervised.

4 Experiments

4.1 Experimental Settings

Network Architecture. We adopt conventional convolution layers and the residual block from [30] to construct our network. The feature extraction branch of the source image (source branch) and the feature extraction branch of the guide image (guide branch) each consists of two convolution layers and two residual blocks. The image prediction part contains three residual blocks and one convolution layer. In the source branch and the prediction part, the convolution kernel size is set as 1×1 . In the guide branch, the kernel size is 3×3 . Before the prediction part, a 1×1 convolution is adopt to fuse \mathbf{F}_{s2g} and \mathbf{F}_{g2s} . The number of channels is 1 for the first convolution in source branch and the last convolution in prediction part; is 3 for the first convolution in guide branch; is 64 for the other convolutions.

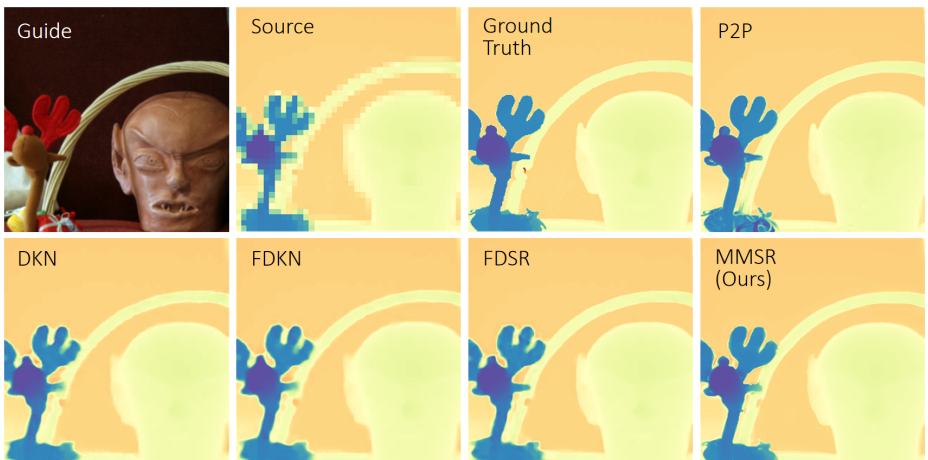
Implementation Details. We implement our MMSR model with PyTorch on an NVIDIA RTX 3090 GPU, and train it through the cycle consistency loss in Equation (5) for 1000 epochs on each combination of LR source and HR guide. Adam optimizer [20] is employed. The learning rate starts from 0.002 and decays by 0.9998 every 5 epochs. We do not use data augmentation techniques [47].

Comparison Methods. We compare MMSR with five state-of-the-art cross-modal SR methods, including three supervised methods (FDSR [16], DKN [19], and FDKN [19]) and two self-supervised methods (CMSR [47] and P2P [36]). We implement these methods fully following the settings suggested in their papers.

Datasets and Evaluation Metric. We conduct experiments on depth, DEM, and thermal modalities. For depth SR, we sample three test sets from the widely used Middlebury 2003 [46], 2005 [45], and 2014 [44] benchmarks. These three test sets include 14, 37, and 43 visible-depth pairs of size 320×320 , respectively. The three supervised comparison methods are trained on 1000 data pairs from the NYU v2 benchmark [21]. For DEM SR, we choose the remote sensing data used in the 2019 IEEE GRSS Data Fusion Contest (DFC) [22] and create a test set that includes 54 visible-DEM pairs. We train the supervised methods on 1000 data pairs. We follow the protocols in [35,36] and adopt pooling to generate LR depth and DEM. For thermal SR, we use the visible and thermal hyperspectral remote sensing data from the 2014 IEEE GRSS DFC [29], and select one band from the original thermal hyperspectral imagery as LR source. As the evaluation metric, we use the Root Mean Squared Error (RMSE).

Table 1. Depth SR on the Middlebury 2003, 2005, and 2014 datasets. We report the average RMSE. The best and the second best results are in red and blue, respectively

Dataset	Scale	Supervised			Self-Supervised		
		DKN [19]	FDKN [19]	FDSR [16]	P2P [36]	CMSR [47]	Ours
2003	$\times 4$	2.11	1.84	1.83	2.94	2.52	1.78
	$\times 8$	2.71	2.74	2.55	3.03	-	2.63
2005	$\times 4$	3.14	2.79	2.74	3.78	3.51	2.47
	$\times 8$	4.45	4.52	4.27	3.99	-	3.92
2014	$\times 4$	2.88	2.51	2.41	3.90	2.87	2.30
	$\times 8$	4.21	4.06	4.00	4.13	-	3.60

**Fig. 5.** $\times 8$ depth SR results on the Middlebury 2005 dataset. All the compared methods [47,16,19], except for P2P [36], take both LR source and HR guide as input

4.2 Evaluation on Benchmark Depth Data

In this section, we compare MMSR with five state-of-the-art methods [47,16,19,36] on the Middlebury benchmarks. In the source-to-guide modulation of MMSR, the neighborhood size for filtering is set as 11×11 . In the guide-to-source modulation, it is 5×5 . The effect of the neighborhood size is analyzed in Section 4.3.

Table 1 quantitatively reports $\times 4$ and $\times 8$ SR results. We do not provide the $\times 8$ SR results of CMSR [47] because its training settings for high scale factors is not reported in its paper. As we can see, our MMSR consistently outperforms previous self-supervised methods [47,36], as well as fully supervised methods [16,19] that are trained under the supervision from HR source.

Fig. 5 and Fig. 6 visualize $\times 8$ SR results on the Middlebury 2005 and 2014 datasets. We can observe that performing cross-modal SR as weakly-supervised cross-modal transformation allows P2P [36] to maintain the resolution of the

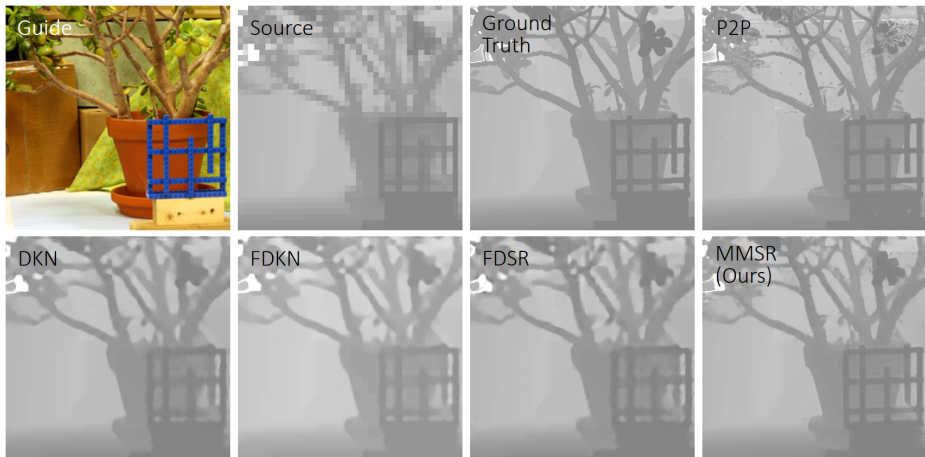


Fig. 6. $\times 8$ depth SR results on the Middlebury 2014 dataset

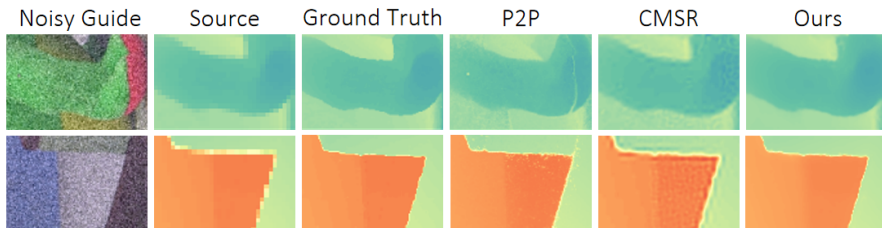


Fig. 7. $\times 4$ depth SR on the Middlebury 2003 dataset under noisy guidance. ‘Noisy Guide’ is generated by adding Gaussian noise with noise level 50

guide yet incurs serious discrepancy artifacts. FDSR [16], DKN [19], and FDKN [19] produce results that are faithful to the source modality but spatially blurry, because supervised methods cannot easily generalize well the test data. In contrast, our MMSR does not require external training and optimizes both the source and the guide with a cycle-consistent constraint, thus achieving strong generalization performance and resulting in both high spatial resolution and faithful modality characteristics.

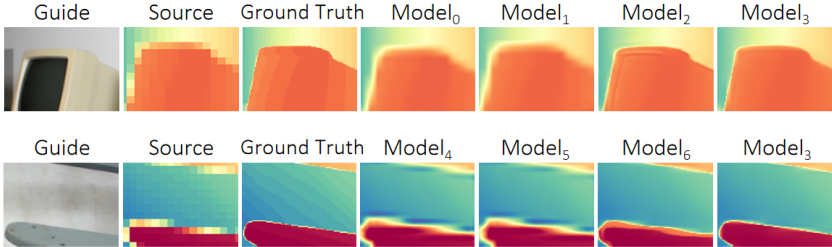
In Fig. 7, we further compare our MMSR with the two self-supervised methods (CMSR [47] and P2P [36]) to study their robustness to noisy guidance. Thanks to our mutual modulation strategy which filters and updates multimodal inputs by considering their correlation at a pixel level, our MMSR shows stronger robustness to guide images with heavy noise.

4.3 Ablation Study

We analyze MMSR by observing $\times 8$ SR results on the Middlebury 2014 dataset.

Table 2. Effectiveness study of mutual modulation strategy

	Model ₀	Model ₁	Model ₂	Model ₃
Source-to-Guide	✗	✗	✓	✓
Guide-to-Source	✗	✓	✗	✓
RMSE	4.30	4.08	3.72	3.67

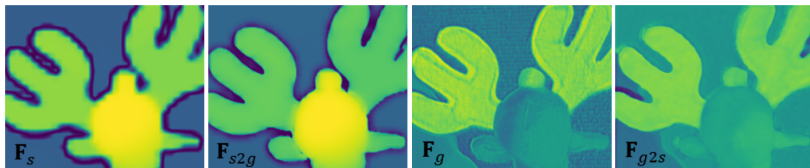
**Fig. 8.** SR results from models with different modulation settings (upper row) or variant filters (lower row)

Mutual Modulation Strategy. We first clarify the effectiveness and necessity of the proposed source-to-guide and guide-to-source modulations with setting the both neighborhood sizes for filtering as 11×11 . Table 2 reports models that adopt different modulation settings. Visual examples are shown in Fig. 8 (upper row). We can see Model₀ without mutual modulation tends toward the unimproved solution of Equation (5). Model₁ with only the guide-to-source modulation also produces blurred results since the input source stays unimproved. Model₂ yields better performance as the source-to-guide modulation increases the resolution of the source. By performing mutual modulation between source and guide, Model₃ successfully overcomes the challenge that limits existing self-supervised methods, and yields results that are HR and faithful to the source modality, though no any supervision from ground truth is given.

Cross-Domain Adaptive Filtering. Based on Model₃, we highlight our developed filters with regard to two factors: (1) cross-domain learning (our filters measure the correlation across the source and guide modality domains) and (2) adaptiveness (our filter weights are variant for different pixel neighborhoods). Table 3 compares Model₃ and three models with variant filters. Visual examples are in Fig. 8 (lower row). Note that, in Model₃, our filters $f_{(i,j)}^{s2g}$ and $f_{(p,q)}^{g2s}$ are of size 11×11 . In Model₄, we replace $f_{(i,j)}^{s2g}$ with an 11×11 convolution to filter \mathbf{F}_s , replace $f_{(p,q)}^{g2s}$ with another 11×11 convolution to filter \mathbf{F}_g , and fuse the filtering products using a 1×1 convolution. In Model₅, we first concatenate \mathbf{F}_s and \mathbf{F}_g , then fuse them using an 11×11 convolution. Due to insufficient consideration of the dependency between source and guide domains and the weight invariance of conventional convolutions [9], Model₄ and Model₅ show inferior performance. In

Table 3. Effectiveness study of cross-domain adaptive filters

	Model ₄	Model ₅	Model ₆	Model ₃
Cross-Domain	✗	✓	✗	✓
Adaptive	✗	✗	✓	✓
RMSE	4.87	4.88	3.84	3.67

**Fig. 9.** Visualization of features before and after cross-domain adaptive filtering

Model₆, we change the target pixel of $f_{(i,j)}^{s2g}(\cdot)$ (i.e., $\mathbf{g}_{(i,j)}$) to $\mathbf{s}_{(i,j)}$, and change the target pixel of $f_{(p,q)}^{g2s}(\cdot)$ (i.e., $\mathbf{s}_{(p,q)}$) to $\mathbf{g}_{(p,q)}$, resulting in adaptive filters $f_{(i,j)}^{s2s}(\cdot)$ and $f_{(p,q)}^{g2g}(\cdot)$ similar to non-local filtering [31]. As we can observe, Model₆ gets improvement by suppressing artifacts caused by bilinear interpolation, but is still inferior as its filters cannot update \mathbf{F}_s and \mathbf{F}_g properly due to the no measurement of cross-domain correlations. In Model₃, our cross-domain adaptive filters $f_{(i,j)}^{s2g}(\cdot)$ and $f_{(p,q)}^{g2s}(\cdot)$ fully exploit cross-modal spatial dependency, update \mathbf{F}_s by considering pixel correlation to \mathbf{F}_g from the HR guide domain, and update \mathbf{F}_g by considering pixel correlation to \mathbf{F}_s from the source modality domain. Fig. 9 visualizes features before and after filtering. Our filters optimize the resolution of the source feature and suppress the discrepancy of the guide feature, enabling an effective inductive bias and the superior performance of Model₃.

Effect of Asymmetric Neighborhood Sizes. As introduced in Sec. 3.1, our mutual modulation is driven by filtering confined neighborhoods in the source and guide features. In general image restoration, properly increasing filtering size benefits model performance [31]. In our experiments, to Model₃ in Table 2, if the neighborhood sizes in the source-to-guide and guide-to-source modulations are both set as 3×3 or 7×7 , the obtained RMSE values are correspondingly 4.23 and 3.91. When both are increased to 11×11 , as in Table 2, the RMSE is 3.67. Considering GPU memory limitations, we did not increase the sizes further. Since our modulation strategy is bidirectional, we further investigate the effect of asymmetric neighborhood sizes. Based on Model₃, we fix the neighborhood size in the source-to-guide modulation as 11×11 , while reducing that in the guide-to-source modulation, as shown in Fig. 10. The performance peaks at 5×5 , which shows that there is an optimal setting on specific types of image data. On the Middlebury data, when the neighborhood sizes in the source-to-guide and guide-to-source modulations are respectively 11×11 and 5×5 , our model can

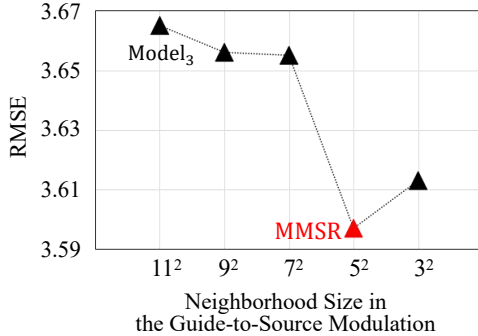


Fig. 10. Effect of asymmetric neighborhood sizes. The neighborhood in the source-to-guide modulation is fixed as 11×11

Table 4. DEM SR evaluation by observing average RMSE. The best and the second best results are in red and blue, respectively

Scale	Supervised			Self-Supervised		
	DKN [19]	FDKN [19]	FDSR [16]	P2P [36]	CMSR [47]	MMSR (Ours)
$\times 4$	0.80	0.80	0.81	1.57	0.78	0.73
$\times 8$	1.39	1.25	1.55	1.70	-	1.02

modulate the source and the guide optimally. Therefore, we adopt this setting to our model in Section 4.2. When we fixed the neighborhood in the guide-to-source modulation and reduced that in the source-to-guide modulation, the results were overly decided by the source. Visual results and more analyses of these two cases are in the supplementary material.

4.4 Validation on Real-World DEM and Thermal

Given the importance of SR techniques in Earth observation, we apply our MMSR to real-world remote sensing data that covers DEM and thermal modalities. The neighborhood sizes in the source-to-guide and guide-to-source modulations are set as 5×5 and 3×3 , respectively.

We compare MMSR with the five state-of-the-art methods [47,16,19,36] in terms of $\times 4$ and $\times 8$ DEM SR. As reported in Table 4, our MMSR yields the best quantitative performance under both scale factors, and outperforms the three supervised methods by a large margin. It can be seen qualitatively from Fig. 11 that our MMSR shows superiority by preserving finer spatial details for the buildings and plants and the modality characteristics of DEM.

We further compare MMSR with CMSR [47] on the visible-thermal data from [29]. We do not provide numerical evaluation since only LR thermal data is available. As presented in Fig. 12, our MMSR shows robust performance and superior generalizability. More visual results are in the supplementary material.

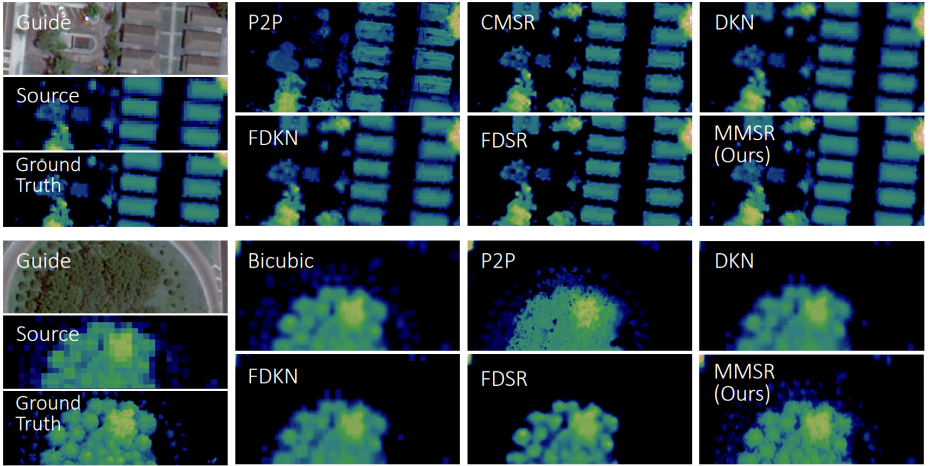


Fig. 11. DEM SR. The upper and lower rows show $\times 4$ and $\times 8$ SR results, respectively



Fig. 12. $\times 5$ thermal SR. (a) Guide. (b) Source. Results from (c) CMSR and (d) Ours

5 Conclusions

We study cross-modal SR and present a robust self-supervised MMSR model. Within MMSR, we introduce a mutual modulation strategy to overcome the LR problem of the source and the discrepancy problem of the guide, and adopt a cycle consistency constraint to conduct training in a fully self-supervised manner, without accessing ground truth or external training data. We demonstrate the superior generalizability of our MMSR on depth, DEM, and thermal modalities, and its applicability to noisy data and real-world remote sensing data. Extensive experiments demonstrate the state-of-the-art performance of our MMSR.

We believe our concept of modulating different modalities to achieve self-supervised cross-modal SR can inspire further progress in this field, and believe our MMSR can contribute to Earth observation applications where images in various modalities are available but HR ones are rare and expensive.

Acknowledgements. XD was supported by the RIKEN Junior Research Associate (JRA) Program. NY was supported by JST, FOREST Grant Number JPMJFR206S, Japan.

References

1. Adriano, B., Yokoya, N., Xia, J., Miura, H., Liu, W., Matsuoka, M., Koshimura, S.: Learning from multimodal and multitemporal earth observation data for building damage mapping. *ISPRS Journal of Photogrammetry and Remote Sensing* **175**(1), 132–143 (2021)
2. Almasri, F., Debeir, O.: Multimodal sensor fusion in single thermal image super-resolution. In: *ACCV* (2018)
3. Arar, M., Ginger, Y., Danon, D., Bermano, A.H., Cohen-Or, D.: Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In: *CVPR*. pp. 13407–13416 (2020)
4. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: *CVPR* (2005)
5. Burger, H., Schuler, C., Harmeling, S.: Image denoising: Can plain neural networks compete with bm3d? In: *CVPR* (2012)
6. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing* **16**(8), 2080–2095 (2007)
7. Diebel, J., Thrun, S.: An application of markov random fields to range sensing. In: *NeurIPS* (2005)
8. Dong, J., Pan, J., Ren, J., Lin, L., Tang, J., Yang, M.H.: Learning spatially variant linear representation models for joint filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021)
9. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: *CVPR* (2021)
10. Ferstl, D., Reinbacher, C., Ranftl, R., Ruether, M., Bischof, H.: Image guided depth upsampling using anisotropic total generalized variation. In: *ICCV* (2013)
11. Gu, S., Guo, S., Zuo, W., Chen, Y., Timofte, R., Van Gool, L., Zhang, L.: Learned dynamic guidance for depth image reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**(10), 2437–2452 (2020)
12. Gu, S., Zuo, W., Guo, S., Chen, Y., Chen, C., Zhang, L.: Learning dynamic guidance for depth image enhancement. In: *CVPR*. pp. 712–721 (2017)
13. Hajjar, A.J.: In-depth guide to self-supervised learning: Benefits & uses. (2020) <https://research.aimultiple.com/self-supervised-learning/>
14. He, J., Dong, C., Qiao, Y.: Modulating image restoration with continual levels via adaptive feature modification layers. In: *CVPR*. p. 11056–11064 (2019)
15. He, K., Sun, J., Tang, X.: Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(6), 1397–1409 (2013)
16. He, L., Zhu, H., Li, F., Bai, H., Cong, R., Zhang, C., Lin, C., Liu, M., Zhao, Y.: Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In: *CVPR* (2021)
17. Hui, T.W., Loy, C.C., Tang, X.: Depth map super-resolution by deep multi-scale guidance. In: *ECCV* (2016)
18. Kalal, Z., Mikolajczyk, K., Matas, J.: Forward-backward error: Automatic detection of tracking failures. In: *ICPR*. pp. 2756–2759 (2010)
19. Kim, B., Ponce, J., Ham, B.: Deformable kernel networks for joint image filtering. *International Journal of Computer Vision* **129**(4), 579–600 (2021)
20. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: *ICLR* (2015)
21. Kohli, P., Silberman, N., Hoiem, D., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: *ECCV*. pp. 746–760 (2012)

22. Kunwar, S., Chen, H., Lin, M., Zhang, H., D’Angelo, P., Cerra, D., Azimi, S.M., Brown, M., Hager, G., Yokoya, N., Hänsch, R., Le Saux, B.: Large-scale semantic 3-d reconstruction: Outcome of the 2019 ieee grss data fusion contest—part a. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **14**, 922–935 (2020)
23. Kwon, H., Tai, Y.W., Lin, S.: Data-driven depth map refinement via multi-scale sparse representation. In: *CVPR*. p. 159–167 (2015)
24. Lee, J., Chung, S.W., Kim, S., Kang, H.G., Sohn, K.: Looking into your speech: Learning cross-modal affinity for audio-visual speech separation. In: *CVPR*. pp. 1336–1345 (2021)
25. Lee, S., Kim, D., Han, B.: CoSMo: Content-style modulation for image retrieval with text feedback. In: *CVPR*. pp. 802–812 (2021)
26. Li, Y., Xue, T., Sun, L., Liu, J.: Joint example-based depth map super-resolution. In: *ICME*. p. 152–157 (2012)
27. Li, Y., Huang, J.B., Ahuja, N., Yang, M.H.: Deep joint image filtering. In: *ECCV*. pp. 154–169 (2016)
28. Li, Y., Huang, J.B., Ahuja, N., Yang, M.H.: Joint image filtering with deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(8), 1909–1923 (2019)
29. Liao, W., Huang, X., Van Coillie, F., Gautama, S., Pižurica, A., Philips, W., Liu, H., Zhu, T., Shimoni, M., Moser, G., Tuia, D.: Processing of multiresolution thermal hyperspectral and digital color data: Outcome of the 2014 ieee grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **8**(6), 2984–2996 (2015)
30. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: *CVPRW* (2017)
31. Liu, D., Wen, B., Fan, Y., Loy, C.C., Huang, T.S.: Non-local recurrent network for image restoration. In: *NeurIPS*. pp. 1680–1689 (2018)
32. Liu, J., Lu, M., Chen, K., Li, X., Wang, S., Wang, Z., Wu, E., Chen, Y., Zhang, C., Wu, M.: Overfitting the data: Compact neural video delivery via content-aware feature modulation. In: *ICCV* (2021)
33. Liu, M.Y., Tuzel, O., Taguchi, Y.: Joint geodesic upsampling of depth images. In: *CVPR*. p. 169–176 (2013)
34. Lu, J., Forsyth, D.: Sparse depth super resolution. In: *CVPR*. p. 2245–2253 (2015)
35. Lutio, R.d., Becker, A., D’Aronco, S., Russo, S., Wegner, J.D., Schindler, K.: Learning graph regularisation for guided super-resolution. In: *CVPR* (2022)
36. Lutio, R.d., D’Aronco, S., Wegner, J.D., Schindler, K.: Guided super-resolution as pixel-to-pixel transformation. In: *ICCV*. pp. 8828–8836 (2019)
37. Maeda, S.: Unpaired image super-resolution using pseudo-supervision. In: *CVPR* (2020)
38. Mei, Y., Fan, Y., Zhou, Y.: Image super-resolution with non-local sparse attention. In: *CVPR*. pp. 3517–3526 (2021)
39. Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T.S., Shi, H.: Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In: *CVPR* (2020)
40. Pan, J., Dong, J., Ren, J., Lin, L., Tang, J., Yang, M.H.: Spatially variant linear representation models for joint filtering. In: *CVPR*. pp. 1702–1711 (2019)
41. Park, J., Kim, H., Tai, Y.W., Brown, M.S., Kweon, I.: High quality depth map upsampling for 3d-tof cameras. In: *ICCV*. p. 1623–1630 (2011)
42. Perez, P., Gangnet, M., Blake, A.: Guided image filtering. *ACM Transactions on Graphics* **22**(3), 313–318 (2003)

43. Rudin, L., Osher, S.: Total variation based image restoration with free local constraints. In: *ICIP* (1994)
44. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nesci, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: *GCPR* (2014)
45. Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: *CVPR* (2007)
46. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: *CVPR*. pp. 195–202 (2003)
47. Shacht, G., Danon, D., Fogel, S., Cohen-Or, D.: Single pair cross-modality super resolution. In: *CVPR* (2021)
48. Shao, Y., Li, L., Ren, W., Gao, C., Sang, N.: Domain adaptation for image dehazing. In: *CVPR* (2020)
49. Shim, G., Park, J., Kweon, I.S.: Robust reference-based super-resolution with similarity-aware deformable convolution. In: *CVPR* (2020)
50. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgb-d images. In: *ECCV*. pp. 746–760 (2012)
51. Su, H., Jampani, V., Sun, D., Gallo, O., Learned-Miller, E., Kautz, J.: Pixel-adaptive convolutional neural networks. In: *CVPR*. pp. 11166–11175 (2019)
52. Sun, J., Jia, J., Tang, C.K., Shum, H.Y.: Poisson matting. In: *ACM Siggraph* (2004)
53. Sundaram, N., Brox, T., Keutzer, K.: Dense point trajectories by gpu-accelerated large displacement optical flow. In: *ECCV* (2010)
54. Tang, J., Chen, X., Zeng, G.: Joint implicit image function for guided depth super-resolution. In: *ACMMM* (2021)
55. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: TDAN: Temporally-deformable alignment network for video super-resolution. In: *CVPR* (2020)
56. Uezato, T., Hong, D., Yokoya, N., He, W.: Guided deep decoder: Unsupervised image pair fusion. In: *ECCV* (2020)
57. Wang, L., Guo, Y., Dong, X., Wang, Y., Ying, X., Lin, Z., An, W.: Exploring fine-grained sparsity in convolutional neural networks for efficient inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022)
58. Wang, L., Wang, Y., Dong, X., Xu, Q., Yang, J., An, W., Guo, Y.: Unsupervised degradation representation learning for blind super-resolution. In: *CVPR* (2021)
59. Wang, W., Guo, R., Tian, Y., Yang, W.: CFSNet: Toward a controllable feature space for image restoration. In: *ICCV*. p. 4140–4149 (2019)
60. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: *CVPR* (2021)
61. Wang, X., Yu, K., Dong, C., Tang, X., Loy, C.C.: Deep network interpolation for continuous imagery effect transition. In: *CVPR*. p. 1692–1701 (2019)
62. Wang, Y., Wang, L., Liang, Z., Yang, J., An, W., Guo, Y.: Occlusion-aware cost constructor for light field depth estimation. In: *CVPR* (2022)
63. Wang, Y., Wang, L., Wu, G., Yang, J., An, W., Yu, J., Guo, Y.: Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022)
64. Xu, G., Xu, J., Li, Z., Wang, L., Sun, X., Cheng, M.M.: Temporal modulation network for controllable space-time video super-resolution. In: *CVPR*. pp. 6388–6397 (2021)
65. Yang, L., Wang, Y., Xiong, X., Yang, J., Katsaggelos, A.K.: Efficient video object segmentation via network modulation. In: *CVPR* (2018)

66. Yang, Q., Yang, R., Davis, J., Nister, D.: Spatial-depth super resolution for range images. In: CVPR (2007)
67. Zhang, K., Li, Y., Zuo, W., Zhang, L., Van Gool, L., Timofte, R.: Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021)
68. Zhang, P., Zhang, B., Chen, D., Yuan, L., Wen, F.: Cross-domain correspondence learning for exemplar-based image translation. In: CVPR (2020)
69. Zhang, Y., Li, K., Li, K., Fu, Y.: MR image super-resolution with squeeze and excitation reasoning attention network. In: CVPR. pp. 13425–13434 (2021)
70. Zhang, Y., Li, K., Li, K., Zhong, B., Fu, Y.: Residual non-local attention networks for image restoration. In: ICLR (2019)
71. Zhou, T., Jae Lee, Y., Yu, S.X., Efros, A.A.: FlowWeb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In: CVPR (2015)
72. Zhou, T., Krahenbuhl, P., Aubry, M., Huang, Q., Efros, A.A.: Learning dense correspondence via 3d-guided cycle consistency. In: CVPR (2016)
73. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)

Supplementary Material: Learning Mutual Modulation for Self-Supervised Cross-Modal Super-Resolution

Xiaoyu Dong^{1,2}, Naoto Yokoya^{1,2}(✉), Longguang Wang³, and Tatsumi Uezato⁴

¹ The University of Tokyo, Tokyo, Japan

² RIKEN AIP, Tokyo, Japan

³ National University of Defense Technology, Changsha, China

⁴ Hitachi, Ltd, Tokyo, Japan

dong@ms.k.u-tokyo.ac.jp, yokoya@k.u-tokyo.ac.jp

<https://github.com/palmdong/MMSR>

Section I analyzes mutual modulation with asymmetric neighborhood sizes. Section II studies different feature fusion approaches. Section III compares the time cost of different self-supervised cross-modal super-resolution (SR) methods, and further compares their performance under noisy guidance. Section IV provides more discussions. Section V provides more qualitative results.

I Modulation with Asymmetric Neighborhood Sizes

In Section 4.3 Ablation Study, we have discussed the effect of the asymmetric neighborhood sizes in our mutual modulation.

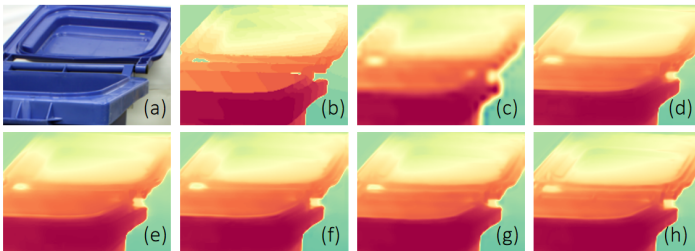


Fig. I. (a) Guide. (b) Ground truth. (c) Bicubic source. Results from models of which the neighborhood sizes for the guide-to-source modulation are (d) 11×11 , (e) 9×9 , (f) 7×7 , (g) 5×5 , and (h) 3×3 , respectively

Fig. I provides visual examples of the case in which we fixed the neighborhood size in the source-to-guide modulation as 11×11 and reduced that in the guide-to-source modulation. When the neighborhood size is reduced to 9×9 , the result (Fig. I(e)) is lack of details, because the spatial suppression to the guide is strong. When it is further reduced to 3×3 , the result (Fig. I(h)) has extraneous structures, because the suppression to the spatial discrepancy in the guide is

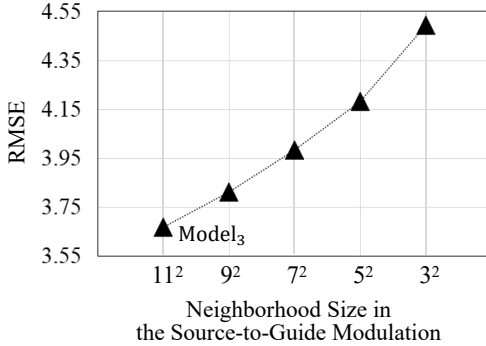


Fig. II. Effect of asymmetric neighborhood sizes. The neighborhood in the guide-to-source modulation is fixed as 11×11

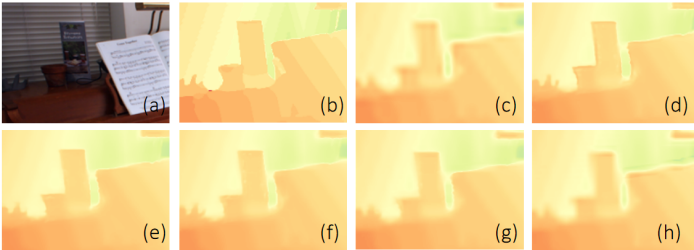


Fig. III. (a) Guide. (b) Ground truth. (c) Bicubic source. Results from models of which the neighborhood sizes for the source-to-guide modulation are (d) 11×11 , (e) 9×9 , (f) 7×7 , (g) 5×5 , and (h) 3×3 , respectively

weak. When it is set as 5×5 , the result (Fig. I(g)) is close to the ground truth (Fig. I(b)) with respect to both spatial resolution and modality characteristics.

Fig. II reports the quantitative results of the case in which we fixed the neighborhood size in the guide-to-source modulation as 11×11 and reduced that in the source-to-guide modulation. Fig. III provides visual examples. As the neighborhood size reduces, the results become blurry. This is because the strength to increase the resolution of the source is reduced.

In summary, our mutual modulation allows to handle different types of multi-modal data flexibly. With setting a large neighborhood size for the source-to-guide modulation and a properly small neighborhood size for the guide-to-source modulation, models can optimally increase the resolution of the source and capture and suppress the spatial discrepancy in the guide.

II Different Feature Fusion Approaches

In MMSR, the modulated features \mathbf{F}_{s2g} and \mathbf{F}_{g2s} are fused by a 1×1 convolution. We additionally studied other fusion approaches, including naive summation

Table I. $\times 4$ depth SR on the Middlebury 2003 dataset

	Sum.	Att. + Sum.	Att. + Conv $_{1\times 1}$	Conv $_{1\times 1}$
RMSE	1.88	1.92	1.84	1.78
Training Time	140s	150s	147s	137s

and attentional fusion (spatial attention and channel attention were performed before summation or 1×1 convolution), as reported in Table I. The model with only a 1×1 convolution as fusion approach achieves the best performance and the shortest training time. Therefore, a 1×1 convolution is adopt to fuse the modulated features in our MMSR.

III Comparisons with Other Self-Supervised Methods

Time Cost. As introduced in Section 1, self-supervised cross-modal SR methods, including CMSR [47], P2P [36], and our MMSR, perform online learning on each combination of low-resolution (LR) source and high-resolution (HR) guide. Table II compares their training time cost. Note that, the time cost of our MMSR is influenced by the modulation neighborhood sizes (i.e., modulation with larger neighborhood sizes results in higher time cost). For depth SR, the neighborhood sizes for the source-to-guide modulation and the guide-to-source modulation in our MMSR were set as 11×11 and 5×5 , respectively. The training time of MMSR/P2P/CMSR on each depth-visible input (of size 320×320) is 137s/131s/90s. Our MMSR runs slightly slower yet shows obvious performance superiority, as shown in Table II. For digital elevation model (DEM) SR, the neighborhood sizes for the source-to-guide modulation and the guide-to-source modulation in our MMSR were set as 5×5 and 3×3 , respectively. The training time of MMSR/P2P/CMSR on each DEM-visible input (of size 320×320) is 49s/131s/90s. Our MMSR requires much less time and still obtains obvious performance superiority.

Table II. t shows training time on an NVIDIA RTX 3090 GPU. RMSE_{2003} , RMSE_{2005} , and RMSE_{2014} denote the average RMSE on the Middlebury 2003 [46], 2005 [45], and 2014 [44] datasets, respectively. Numbers in brackets show the performance improvement achieved by our MMSR

	$\times 4$ Depth SR				$\times 4$ DEM SR	
	t	RMSE_{2003}	RMSE_{2005}	RMSE_{2014}	t	RMSE
P2P [36]	131s	2.94 (\uparrow 39.5%)	3.78 (\uparrow 34.7%)	3.90 (\uparrow 41.0%)	131s	1.57 (\uparrow 53.5%)
CMSR [47]	90s	2.52 (\uparrow 29.4%)	3.51 (\uparrow 29.6%)	2.87 (\uparrow 19.9%)	90s	0.78 (\uparrow 6.4%)
Ours	137s	1.78 (-)	2.47 (-)	2.30 (-)	49s	0.73 (-)

Performance under Noisy Guidance. Fig. IV further compares our MMSR with CMSR [47] and P2P [36] under noisy guidance. As we can see, under even heavy noise, our MMSR still outperforms CMSR and P2P by a large margin and can produce results that are closer to ground truth. This demonstrates the robustness of our MMSR and the effectiveness of our mutual modulation with cross-domain adaptive filtering.

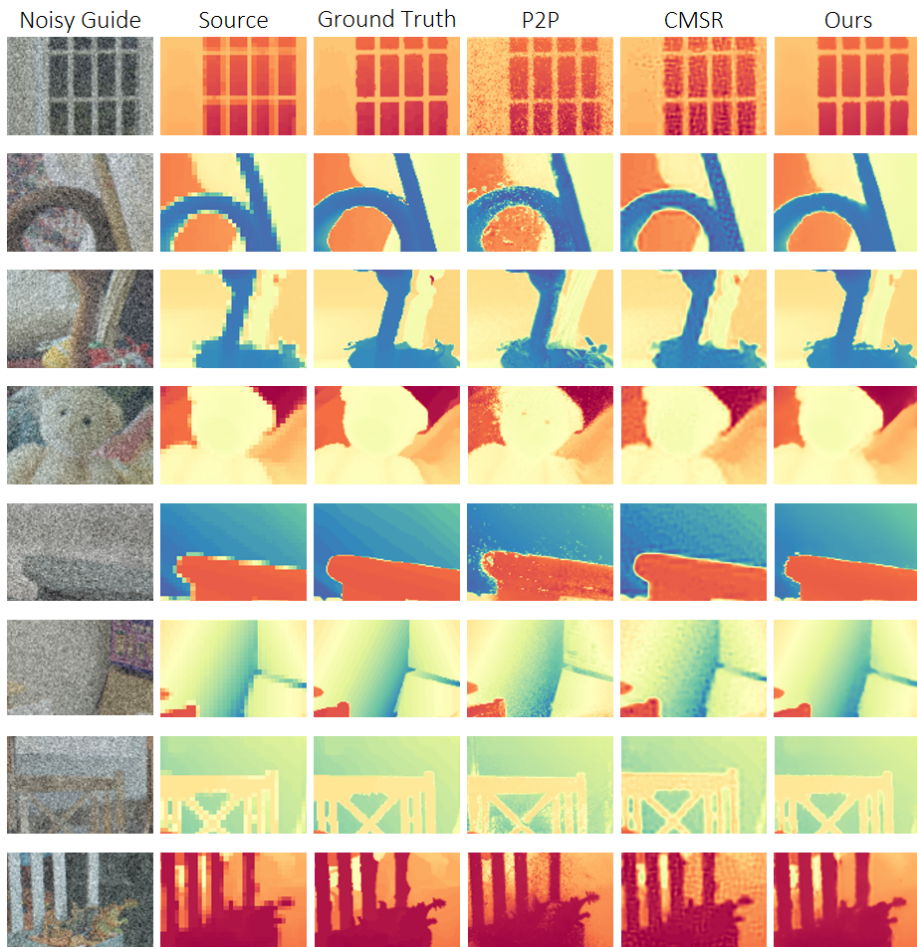


Fig. IV. $\times 4$ depth SR under noisy guidance. The first four and the second four rows show results on the Middlebury 2005 and 2014 datasets, respectively. ‘Noisy Guide’ is generated by adding Gaussian noise with noise level 50

IV More Discussions

What Is Important in Cross-Modal SR? Given an LR source and an HR guide from different modalities, cross-modal SR aims at achieving an image product that has spatial resolution comparable with the guide and modality characteristics faithful to the source. We argue both the structural cues from the HR guide and the modality constraint from the LR source are important in the task. Thus we develop a mutual modulation strategy and adopt cycle consistency constraint to fully exploit the guide and also the source, enabling a robust self-supervised MMSR model.

Why Can MMSR Outperform Supervised Methods? Supervised cross-modal SR methods have shown promising performance. However, they have two problems: **(1)** They suffer limited performance in real-world scenes because large-scale paired training data is hard to acquire. **(2)** They cannot easily generalize well to test data that is not in the same domain as the training data. The reasons of our superior performance are twofold: **(1)** Our mutual modulation strategy and cycle-consistent self-supervised learning effectively facilitate our MMSR to achieve state-of-the-art performance. **(2)** The employed online learning scheme allows our MMSR a strong generalization capability to any given input. With robust performance and strong generalizability, MMSR can outperform even supervised methods.

Contributions beyond Superior Performance. Our MMSR outperforms previous supervised and self-supervised methods on various tasks. Moreover, our work also has the following three major contributions: **(1)** The state-of-the-art performance of our MMSR bridges the gap of robust self-supervised cross-modal SR. **(2)** For the first time, our mutual modulation effectively overcomes the spatial discrepancy and resolution gap of multi-modal images, and show correlation-based filtering provides an effective inductive bias for deep cross-modal SR. This benefits further progress in research fields. **(3)** Our MMSR shows superior generalization capability to diverse modalities, robustness to noise, and applicability to real-world scenarios. This is beneficial to real-world applications.

Limitation. Like other methods, MMSR produces ghosting artifacts on some samples. In Fig. 5, ghosting artifacts can be observed around the antlers in the results of FDSR [16], FDKN [19], DKN [19], and MMSR. This is caused by the bicubic/bilinear upsampled source input. Since P2P [36] inputs only the guide image, it does not suffer from ghosting artifacts but produces discrepancy artifacts. Likewise, in Fig. 9, in feature \mathbf{F}_{g2s} , the ghosting along the antler is because the guide-to-source modulation induces \mathbf{F}_g to mimic \mathbf{F}_s which has bilinear ghosting. However, compared with previous state-of-the-art methods [16,47,19,36], our MMSR achieves final predictions that are closer to ground truth. Exploring the upperbound performance of self-supervised cross-modal SR models would be an interesting and challenging research problem.

V More Qualitative Results

We provide more visual comparisons between our MMSR and the five cross-modal SR methods [47,16,19,36]. Fig. V, Fig. VI, and Fig. VII show SR results on the depth-visible data from the Middlebury 2003 [46], 2005 [45], and 2014 [44] benchmarks, respectively. Fig. VIII shows SR results on the real-world DEM-visible data from [22]. For depth SR, error maps are provided for better visual comparison. As we can see, our MMSR produces lower errors and finer edge details. Overall, as a self-supervised method, our MMSR achieves state-of-the-art performance on various tasks, and outperforms fully supervised methods (FDSR [16], DKN [19], and FDKN [19]) and previous self-supervised methods (CMSR [47] and P2P [36]) consistently.

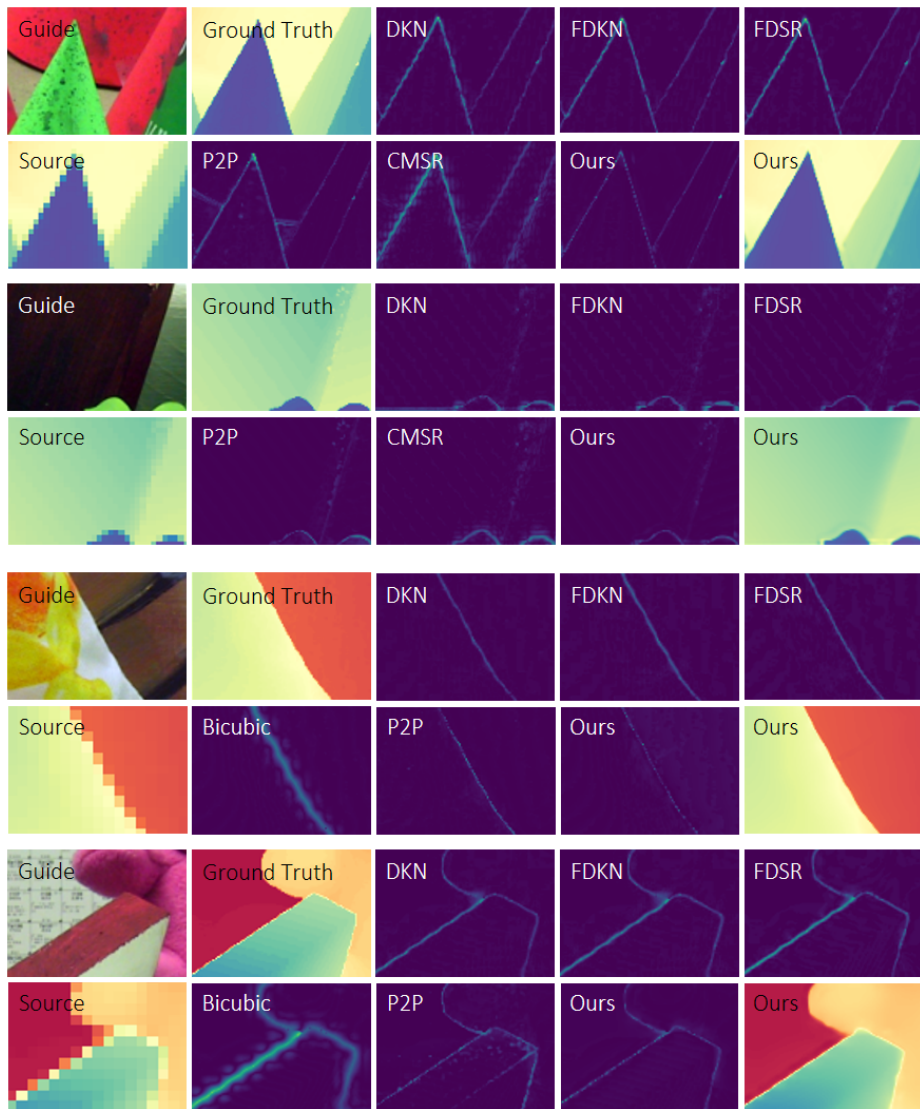


Fig. V. Depth SR on the Middlebury 2003 dataset. The first and second rows show $\times 4$ SR results. The third and fourth rows show $\times 8$ SR results.

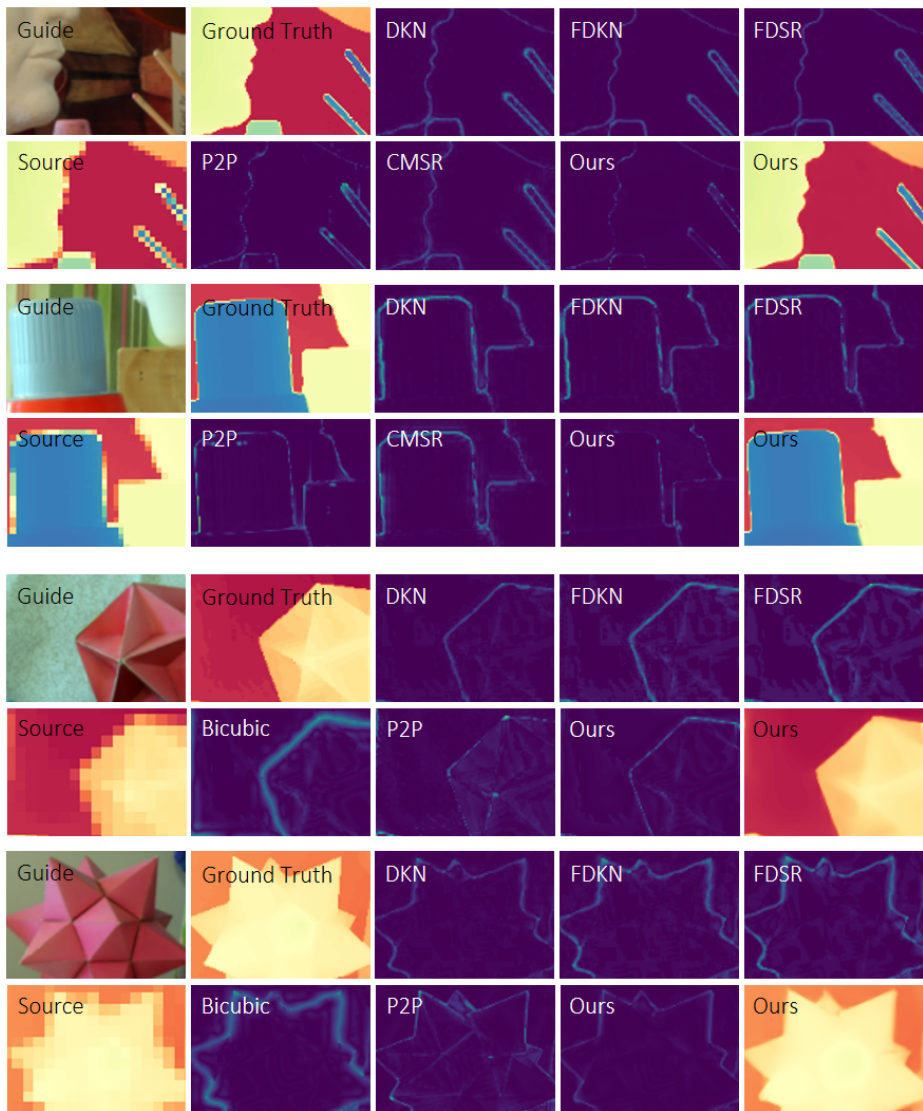


Fig. VI. Depth SR on the Middlebury 2005 dataset. The first and second rows show $\times 4$ SR results. The third and fourth rows show $\times 8$ SR results

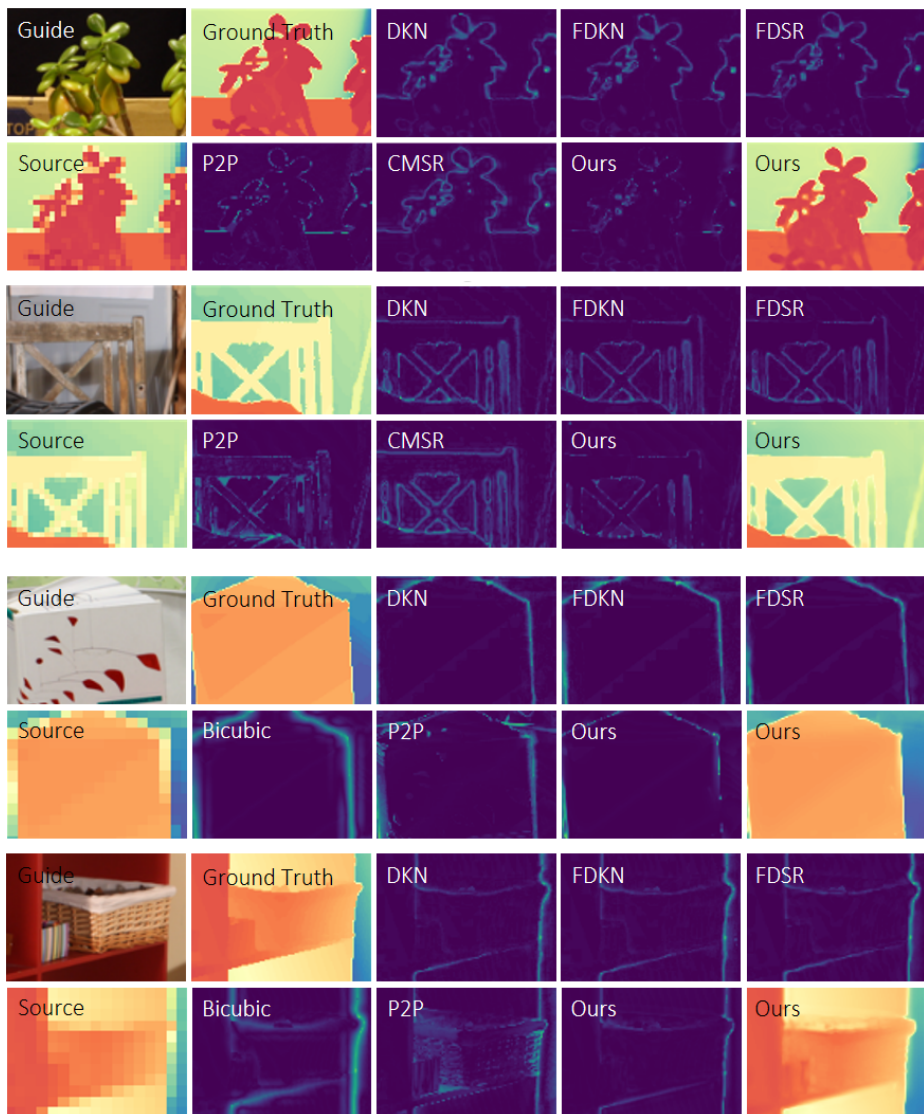


Fig. VII. Depth SR on the Middlebury 2014 dataset. The first and second rows show $\times 4$ SR results. The third and fourth rows show $\times 8$ SR results.

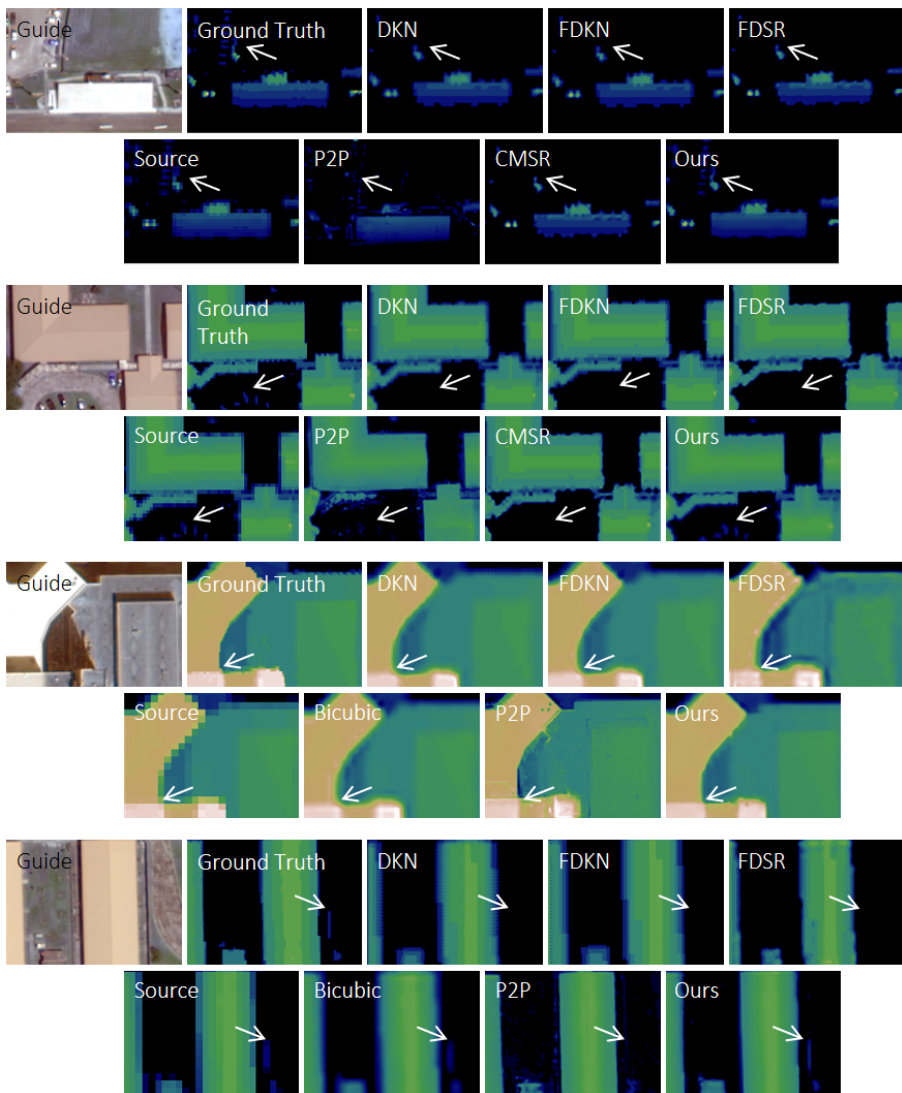


Fig. VIII. DEM SR. The first and second rows show $\times 4$ SR results. The third and fourth rows show $\times 8$ SR results