Metadata-Based RAW Reconstruction via Implicit Neural Functions

Leyi Li[1] Huijie Qiao[2] Qi Ye[1,3] Qinmin Yang[1]
[1]Zhejiang University [2]Chinese Academy of Sciences [3]Key Lab of CS&AUS of Zhejiang Province
{lileyi,qi.ye,qmyang}@zju.edu.cn qiaohj@ioz.ac.cn

Abstract

Many low-level computer vision tasks are desirable to utilize the unprocessed RAW image as input, which remains the linear relationship between pixel values and scene radiance. Recent works advocate to embed the RAW image samples into sRGB images at capture time, and reconstruct the RAW from sRGB by these metadata when needed. However, there still exist some limitations in making full use of the metadata. In this paper, instead of following the perspective of sRGB-to-RAW mapping, we reformulate the problem as mapping the 2D coordinates of the metadata to its RAW values conditioned on the corresponding sRGB values. With this novel formulation, we propose to reconstruct the RAW image with an implicit neural function, which achieves significant performance improvement (more than 10dB average PSNR) only with the uniform sampling. Compared with most deep learning-based approaches, our method is trained in a self-supervised way that requiring no pre-training on different camera ISPs. We perform further experiments to demonstrate the effectiveness of our method, and show that our framework is also suitable for the task of guided super-resolution.

1. Introduction

Low-level computer vision tasks benefit a lot from the scene-referred RAW images [7, 39, 19, 17, 16], which is rendered to the display-referred standard RGB (sRGB) images via camera image signal processors (ISPs). Compared with sRGB images, typical RAW images has the advantages of linear relationship between pixel values and scene radiance, as well as higher dynamic range. However, RAW images occupy obviously more memory than the sRGB images in common format like JPEG, which is unfavourable for transferring and sharing. Moreover, since most display and printing devices are designed for images stored and shared in sRGB format, it is inconvenient to directly replace sRGB with RAW. Consequently, mapping sRGB images back to their RAW counterparts, which is also called RAW reconstruction, is regarded as the appropriate way to utilize the advantage of RAW data [23, 25, 36, 10, 28, 24].

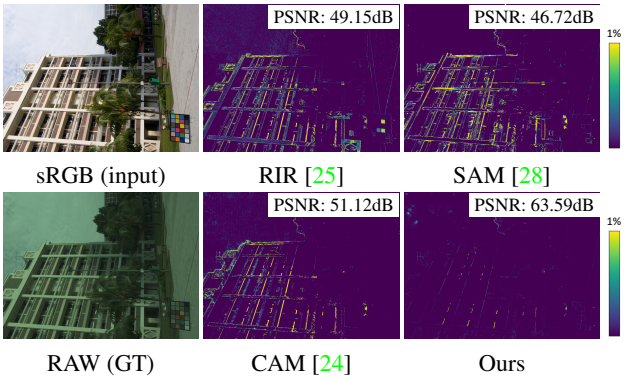


Figure 1. As RAW images are beneficial to many low-level computer vision tasks, we aim to reconstruct the RAW image from the corresponding sRGB image with the assistance of extra metadata. In this figure, the reconstructed RAW images are visualized through error maps. As can be seen, our method remarkably outperforms other related methods with the improvement of more than 10 dB PSNR. We owe this performance boost to the effectiveness of our proposed *implicit neural function (INF)* .

Early RAW reconstruction methods focus on building standard models to reverse ISPs, which is parameterized by either explicit functions [4, 18, 14, 5] or neural networks [23, 36, 10]. However, these approaches are faced with the same issue that a parameterized model is only suitable for a specific ISP. Meanwhile, a series of methods [25, 26, 28, 24] propose to overcome this problem by embedding extra metadata into sRGB images at capture time. For such methods, the main challenge is to improve the accuracy with lower metadata generation cost. RIR [25] implements complex optimization algorithm to estimate the global mapping parameters as metadata, but suffers high computational cost. SAM [28] adopts a uniform sampling on RAW images to generate the metadata, which is further replaced with a sampler network by CAM [24].

For the metadata-based methods of SAM [28] and CAM [24], the embedded RAW samples stores partial information of ISPs which helps to reconstruct the RAW images better; also, by conditioning the reconstruction algorithm on

these metadata, the recovery of the RAW data turns into a conditional mapping function instead of a function fitted to a specific case, enabling the potential to achieve better generalization. Therefore, we adopt this strategy in the paper.

Despite the progress that SAM [28] and CAM [24] have made, there still exist some limitations for the metadata-based RAW reconstruction methods. SAM utilizes RBF interpolation [3], the main idea of which is to calculate the difference between sampling and target points by a kernel function. However, a fixed kernel function lacks the flexibility to model various sRGB-to-RAW mappings. CAM directly uses a neural network for reconstruction but requires pre-training on pairs of sRGB and raw data from different types of ISPs. Also, we observe that the results of these methods fail to recover the saturated regions [26] (i.e., pixels with any channel value close to the maximum), as is shown in Figure 1.

To address the limitation, we propose a two-way RAW reconstruction algorithm based on an implicit neural function (INF). Previously, RAW reconstruction is formulated as mapping a sRGB image and the metadata to its RAW image. In this paper, we reformulate the problem as mapping the 2D coordinates of the metadata to its RAW values conditioned on the corresponding sRGB values, i.e. an implicit function. With this novel formulation, we can also decompose the problem into two aspects: a mapping function from the sRGB values to the corresponding raw values; a super-resolution function to interpolate the RAW image from the sparse samples. We observe that the super-resolution part usually exhibits much higher errors, indicating the latter a more challenging task. Accordingly, two branches are designed for each task inside an implicit neural network and the hyper-parameters for these branches are tuned to accommodate the difficulty of the tasks. Also, notice that with this formulation, the network can be trained in a self-supervised way, without the need of corresponding RAW images.

Our contribution can be summarized as follows:

- We reformulate the RAW reconstruction problem as a RAW image approximation problem that learns the 2D-to-RAW mapping of image coordinates to RAW values conditioned on its sRGB image.
- We decompose the reconstruction into two aspects and design the implicit neural network accordingly.
- We conduct extensive experiments on different cameras and demonstrate our algorithm outperforms existing work significantly.

2. Related Work

Blind RAW reconstruction. Early works of RAW reconstruction were blindly taken sRGB images as input without extra metadata. Since the processing stages of modern ISPs are designed more complicated, more complex models are proposed [4, 18, 14, 5]. However, these methods suffer the inconvenient calibration procedures that need to be repeated on each camera or even each camera setting. Deep learning-based methods (e.g., [23, 36]), similarly, are faced with the parallel issues of camera-specific models, which required abundant training data captured for each camera. Recent methods that simulate ISP architectures by assuming a classical set of ISP operations, such as [37, 10], cannot handle different camera settings as they take fixed parameters (e.g., gamma correction) and ISP length as a priori knowledge.

RAW reconstruction with metadata. Compared with blind RAW reconstruction, a series of recent methods [25, 26, 28, 24] benefit from additional metadata embedded into sRGB images at capture time. Nguyen and Brown [25, 26] propose to extract and store the necessary parameters for recovering a RAW image from the sRGB counterpart. These parameters model the specific sRGB-to-RAW mapping and are restricted to a small memory (e.g., 128KB). However, their main algorithm deployed on the device has high computational cost and only considers the global tone mapping. Punnappurath and Brown [28], on the contrary, implement a uniform sampling on RAW images to save as metadata. They employ radial basis function (RBF) interpolation using sRGB pixel values and coordinates as input to reconstruct the RAW values. Nevertheless, the fixed type of RBF interpolation does not take full use of the information of sampled data. A further work [24] improves the effectiveness of sampling and recovering by taking a U-Net architecture [29] as both sampler and reconstruction networks, but their method requires to run a deep neural network on the device, which would lead to high computational cost and additional memory cost to save the pixel positions into metadata. We follow the uniform sampling in [28], but improve the reconstruction performance by INF.

Implicit neural representation. Implicit neural representation (INR) has recently been introduced to represent 2D images and 3D objects using coordinate-based multi-layer perceptron (MLP). To overcome the problem that conventional MLPs are incapable of representing high-frequency details of signals, two methods have been proposed. Sitzmann *et al*. [31] introduce SIREN, which replace ReLU activation with periodic activation (e.g., sine function). They demonstrate that the representation power of SIREN comes from the derivation invariance of sine function, and provide a number of potential applications and future works. A concurrent work [32] leverages random Fourier feature mapping on input coordinates to enable the MLP with ReLU to learn high-frequency details. It indicates that a Fourier feature mapping can beneficially address the spectral bias of a conventional MLP. Based on these breakthroughs, INR has been successfully adopted in various tasks [8, 1, 6, 35]. In this work, the proposed INF is based on [31], which shows

remarkably improvement in reconstruction accuracy.

Guided super-resolution. Another topic similar to RAW reconstruction with metadata is guided image super-resolution, which aims to convert a low-resolution (LR) source image to the high-resolution (HR) target with a guided HR image [20, 30, 33]. The difference of these two topics is that RAW reconstruction has obvious mapping of pixel values between two HR images. On this aspect, PixTransform [20] is more related to our method. It uses an MLP with ReLU to learn the pixel-to-pixel mapping of guided HR image to target HR image, where the LR image is treated like the metadata in RAW reconstruction. Our method differs in the MLP layer, where we use SIREN layer [31] to enhance the expressive power [38].

3. Method

3.1. Problem Formulation with Implicit Function

Let $\boldsymbol{S}$ and $\boldsymbol{R}$ denote an sRGB image and the corresponding RAW image, respectively. Since $\boldsymbol{R}$ is converted to $\boldsymbol{S}$ through a series of operations in the camera ISP, classical methods aim to build a model g to map $\boldsymbol{S}$ back to $\boldsymbol{R}$, i.e., $\boldsymbol{R} = g(\boldsymbol{S})$. However, the model g is specific to a camera ISP or even a set of ISP parameters. Metadata-based methods introduce extra data $\boldsymbol{M}$ for RAW reconstruction, i.e., $\boldsymbol{R} = f(\boldsymbol{S}, \boldsymbol{M})$. Here f is supposed to be a general function, and $\boldsymbol{M}$ includes the image-specific information generated at capture time, which is commonly implemented by sampling on the RAW images for low computation cost. Hence $\boldsymbol{M} = \{\mathbf{p}_i, \mathbf{r}_i\}_{i=0}^{N}$, where $\mathbf{p}_i = (x_i, y_i)$ is the coordinate of a sampled RAW pixel and $\mathbf{r}_i = (r_{R_i}, r_{G_i}, r_{B_i})$ its RAW value. N represents the number of sampled RAW pixels and i refers to the i^{th} sampling one.

In contrast to previous work that establishes a mapping from RGB to RGB values, $f : (r, s) \rightarrow r$, we reformulate the problem as a mapping from the 2D coordinates to the RAW values conditioning on the sRGB value

$$f : (p; s) \rightarrow r, \tag{1}$$

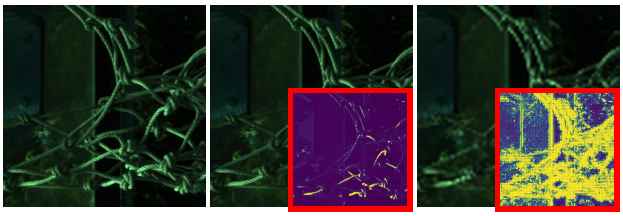
where r, s, p denote the variables for the RAW value, the sRGB value and coordinates. They can either represent the value for a pixel or a patch, or a set of coordinates.

Specifically, for the sRGB image, we only use the the sRGB value $(s_{R_i}, s_{G_i}, s_{B_i})$ at $\mathbf{p}_i$, queried by $S(\mathbf{p}_i)$. Therefore we aim to learn an implicit neural function f_θ taking a coordinate $\mathbf{p}_i$ and its sRGB value and output $\mathbf{r}_i$.

$$f_\theta : (\mathbf{p}_i; \mathbf{s}_i) \rightarrow \mathbf{r}_i. \tag{2}$$

To learn the function, we define the loss function below to find the best configuration for the parameters $\boldsymbol{\theta}$ of f

$$\mathcal{L} = \sum_{i=1}^{N} \| f_{\boldsymbol{\theta}}(\mathbf{s}_i, \mathbf{p}_i) - \mathbf{r}_i \|, \tag{3}$$



RAW (GT) (a) $f : \mathbf{s}_i \rightarrow \mathbf{r}_i$ (b) $f : \mathbf{p}_i \rightarrow \mathbf{r}_i$

Figure 2. Illustration of the two aspects of reconstruction. It can be observed that the reconstruction from (a) pixel value mapping is much more accurate than (b) spatial super-resolution. Our method utilizes the information from both $\mathbf{s}_i$ and $\mathbf{p}_i$, but give a different constraint to them to control their impacts.

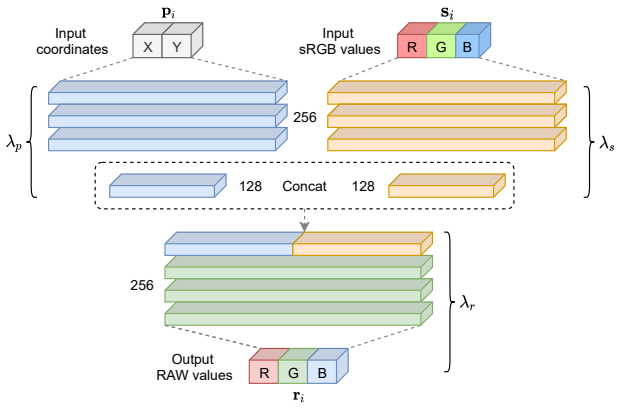


Figure 3. The network structure of our proposed implicit neural function (INF). The inputs are divided into sRGB values and coordinates, as their branches are regularized with different parameters of weight decay. Each cube represents the output of each linear layer with sine activation proposed in [31].

The overview of our method is shown in Figure 4. As mentioned above, considering the on-device computational cost, we follow SAM [28] to implement a uniform sampling on the RAW image to get the metadata $\boldsymbol{M}$. For the reconstruction stage, we implement the same sampling on the sRGB image, which is used to train the implicit neural function (INF) together with $\boldsymbol{M}$. The sRGB image is then processed by the trained INF to reconstruct the RAW image.

3.2. Two-way Raw Reconstruction

We note that Equation (2) contains the reconstruction from two aspects: (a) a pixel values mapping (i.e., $f : \mathbf{s}_i \rightarrow \mathbf{r}_i$) and (b) a spatial super-resolution (i.e., $f : \mathbf{p}_i \rightarrow \mathbf{r}_i$). Since $\boldsymbol{M}$ will cost additional memory, the sampling rate is required to be constrained (e.g., no more than 1.5%), which limits the accuracy of (b). On the contrary, the sRGB-to-RAW mapping is *piece-wise smooth* [27], hence only such small amount of samples is able to model the whole mapping function theoretically. We provide an example in Fig-

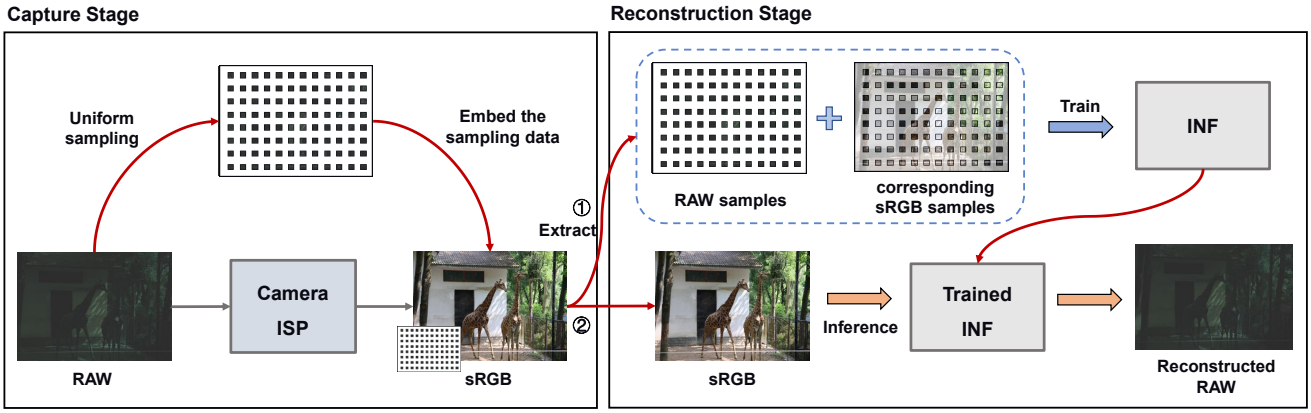


Figure 4. Overview of our proposed RAW reconstruction method. At capture stage, we follow [28] to implement an uniform sampling on the RAW image to generate the metadata, which does not require to store the pixel positions. The metadata is then embedded into the sRGB image for storage and transmission. At reconstruction stage, the metadata together with corresponding sRGB samples are used to train the implicit neural functions (INFs), which then convert the sRGB image to the RAW. Compared with the state-of-the-art metadata-based RAW reconstruction approach [24], our method achieves both lower on-device computational cost and much higher reconstruction accuracy.

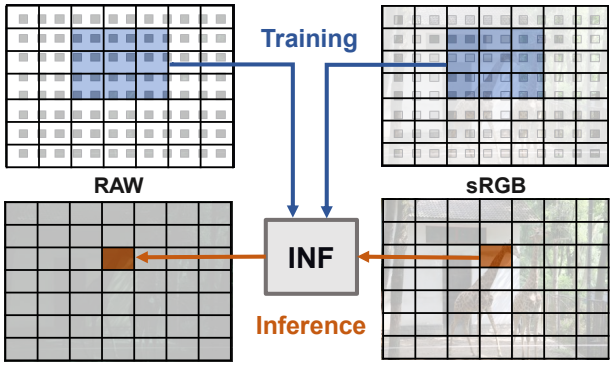


Figure 5. Illustration of the patch-specific INF. Inspired by [28], we split the image into patches, and train the INF for each patch with its neighbours to make it robust to the local mapping. We further discuss the impact of patch-size in the ablation study.

ure 2 to intuitively illustrate this phenomenon. Therefore (a) is supposed to be more significant than (b) in f.

Based on the analysis of the sub-tasks, we design an implicit neural network consisting of two branches as shown in Figure 3. INF is an MLP structure built with SIREN layers [31], where the inputs are separated into two branches and then concatenated into the output branch. Each branch contains 4 layers, with 256 channels for the hidden layer. The output channels of two input branches are reduced to half, which are then concatenated into the output branch. Here $\boldsymbol{\theta}$ refers to the network weights, and Equation (4) can then be solved by training the INF.

Also, considering the difficulty of the tasks, we introduce regularisation parameters to Equation (3) to control the complexity of different branches. That is,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left\| f_{\boldsymbol{\theta}}\left(\mathbf{s}_i, \mathbf{p}_i\right) - \mathbf{r}_i \right\| + \lambda_c \left\| \boldsymbol{\theta}_c \right\|^2, \quad (4)$$

where $c \in \{\mathbf{s}, \mathbf{p}, \mathbf{r}\}$ and λ_c is a hyper-parameter to control the strength of regularisation for different parts of $\boldsymbol{M}$.

We note that the structure of INF is most similar to the network of [20]. However, instead of merging the two input branches by adding their output, we first reduce the output size of these two branches and then concatenate them as the next input. This is because the output of SIREN Layers X_l is distributed in $X_l \sim \mathrm{Arcsin}\left(-1, 1\right)$ [31], and adding the outputs would let the output branch treat them equally. Hence we use concatenation to enable the network to learn different weights for the outputs of $\mathbf{p}_i$ and $\mathbf{s}_i$ branches.

Reconstruction conditioned on sRGB patches. Additionally, it is worth noting that local tone mapping is commonly applied as a favourable part of the ISP pipeline, which is adopted on local images to enhance the special objects (e.g., food, flower, sky). In order to make our method robust to local tone mapping, we split the input image into patches, and then train the INF for each patch. Moreover, inspired by [28], considering that the mapping of each patch and its neighbours are supposed to have strong relationship, we use the samples from each patch together with its eight neighbour patches to train the patch-specific INF. Therefore the mapping in Eq. (2) becomes

$$f_\theta : \left(\mathbf{p}_i; \{s_j\}_{j \in \mathcal{N}(i)}\right) \rightarrow \mathbf{r}_i. \quad (5)$$

where $\mathcal{N}(i)$ represents neighbour patches of current pixel $\mathbf{p}_i$. Figure 5 illustrates the training strategy of INF for such patch-wise RAW reconstruction.

4. Experiments

4.1. Experimental Setup

Baselines. We take three metadata-based RAW reconstruction methods for comparison: RIR [25], SAM [28] and CAM [24]. The RIR method computes the parameters of global sRGB-to-RAW mapping as metadata. The SAM and CAM methods adopt different sampling approaches on RAW images to generate metadata. We follow the uniform sampling as SAM [28], and show that it is enough to get much higher performance using uniform samples with the proposed method. Note that neither RIR nor SAM make their source codes publicly available, and only CAM reports their results reproduced on its own dataset. Therefore we only give a comparison with RIR and SAM on the dataset of CAM by inheriting its results in Table 1 and Figure 6.

Dataset. We deploy the same dataset of the previous work [24] to test our proposed method, which contains the preprocessed [13] and downsampled (with a downsampling factor of 4) images of three cameras from the NUS dataset [9]—Samsung NX2000, Olympus E-PL6 and Sony SLT-A57. Note that as our method does not require any pretraining step, the training and validation sets are not used in our method. Besides, since our method is able to directly process full-resolution images, we also compare it with CAM [24] on the original NUS dataset [9]. Additionally, as the NUS dataset [9] includes JPEG-compressed sRGB images rendered by real camera ISPs, we also conduct the experiments on these RAW-JPEG image pairs. These results are reported in the supplementary.

Implementation details. For capture stage, we use the same sampling rate of 1.5% like [24] in all experiments for fair comparison with other metadata-based methods. With respect to reconstruction stage, we split the image into different patches—$(228, 272)$ for Samsung NX2000, $(216, 288)$ for Olympus E-PL6 and $(204, 304)$ for Sony SLT-A57—to make the resolution divisible. The proposed INF is trained by Adam optimizer [15] with L_2 loss in 500 iterations. We set the initial learning rate to 0.0001, and reduce it to half every 200 iterations. The regularisation parameters are set to $\lambda_{\mathbf{s}} = 0.0001, \lambda_{\mathbf{p}} = 0.1, \lambda_{\mathbf{r}} = 0.001$ to produce the final results. It is worth noting that the reconstruction network is only trained for each specific patch, and will be re-initialized for the training of other patches.

4.2. Results

Quantitative results. We report the quantitative comparison results against other three baselines in Table 1. It can be observed that our method makes a remarkably improvement in both PSNR and SSIM metrics, especially outperforms the closest competitor CAM [24] by more than 10 dB in average PSNR. This indicates the effectiveness of INF to modeling the sRGB-to-RAW mapping.

Qualitative results. We also provide examples of the qualitative comparison in Figure 6, which gives an intuitive explain of the success of our method. As can be seen from the error maps, our method significantly reduces the reconstructed error on the whole images, especially the saturated regions [26]. We note that recovering saturated regions is the most difficult problem for all RAW reconstruction methods, and is the **only case** of our method getting large errors ($\geqslant 1\%$). We attribute this to the expressive power [38] of INF, which is further discussed in Section 4.4.

4.3. Ablation study.

RBF vs. INF. We first compare RBF and INF with different input, which is shown in Table 2. As can be seen, the proposed INF outperforms the RBF interpolation [28] whatever the input is. We attribute the main performance boost to the basic structure, i.e. SIREN [31], as INF is able to modelling the sRGB-to-RAW mapping more accurately (59.62dB) only with the pixel value as input. Moreover, compared with RBF, our proposed INF can make further improvement by splitting the input into separate branches and regularize them with different weights.

The structure of INF. We also conduct experiments on the structure of INF, which is reported in Figure 7. We compare the different activation functions and the layer location where two input branches are merged. It is obviously shown that the INF with sine activation remarkably outperforms the INF with ReLU function, which can be attributed to the expressive power of INRs [38]. As to the location of fusion layer, we find there is no significant difference except the final layer for the INF with sine activation, since it directly maps the merged vector to the output by a fully connection layer without activation. Therefore, we merge the two branches at fourth layer to balance the complexity of each branch[1]. On the contrary, due to lack of such expressive power, the INF with ReLU requires more layers to model the mapping. Hence the performance increases when the input branch goes deep.

Regularisation parameters. We further discuss the influence of key setups to INF, which mainly focuses on the regularisation parameters and patch size. As is shown in Table 3, it is obvious that the regularisation parameters play a vital role in the final accuracy. Here we use the grid search with a commonly-used tuner TPE [2] (implemented by [22]) in a grid of $\left[1, 10^{-1}, \ldots, 10^{-5}\right]$ to determine each parameter. In our experiments, we find the regulation parameters are mainly related to the **complexity** of camera ISP rather than data dependent, which means we only need

[1] In fact, we find the fusion at third layer would achieve a slightly higher performance, but which can be ignored compared with the instability of random initialization. See more detail in supplementary material.

Dataset	Method	Samsung NX2000		Olympus E-PL6		Sony SLT-A57	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CAM [24]	RIR [25]	45.66	0.9939	48.42	0.9924	51.26	0.9982
	SAM [28]	47.03	0.9962	49.35	0.9978	50.44	0.9982
	CAM [24]	49.57	0.9975	51.54	0.9980	53.11	0.9985
	Ours	**53.53**	**0.9985**	**56.71**	**0.9991**	**58.36**	**0.9992**
NUS [9]	CAM [24]	49.51	0.9937	52.87	0.9961	53.34	0.9959
	Ours	**61.48**	**0.9991**	**63.38**	**0.9993**	**63.80**	**0.9995**

Table 1. Quantitative comparison with [25, 28, 24]. Here CAM dataset [24] involves downsampling operation (with the factor of 4) on the original NUS dataset [9], which undermines the strength of local mapping. For CAM dataset, we conduct the experiments on their test set to inherit the results of [24]. For NUS dataset, we report the results tested on the whole dataset. The best score for each column is in bold.

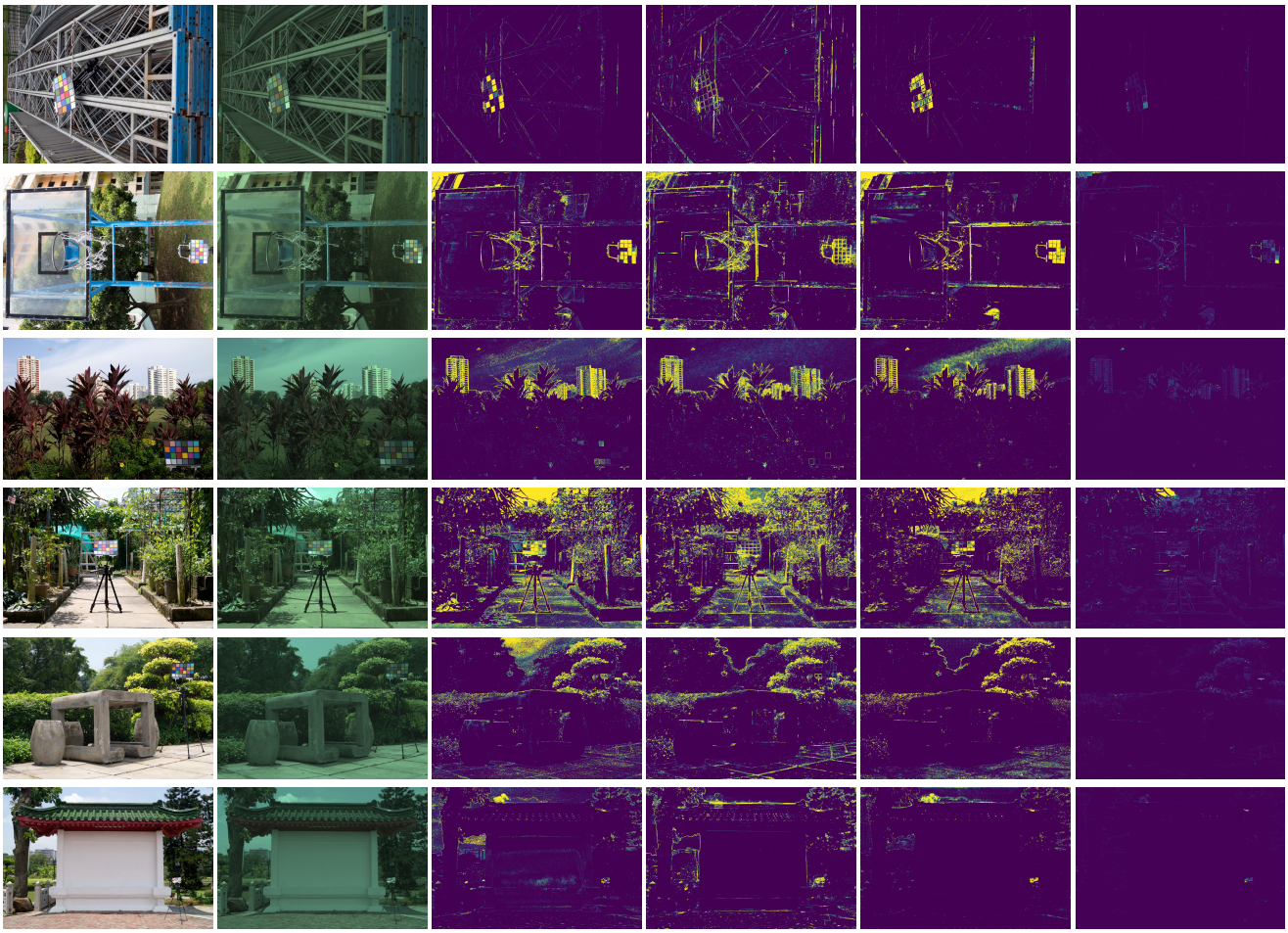


Figure 6. Qualitative comparison with [25, 28, 24]. Each two rows from top to bottom are respectively selected from Olympus E-PL6, Samsung NX2000 and Sony SLT-A57. The GT RAW images are processed with a gamma function for better visualization, and the reconstructed RAW images are visualized through error maps. This figure is best viewed in the electronic version.

to tune it for different ISPs, e.g., digital camera ISP, smart phone ISP, or software ISP. It also means we just need to use one or quite a few images for tuning, resulting in a fast tuning process—within 2 hours on a single RTX3090. The

Method	Input	PSNR
RBF [28]	$\mathbf{s}$	29.10
RBF [28]	$\mathbf{p}$	21.71
RBF [28]	$(\mathbf{s}, \mathbf{p})$	49.69
INF	$\mathbf{s}$	59.62
INF	$\mathbf{p}$	31.09
INF	$(\mathbf{s}, \mathbf{p})$	*60.91*
INF	$\{\mathbf{s}; \mathbf{p}\}$	**62.24**

Table 2. Ablation study on different reconstruction methods. The best score is in bold and the second in italic. $(\mathbf{s}, \mathbf{p})$ refers to combine $\mathbf{s}$ and $\mathbf{p}$ into a single input vector, while $\{\mathbf{s}; \mathbf{p}\}$ denotes separating them as different input. For the RBF interpolation, we use the linear RBF kernel as proposed in [28].

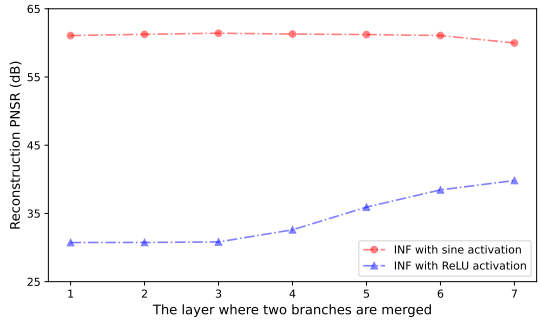


Figure 7. Ablation study on the structure of INF. The abscissa value represents the layer where we merge the two input branches, e.g., the branches are merged at fourth layer in Figure 3.

tuned parameters can then be used for all scenes and all other images from the ISPs with similar complexity, hence we use the same parameters for the three reported cameras.

As shown in Table 3, the proposed INF achieves the best performance at $\lambda_{\mathbf{s}} = 0.0001, \lambda_{\mathbf{p}} = 0.1, \lambda_{\mathbf{r}} = 0.001$, which indicates the spatial branch is not supposed to be more complex. This is because the output $\mathbf{r}$ has weaker relationship with the coordinates $\mathbf{p}$ than sRGB values $\mathbf{s}$, and regularisation could help to control the contribution of $\mathbf{s}$ and $\mathbf{p}$. Therefore when we set $\lambda_{\mathbf{p}} = 0$, the reconstruction performance dramatically drops down due to inaccurate information from $\mathbf{p}$ is expanded. We believe such intuitive relationship would be helpful to set the initial values for regularisation parameters. Besides, for different ISPs with similar complexity, the results are insensitive to the parameters in a wide range (for example, where all branches are regularized by $\lambda \geqslant 10^{-4}$). In this sense, the risk of over-fitting due to these hyper-parameter settings is very small.

Impact of patch size. Moreover, we find that different size of the patch would lead to an obvious accuracy gap. For the

Patch size	$(\lambda_{\mathbf{s}}, \lambda_{\mathbf{p}}, \lambda_{\mathbf{r}})$	PSNR
$\times 1$	$(0, 0, 0)$	47.48
$\times 2$		49.17
$\times 4$		50.53
$\times 1$	$(0.0001, 0.001, 0.0001)$	61.13
$\times 2$		60.38
$\times 4$		57.80
$\times 1$	$(0.0001, 0.1, 0.001)$	**62.58**
$\times 2$		*62.24*
$\times 4$		59.64

Table 3. Ablation study on the patch size and regularisation parameters. The best score is in bold and the second in italic. For the patch size, $\times 2$ refers to the sizes we report in our main experiments, and the rescale factor of $1, 2, 4$ corresponds to the length of patch sides. We note that the accuracy gap less than 0.5dB can not represent an improvement, which may be caused by the different calculations of PSNR and the random initialization of INF. This is further discussed in our supplementary material.

case of all regularisation parameters set to 0, we find small patch would lead to accuracy reduction. When the regularisation is correctly set, reconstruction on a small patch is more accurate than on a large one. This can be explained by the various local mapping functions in different image regions, which means modelling the local mapping for a small patch would be more accurate than a larger one. Note that though reducing the patch size would be helpful to the reconstruction accuracy, it would take much more time to recover an image. Therefore, we compromise to use the patch size around 200-300 in the main experiments.

4.4. Discussion

Why INF works in sRGB-to-RAW mapping. We first rethink the RBF interpolation [3] used in SAM [28], by which the sRGB-to-RAW mapping is built in the following form:

$$\hat{y} = \mathbf{M}^{test}\mathbf{M}^{-1}y \tag{6}$$

where $\mathbf{M}$ denotes the RBF interpolation matrix with linear kernel function [27], and $test$ refers to the pixels that need to be reconstructed. As discussed in [32], the output of an MLP can be approximated as:

$$\hat{y}^{(t)} = \mathbf{K}^{test}\mathbf{K}^{-1}\left(\mathbf{I} - e^{-\eta\mathbf{K}t}\right)y \tag{7}$$

where $\mathbf{K}$ denotes the NTK matrix [12] and t refers to training iterations. In our experiments, an MLP (typical with ReLU activation) tends to behave like RBF when t is large enough (e.g., $t > 1000$), which meets the results of Equation (6) and Equation (7). The reason why RBF interpolation works can be attributed to that sRGB-RAW mapping

Method	×4	×8	×16
Bicubic	2.62	4.19	6.68
PixTransform [20]	3.51	3.94	4.94
INF	2.50	3.06	5.23
INF with ReLU	**2.39**	**2.69**	**3.57**

Table 4. Quantitative comparison on guided depth super-resolution. We use the Middlebury dataset provided by [33] for testing. The results are reported in terms of average RMSE.

is piecewise smooth [27], but there still exits errors when using a linear kernel function, which is also suffered by MLPs with ReLU. On the other hand, Sitzmann *et al*. [31] demonstrate that MLPs with periodic activation functions (i.e., SIRENs) are more suitable for representing complicated signals and their derivatives compared with the typical MLP architectures. SIRENs have the advantage of representing the derivatives of the signal [31], which is more helpful to reconstruct fine details. Therefore, recovering RAW images by SIRENs can effectively improve the performance, which is also proved by Table 2.

Limitations and future research. One of the unresolved problem of our method is that the regularisation parameters of weight decay require pre-experiments to determine. These three parameters play a significant role to control the final result by limiting the complexity of network. In our experiments, we find the parameters are strongly related to the correlation of input and output data. For example, compared with the coordinates, sRGB values have a more obvious relationship with the RAW values, where the regularisation parameters are respectively set to $\lambda_p = 0.1$ and $\lambda_s = 0.0001$. Such correlation is suitable for modelling with another neural network, and would lead to the performance improvement with optimized parameters. Therefore, learning the regularisation parameters from the input and output is supposed to be a valuable topic of future work.

4.5. Applications

Guided depth super-resolution As mentioned in Section 2, our framework is related to guided super-resolution, therefore we also provide an experiment to illustrate the applicability of our framework. Considering that our method is only self-supervised without any priori knowledge, we give a comparison with another self-supervised method PixTransform [20]. As shown in Table 4, INF with ReLU outperforms other methods, which indicates the ReLU activation is more suitable for this task than the sine activation. The reason for ReLU working better can be attributed to the sharp changes of the depth near the object edges, as ReLU is piecewise and discontinuous, which fits this depth discontinuity. In contrast, the discontinuity in RAW-to-sRGB mapping is moderate and the smooth sine activation function is a better option. We also provide qualitative comparison in Figure 8. which shows that the result of INF, though does not achieve higher accuracy, obtains better visual quality compared with the result of INF with ReLU.

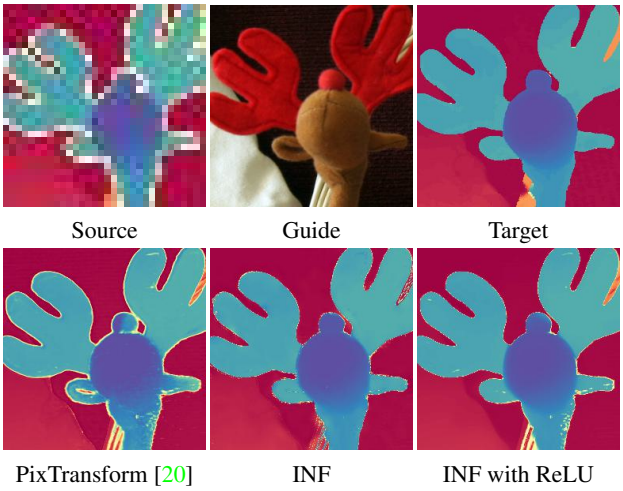


Figure 8. Qualitative comparison on guided depth super-resolution.

Low-light image enhancement. As a potential application of RAW reconstruction, we test our method on low-light image enhancement (LLIE) task. We find that the LLIE task can be simplified into a linear degradation problem (e.g., enlarging the pixel values by the same multiple) on the reconstructed RAW images, which achieves equal visual quality compared with the state-of-the-art deep learning-based approaches [11, 34, 21]. The results are reported in our supplementary. Note that such comparison is merely aimed to illustrate the effectiveness of executing low-light enhancement on reconstructed RAW images.

5. Conclusion

We propose a method for recovering the RAW image from the sRGB counterpart with assistance of additional metadata, which is sampled from the RAW image at capture time. We introduce the implicit neural function (INF) to remarkably improve the reconstruction accuracy (10 dB average PSNR) only with uniform sampling. We prove that the structure of INF is beneficial to merging the information from both pixel values and coordinates. Further experiments indicate that our framework is also suitable for the task of guided super-resolution.

Acknowledgement. This work was partly supported by National Key R&D Program of China (Grant No. 2021YFD1400200) and National Natural Science Foundation of China (Grant No. 62088101).

References

[1] Yizhak Ben-Shabat, Chamin Hewa Koneputugodage, and Stephen Gould. Digs: Divergence guided shape implicit neural representation for unoriented point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19323–19332, 2022. 2

[2] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011. 5

[3] Marco Evangelos Biancolini. *Fast radial basis functions for engineering applications*. Springer, 2017. 2, 7

[4] Ayan Chakrabarti, Daniel Scharstein, and Todd E Zickler. An empirical camera model for internet color vision. In *BMVC*, volume 1, page 4. Citeseer, 2009. 1, 2

[5] Ayan Chakrabarti, Ying Xiong, Baochen Sun, Trevor Darrell, Daniel Scharstein, Todd Zickler, and Kate Saenko. Modeling radiometric uncertainty for vision with tone-mapped color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2185–2198, 2014. 1, 2

[6] Bowei Chen, Tiancheng Zhi, Martial Hebert, and Srinivasa G Narasimhan. Learning continuous implicit representation for near-periodic patterns. *arXiv preprint arXiv:2208.12278*, 2022. 2

[7] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3291–3300, 2018. 1

[8] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 2

[9] Dongliang Cheng, Dilip K Prasad, and Michael S Brown. Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. *JOSA A*, 31(5):1049–1058, 2014. 5, 6

[10] Marcos V Conde, Steven McDonagh, Matteo Maggioni, Ales Leonardis, and Eduardo Pérez-Pellitero. Model-based image signal processors via learnable dictionaries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 481–489, 2022. 1, 2

[11] Chunle Guo Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1780–1789, June 2020. 8

[12] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018. 7

[13] Hakki Can Karaimer and Michael S Brown. A software platform for manipulating the camera imaging pipeline. In *European Conference on Computer Vision*, pages 429–444. Springer, 2016. 5

[14] Seon Joo Kim, Hai Ting Lin, Zheng Lu, Sabine Süsstrunk, Stephen Lin, and Michael S Brown. A new in-camera imaging model for color computer vision and its application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2289–2302, 2012. 1, 2

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015. 5

[16] Shayan Kousha, Ali Maleky, Michael S Brown, and Marcus A Brubaker. Modeling srgb camera noise with normalizing flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17463–17471, 2022. 1

[17] Zhihao Li, Si Yi, and Zhan Ma. Rendering nighttime image via cascaded color and brightness compensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 897–905, 2022. 1

[18] Hai Ting Lin, Zheng Lu, Seon Joo Kim, and Michael S Brown. Nonuniform lattice regression for modeling the camera imaging pipeline. In *European Conference on Computer Vision*, pages 556–568. Springer, 2012. 1, 2

[19] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1651–1660, 2020. 1

[20] Riccardo de Lutio, Stefano D'aronco, Jan Dirk Wegner, and Konrad Schindler. Guided super-resolution as pixel-to-pixel transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8829–8837, 2019. 3, 4, 8

[21] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5637–5646, 2022. 8

[22] Microsoft. Neural Network Intelligence, 2021. 5

[23] Seonghyeon Nam and Seon Joo Kim. Modelling the scene dependent imaging in cameras with a deep neural network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1717–1725, 2017. 1, 2

[24] Seonghyeon Nam, Abhijith Punnappurath, Marcus A Brubaker, and Michael S Brown. Learning srgb-to-raw-rgb de-rendering with content-aware metadata. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17704–17713, 2022. 1, 2, 4, 5, 6

[25] Rang MH Nguyen and Michael S Brown. Raw image reconstruction using a self-contained srgb-jpeg image with only 64 kb overhead. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1655–1663, 2016. 1, 2, 5, 6

[26] Rang MH Nguyen and Michael S Brown. Raw image reconstruction using a self-contained srgb–jpeg image with small memory overhead. *International journal of computer vision*, 126(6):637–650, 2018. 1, 2, 5

[27] Abhijith Punnappurath and Michael S Brown. Learning raw image reconstruction-aware deep image compressors. *IEEE*

transactions on pattern analysis and machine intelligence, 42(4):1013–1019, 2019. 3, 7, 8

[28] Abhijith Punnappurath and Michael S Brown. Spatially aware metadata for raw reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 218–226, 2021. 1, 2, 3, 4, 5, 6, 7

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[30] Guy Shacht, Dov Danon, Sharon Fogel, and Daniel Cohen-Or. Single pair cross-modality super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6378–6387, 2021. 3

[31] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 2, 3, 4, 5, 8

[32] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 2, 7

[33] Jiaxiang Tang, Xiaokang Chen, and Gang Zeng. Joint implicit image function for guided depth super-resolution. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4390–4399, 2021. 3, 8

[34] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2604–2612, 2022. 8

[35] Keyu Wu, Yifan Ye, Lingchen Yang, Hongbo Fu, Kun Zhou, and Youyi Zheng. Neuralhdhair: Automatic high-fidelity hair modeling from a single image using implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1526–1535, 2022. 2

[36] Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6287–6296, 2021. 1, 2

[37] Ke Yu, Zexian Li, Yue Peng, Chen Change Loy, and Jinwei Gu. Reconfigisp: Reconfigurable camera image processing pipeline. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4248–4257, 2021. 2

[38] Gizem Yüce, Guillermo Ortiz-Jiménez, Beril Besbinar, and Pascal Frossard. A structured dictionary perspective on implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19228–19238, 2022. 3, 5

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