

Learning RAW-to-sRGB Mappings with Inaccurately Aligned Supervision

Zhilu Zhang¹, Haolin Wang¹, Ming Liu¹, Ruohao Wang¹, Jiawei Zhang², Wangmeng Zuo^{1,3} (✉)

¹Harbin Institute of Technology, ²SenseTime Research, ³Pazhou Lab, Guangzhou

{cszlzhang, Why.cs, csmliu, rhwangHIT}@outlook.com {zhjw1988}@gmail.com wmuo@hit.edu.cn

Abstract

Learning RAW-to-sRGB mapping has drawn increasing attention in recent years, wherein an input raw image is trained to imitate the target sRGB image captured by another camera. However, the severe color inconsistency makes it very challenging to generate well-aligned training pairs of input raw and target sRGB images. While learning with inaccurately aligned supervision is prone to causing pixel shift and producing blurry results. In this paper, we circumvent such issue by presenting a joint learning model for image alignment and RAW-to-sRGB mapping. To diminish the effect of color inconsistency in image alignment, we introduce to use a global color mapping (GCM) module to generate an initial sRGB image given the input raw image, which can keep the spatial location of the pixels unchanged, and the target sRGB image is utilized to guide GCM for converting the color towards it. Then a pre-trained optical flow estimation network (e.g., PWC-Net) is deployed to warp the target sRGB image to align with the GCM output. To alleviate the effect of inaccurately aligned supervision, the warped target sRGB image is leveraged to learn RAW-to-sRGB mapping. When training is done, the GCM module and optical flow network can be detached, thereby bringing no extra computation cost for inference. Experiments show that our method performs favorably against state-of-the-arts on ZRR and SR-Raw datasets. With our joint learning model, a light-weight backbone can achieve better quantitative and qualitative performance on ZRR dataset. Codes are available at <https://github.com/cszhilu1998/Raw-to-sRGB>.

1. Introduction

The image signal processing (ISP) pipeline refers to the processing of raw sensor image for producing high quality display-referred sRGB image, and thus is pivotal for a camera system. A representative ISP pipeline usually involves a sequence of steps including demosaicking, white balance, color correction, tone mapping, denoising, sharpening, gamma correction and so on [40]. While hand-

crafted ISP solutions are usually adopted in current camera systems, convolutional networks (CNNs) have exhibited great potential in learning deep ISP model in an end-to-end manner [22, 29, 45].

The end-to-end property of deep ISP makes it very competitive to learn RAW-to-sRGB mapping to generate high quality image for mobile camera [22]. Albeit mobile camera has become the dominant sources of photos, it has a smaller sensor size and limited aperture in comparison to DSLR camera. By learning RAW-to-sRGB mapping to produce DSLR-like sRGB image from mobile raw image, deep ISP model can thus offer an encouraging way to close the gap between mobile camera and DSLR camera. Moreover, in contrast to 8-bit sRGB image, raw image usually has higher-bit (e.g., 10-14 bit) and may convey richer details. Therefore, learning RAW-to-sRGB mapping is beneficial to performance improvement even for other low level vision tasks, e.g., image super-resolution [62], low light image denoising [8] and high dynamic range imaging (HDR) [6].

However, when preparing training data, input raw image and target sRGB image are usually taken using different cameras (e.g., a smartphone and a DSLR) or with different camera configurations (e.g., focal length). Consequently, color inconsistency and spatial misalignment are usually inevitable. On the one hand, the color inconsistency makes it very challenging to generate well-aligned training pairs of input raw and target sRGB images. The input raw and target sRGB images usually cannot be perfectly aligned by existing methods [34, 49], resulting in mild alignment. On the other hand, learning with inaccurately aligned supervision is prone to pixel shift and producing blurry results (see Fig. 1(b)). To alleviate the adverse effect of inaccurate alignment, AWNet [9] adopted global context block [5] at the cost of increasing inference time, while Zhang *et al.* [62] presented a contextual bilateral (CoBi) loss to search the best matching patch for supervision. However, the patch-based alignment is unable to appropriately handle the spatially variant misalignment caused by depth discrepancy between objects. As a result, their method is still prone to producing blurry results as shown in Fig. 1(f).

In order to circumvent inaccurately aligned supervision

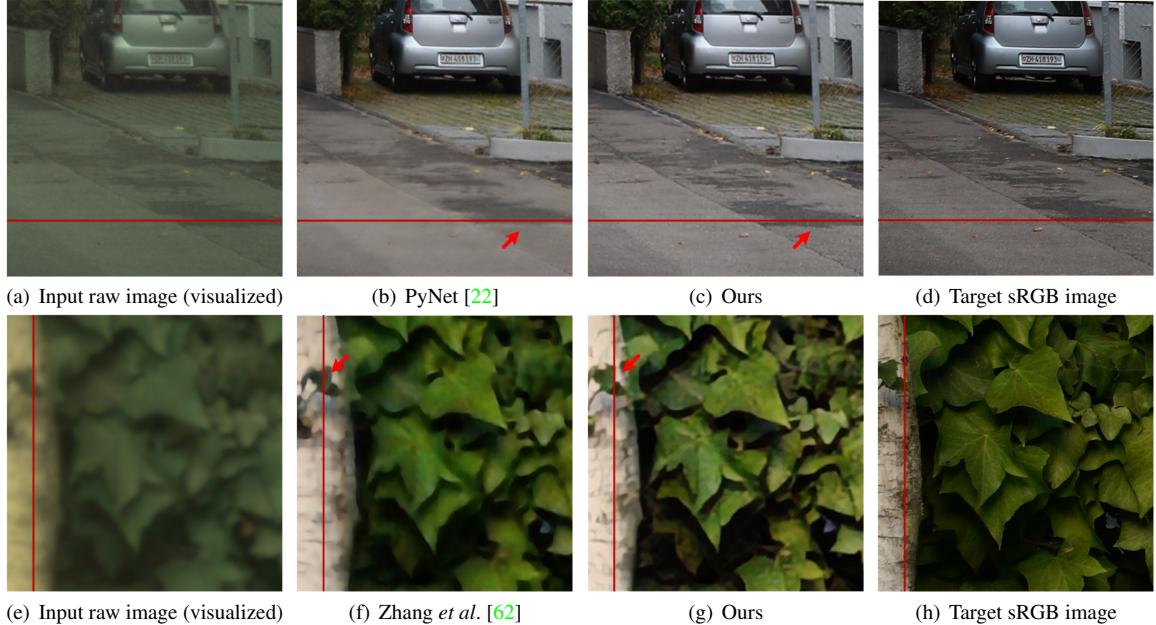


Figure 1: Example of data pairs of ZRR and SR-RAW datasets, where clear spatial misalignment can be observed with the reference line. With such inaccurately aligned training data, PyNet [22] and Zhang *et al.* [62] are prone to generating blurry results with spatial misalignment, while our results are well aligned with the input. Please zoom in for better observation.

problem, this paper presents a joint learning model for image alignment and RAW-to-sRGB mapping. We argue that one major reason that explains the inaccurate/mild alignment is the severe color inconsistency between input raw and target sRGB images. Otherwise, existing optical flow networks [10, 23, 46] can be readily utilized to fulfill the task of image alignment. Thus, we suggest to perform image alignment by concatenating a delicately designed global color mapping (GCM) module with a pre-trained optical flow estimation network (*e.g.*, PWC-Net [46]). In particular, the GCM module involves a stack of 1×1 convolutional layers to ensure that the mapping is spatially independent. To overcome the obstacle of color inconsistency, we constrain the GCM output to approximate the aligned target sRGB image. It is worth noting that GCM is deployed to align target sRGB image only during training. Thus, we can also take the target sRGB image and coordinate map to generate conditional guidance for modulating GCM features towards diminishing color inconsistency. Then, a pre-trained optical flow estimation network (*e.g.*, PWC-Net [46]) can be used to align the target sRGB image with the GCM output, resulting in the well aligned sRGB image.

The aligned target sRGB image can serve as a better supervision for training the RAW-to-sRGB mapping. In particular, we propose a LiteISPNet by reducing the residual channel attention blocks (RCABs) in MW-ISPNet [20]. GCM and LiteISPNet are jointly trained for both the alignment of target sRGB image (*i.e.*, GCM and PWC-Net) and the RAW-to-sRGB mapping (*i.e.*, LiteISPNet). When training is done, GCM and PWC-Net can be detached and only LiteISPNet is required for handling test raw images, thereby

bringing no extra inference cost. Experiments on Zurich RAW to RGB (ZRR) dataset [22] show that our solution is effective in learning with inaccurately aligned supervision and producing more fine details. Our proposed method also outperforms the state-of-the-art method in terms of quantitative metrics, perceptual quality and computational efficiency. Furthermore, using SRResNet as the backbone, experiments also show the effectiveness of our method for image super-resolution on the SR-RAW dataset [62].

The main contributions of this work are three-fold:

- An effective approach is presented to circumvent the task of learning RAW-to-sRGB mapping with inaccurately aligned supervision.
- A global color mapping (GCM) module is delicately designed to tackle the effect of color inconsistency on image alignment. A spatially preserving network (SPN) is leveraged to avoid spatial shift of pixels, and the target sRGB image is adopted to modulate GCM features towards diminishing color inconsistency.
- Quantitative and qualitative results on ZRR and SR-RAW datasets show that our method outperforms the state-of-the-art methods with no extra inference cost.

2. Related Work

2.1. Deep Networks for ISP

The camera ISP pipeline is deployed to produce display-referred sRGB image from raw images. To this end, classical ISP has been extensively studied, which generally involves a sequence of subtasks [40] including demosaicking, white balance, color correction, tone mapping, denois-

ing, sharpening, gamma correction and *etc.* For each sub-task, a number of methods have been proposed in the literature [2, 17, 42, 48]. Motivated by the unprecedented success of deep learning, CNNs have also been investigated to tackle several hard ISP subtasks like image denoising [57, 59, 60], demosaicking [12, 30, 47], auto-white-balance (AWB) [1, 18, 33, 54] and tone mapping [3, 13, 16, 58].

Recently, several attempts have been made to learn a full ISP pipeline with a deep model. Schwartz *et al.* [45] designed a DeepISP model to produce a well-lit sRGB image given a raw image captured under low-light environment. CameraNet [29] explicitly grouped the sub-tasks into two categories (*i.e.*, restoration and enhancement), and extracted the ground-truths by commercial software. Ignatov *et al.* [22] collected a dataset containing paired raw and sRGB images, which respectively are captured by Huawei P20 smartphone and Canon 5D Mark IV DSLR. With the dataset proposed in [22], two challenges were held [20, 21]. Among the decent methods proposed by the participants, MW-ISPNet [20] leveraged MWCNN [32] and residual channel attention blocks (RCABs) [63], AWNet [9] adopted global context block [5] to learn non-local color mapping, and they won first two places in the perceptual track [20]. In this work, we present a LiteISPNet by reducing the number of RCABs in MW-ISPNet [20] for learning full ISP model. By incorporating LiteISPNet with the joint learning model, better quantitative results and perceptual quality can be attained in comparison to MW-ISPNet [20] and AWNet [9].

2.2. RAW-to-sRGB Mappings for More Tasks

In [22], the raw sensor and sRGB images are captured using different cameras. This makes it feasible to empower low-quality raw sensor to produce high-quality sRGB image by imitating either other cameras or camera with different configurations. Moreover, the pixels of raw images are usually of higher-bit (*e.g.*, 10 ~ 14 bit), spatially independent and linear to brightness, thereby conveying richer details for benefiting image enhancement. Chen *et al.* [8] pioneered this line of work by leveraging paired low-light raw and long-exposure sRGB images with different ISO settings for extreme low-light image enhancement. And they further explored extreme low-light video enhancement in [7]. Analogously, raw images are also utilized in other low level vision tasks such as HDR [6] and image super-resolution (SR) [55, 62]. In this work, our method is also tested on image SR [62] by using SRResNet [27] as backbone, and achieves better quantitative and qualitative results.

Furthermore, studies have also been given for repurposing or merging RAW-to-sRGB mapping with high-level vision tasks. Wu *et al.* [53] designed a visionISP model to generate better input for object detection. Schwartz *et al.* [44] learned a model for image classification with raw images via distilling the knowledge of an ISP pipeline and an sRGB image classification model.

2.3. Alignment of Paired Raw and sRGB Images

For learning RAW-to-sRGB mappings, input raw and target sRGB images are usually taken using different cameras or with different camera configurations [22, 62]. Misalignment caused by multiple cameras and motion in the scene are thus inevitable, hindering the learning of RAW-to-sRGB mappings and giving rise to blurry or even pixel-shifted results. For suppressing the effect of motion in the scene, dual or multiple cameras are deployed to shoot concurrently for some datasets like KITTI [11, 38] and MultiPIE [15]. Beam splitter is also introduced to collect image pairs at the “same position” with different settings for real-world super-resolution [25] and deblurring [41]. Albeit with such equipments, misalignment remains unavoidable.

Several methods have been presented to align images from different sources. SIFT keypoints [34] are adopted for image registration in [19, 22, 52], where the homography can be estimated via RANSAC algorithm [49]. Cai *et al.* [4] designed a pixel-wise registration method that considers luminance adjustment for a real-world super-resolution dataset. Li *et al.* [28] warped the guidance image for facial image restoration via optical flow based methods [10, 23, 46]. However, input raw and target sRGB images usually have severe color inconsistency and cannot be perfectly aligned by existing methods, thereby resulting in mild alignment. Global context block [5, 9] and CoBi loss [62] have been introduced to alleviate the effect of mild alignment, but are still prone to producing blurry results. In this work, we introduce a global color mapping (GCM) module for tackling color inconsistency, and present a joint learning model for both image alignment and RAW-to-sRGB mapping.

3. Proposed Method

In this section, we first give an overview of our model for joint learning of image alignment and RAW-to-sRGB mapping. Then, the design of the global color mapping (GCM) module and LiteISPNet is detailed, and the learning objectives are presented. And we also extend our proposed method to other image enhancement tasks, *e.g.*, image SR.

3.1. Joint Learning Model

Denote by \mathbf{x} and \mathbf{y} a raw image and the corresponding target sRGB image. The RAW-to-sRGB mapping is used to produce a sRGB image $\hat{\mathbf{y}}$ from \mathbf{x} for approximating the color characteristic of target sRGB image \mathbf{y} ,

$$\hat{\mathbf{y}} = \mathcal{I}(\mathbf{x}; \Theta_{\mathcal{I}}), \quad (1)$$

where \mathcal{I} denotes a RAW-to-sRGB mapping (*e.g.*, LiteISPNet in Sec. 3.3) with the parameter $\Theta_{\mathcal{I}}$.

However, \mathbf{x} and \mathbf{y} are usually taken using different cameras or with diverse camera configurations, giving rise to inevitable spatial misalignment between them. Moreover, the severe color inconsistency between \mathbf{x} and \mathbf{y} further makes the image alignment more difficult. On the other hand, the

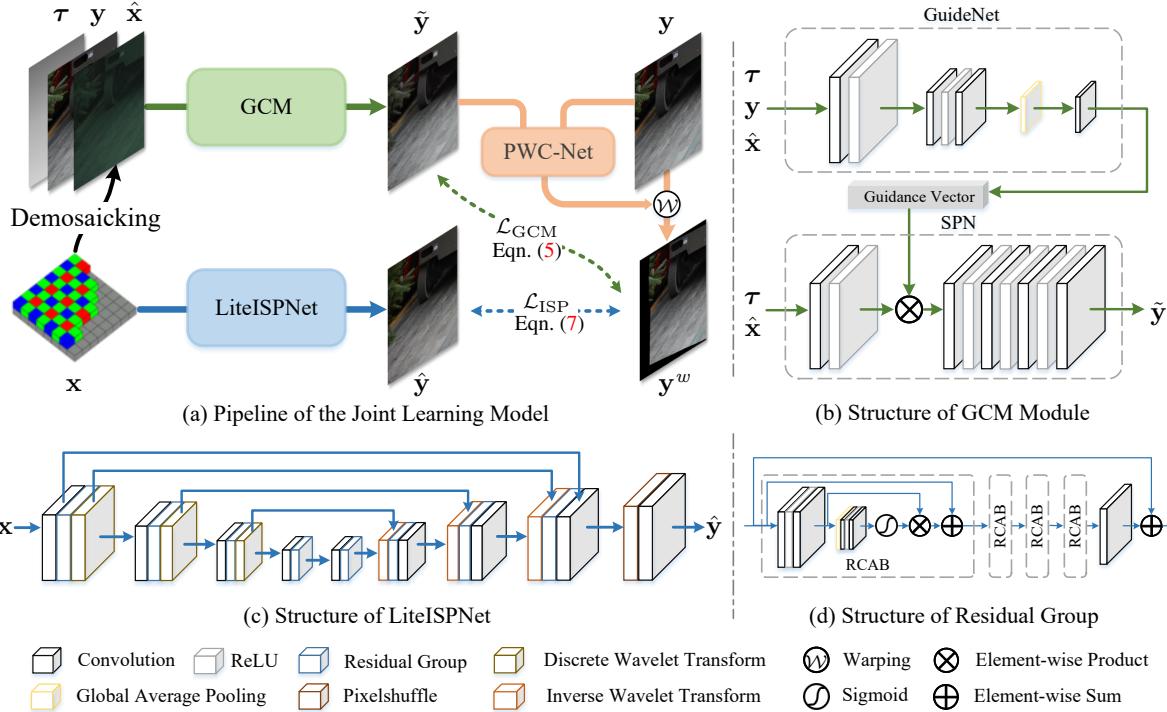


Figure 2: Illustration of the proposed joint learning framework. (a) Pipeline of the joint learning model, where GCM module converts the color of \hat{x} (demosaicked from x via simple interpolation) for more accurate optical flow, then the warped sRGB (y^w) can provide aligned supervision for a joint training of GCM and LiteISPNet. (b) Structure of GCM module, which is composed of a GuideNet and an SPN. The target sRGB (y) and a coordinate map (τ) are deployed for guiding the color conversion. (c) Structure of LiteISPNet, please refer to Sec. 3.3 for more details. (d) Structure of residual group used in (c).

RAW-to-sRGB mapping aims at imitating the color characteristics and fine details of target sRGB image. The misalignment between x and y is harmful to the learning of RAW-to-sRGB mapping, thereby being prone to producing blurry outputs with unfavorable pixel shift. Several approaches have been proposed to improve the alignment or the learning robustness, but are still not sufficient in suppressing blurry outputs.

Joint learning of image alignment and RAW-to-sRGB mapping can offer some new chances to circumvent such issue. On the one hand, RAW-to-sRGB mapping is helpful in diminishing the color inconsistency between x and y , thus easing the difficulty of image alignment. On the other hand, better image alignment also benefits RAW-to-sRGB mapping for suppressing blurry outputs and pixel shift. Unfortunately, the RAW-to-sRGB mapping (e.g., [22]) cannot completely avoid pixel shift (see Fig. 1), thus aligning y with \hat{y} cannot solve the misalignment problem.

Instead of aligning y with \hat{y} , our joint learning model leverages a delicately designed global color mapping (GCM) module to generate a color-adjusted image \tilde{y} for warping y . A simple demosaicking method (e.g., bicubic) is first used to obtain \hat{x} by filling in missing values of x . Then, the GCM module is introduced as a pixel-wise mapping of \hat{x} , and thus can guarantee not to introduce any pixel shift in color correction. Furthermore, GCM is only required dur-

ing training. So we can take target sRGB image and coordinate map as input to generate conditional guidance for modulating GCM features towards diminishing color inconsistency. Thus, the GCM module can be given by,

$$\tilde{y} = \mathcal{C}(\hat{x}, y, \tau; \Theta_C), \quad (2)$$

where \mathcal{C} denotes the GCM module, $\tau \in \mathbb{R}^{2 \times H \times W}$ is the 2D coordinate map containing the coordinate of the pixels, which is normalized to $[-1, 1]$. Given \tilde{y} and y , we use a pre-trained optical flow network (denote by \mathcal{F}), e.g., PWC-Net [46], to estimate the optical flow Ψ ,

$$\Psi = \mathcal{F}(\tilde{y}, y). \quad (3)$$

The estimated optical flow is then used to warp y to form a warped target sRGB image,

$$y^w = \mathcal{W}(y, \Psi), \quad (4)$$

where \mathcal{W} is a warping operation (e.g., bilinear interpolation) [46]. Then, y^w can serve as a well aligned target sRGB image for supervising the RAW-to-sRGB mapping \mathcal{I} in Eqn. (1), resulting in our joint learning model (see Fig. 2).

3.2. GCM Module

For optical flow estimation, the color-adjusted image \tilde{y} is required to satisfy two prerequisites. (i) \tilde{y} should imitate the color of y for diminishing the severe color inconsistency. (ii) The spatial position of the pixels should keep the same

as the input image $\hat{\mathbf{x}}$. According to [16], some commonly-used image processing operations can be approximated or formulated by multi-layer perceptron (MLP), and the pixel-wise nature ensures that the input and output are strictly aligned. Thus, we deploy a spatially preserving network (SPN) as the backbone of our GCM module, which is composed of a stack of 1×1 convolutional layers.

It is worth noting that GCM is only required during training. Thus the target sRGB image can also be safely fed into GCM for better converting $\tilde{\mathbf{y}}$ towards target sRGB image. To this end, we build a GuideNet to generate a conditional guidance vector from the raw ($\hat{\mathbf{x}}$) and target sRGB (\mathbf{y}) pair (see Fig. 2(b)). To alleviate the effect of misalignment in generating guidance vector, we use a relatively large kernel (*i.e.*, 7×7) in GuideNet, and global average pooling is deployed to acquire holistic information.

Besides, dark corner (*a.k.a.*, vignetting) can be observed in the raw images of several datasets (*e.g.*, ZRR [20]), where the brightness is gradually weakened from image center to the borders. Standard global color mapping, however, is not sufficient for compensating the adverse effect of dark corner. Fortunately, the effect of dark corner can be modeled by a pixel-wise function of 2D coordinate map (*i.e.*, τ in Eqn. (2)) [56]. As a remedy, we simply feed τ to both the SPN and the GuideNet for handling anti-vignetting while learning color mapping simultaneously.

With the GCM output $\tilde{\mathbf{y}}$, we use PWC-Net [46] to estimate the optical flow for warping target sRGB image \mathbf{y} . The warped sRGB image \mathbf{y}^w can then be adopted as the supervision for training GCM. Besides, we note that the pixel-wise mapping makes GCM unable to remove the noise in $\hat{\mathbf{x}}$. Nonetheless, benefiting from PWC-Net [46], we can still estimate the optical flow between $\tilde{\mathbf{y}}$ and \mathbf{y} robustly.

3.3. LiteISPNet

For alleviating mild alignment, existing methods usually build large models and exploit specific modules [9, 22], which improves the performance at the cost of increasing inference time. Considering that a better alignment can be attained by joint learning, we can adopt a more efficient network for learning RAW-to-sRGB mapping to achieve comparable or even better performance. Thus, we present a LiteISPNet by simplifying MW-ISPNet [20], which is a U-Net [43] based multi-level wavelet ISP network. In particular, we put convolutional layer and residual group [63] before each wavelet decomposition by referring to [31]. Moreover, we also reduce the number of RCAB from 20 to 4 in each residual group to construct the LiteISPNet backbone. Fig. 2(c) illustrates the network structure of LiteISPNet. Benefited from the structure modification and joint learning, LiteISPNet outperforms MW-ISPNet [20] quantitatively and qualitatively with $\sim 40\%$ model size and $\sim 20\%$ running time.

3.4. Learning Objectives

Using the pre-trained PWC-Net [46] for computing optical flow, GCM and LiteISPNet can be jointly trained for learning image alignment and RAW-to-sRGB mapping. In the following, we respectively introduce the loss terms for GCM and LiteISPNet.

Loss Term for GCM. Denote by $\tilde{\mathbf{y}}$ the GCM output in Eqn. 2 and \mathbf{y}^w the warped target sRGB image in Eqn. 4. The loss term for GCM is given by,

$$\mathcal{L}_{GCM}(\tilde{\mathbf{y}}, \mathbf{y}^w) = \|\mathbf{m} \circ (\tilde{\mathbf{y}} - \mathbf{y}^w)\|_1, \quad (5)$$

where \circ denotes entry-wise product, $\|\cdot\|_1$ is ℓ_1 loss and \mathbf{m} is a mask indicating valid positions of the optical flow. Here, each element m_i of \mathbf{m} is defined as,

$$m_i = \begin{cases} 1, & [\mathcal{W}(\mathbf{1}, \Psi)]_i \geq 1 - \epsilon \\ 0, & \text{otherwise} \end{cases}, \quad (6)$$

where $\mathbf{1}$ denotes an all-1 matrix, ϵ is a threshold set to 0.001, and $[\cdot]_i$ denotes the i -th element of a matrix.

Loss Terms for LiteISPNet. Denote by $\hat{\mathbf{y}}$ the LiteISPNet output in Eqn. 1. The LiteISPNet is trained with a combination of ℓ_1 loss and (VGG-based) perceptual loss [36], which can be written as,

$$\begin{aligned} \mathcal{L}_{ISP}(\hat{\mathbf{y}}, \mathbf{y}^w) = & \lambda_{\ell_1} \|\mathbf{m} \circ (\hat{\mathbf{y}} - \mathbf{y}^w)\|_1 \\ & + \lambda_{VGG} \|\mathbf{m} \circ (\phi(\hat{\mathbf{y}}) - \phi(\mathbf{y}^w))\|_1, \end{aligned} \quad (7)$$

where ϕ denotes the pre-trained VGG-19 [36] network, and we set $\lambda_{\ell_1} = \lambda_{VGG} = 1$. Besides, to further enhance the visual quality, we also train the LiteISPNet with adversarial loss [14]. Following LSGAN [37], the loss function is defined as,

$$\mathcal{L}_{GAN} = \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} [\mathcal{D}(\mathcal{I}(\mathbf{x})) - 1]^2, \quad (8)$$

where \mathcal{D} denotes the discriminator (see the suppl for detailed structure configuration), which is trained by

$$\mathcal{L}_D = \frac{1}{2} \mathbb{E}_{\mathbf{y} \sim p_{\mathbf{y}}} [\mathcal{D}(\mathbf{y}) - 1]^2 + \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} [\mathcal{D}(\mathcal{I}(\mathbf{x}))]^2. \quad (9)$$

Then, LiteISPGAN is provided by training with the loss,

$$\mathcal{L}_{ISPGAN} = \mathcal{L}_{ISP} + \lambda_{GAN} \mathcal{L}_{GAN}, \quad (10)$$

where $\lambda_{GAN} = 0.01$.

Learning Objective. With the above loss terms, the overall learning objective of our model can be defined by,

$$\mathcal{L} = \mathcal{L}_{GCM} + \mathcal{L}_{ISP/ISPGAN}. \quad (11)$$

3.5. Extension to Other Image Enhancement Tasks

As previously discussed in Sec. 2.2, RAW-to-sRGB mapping has been combined with many other tasks, where considerable efforts have been paid to mitigate the influence of misalignment [6, 55, 62]. In these scenarios, the main obstacle to alignment is similar with the ISP problem [22], thus the proposed joint training framework can be naturally extended to many image enhancement tasks. In this paper, we conduct experiments on the raw image SR [62] task to show the generalization ability of our method.

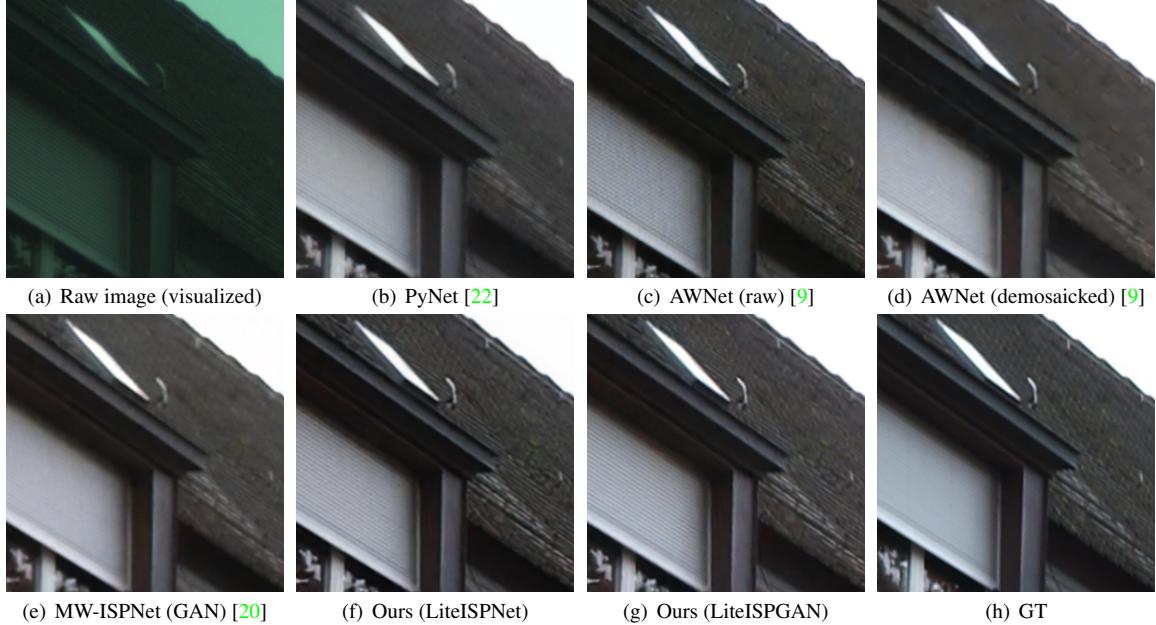


Figure 3: Visual comparisons on ZRR dataset. Please zoom in for better observation.

Table 1: Quantitative results on ZRR dataset for methods trained without and with adversarial loss.

Method	# Params (M)	Time (ms)	Original GT	Align GT with raw	Align GT with result
			PSNR↑ / SSIM↑ / LPIPS↓	PSNR↑ / SSIM↑ / LPIPS↓	PSNR↑ / SSIM↑ / LPIPS↓
PyNet [22]	47.6	62.7	21.19 / 0.7471 / 0.193	22.73 / 0.8451 / 0.152	22.97 / 0.8510 / 0.152
AWNet (raw) [9]	52.2	55.7	21.42 / 0.7478 / 0.198	23.27 / 0.8542 / 0.151	23.35 / 0.8559 / 0.151
AWNet (demosiaicked) [9]	50.1	62.7	21.53 / 0.7488 / 0.212	23.38 / 0.8497 / 0.164	23.41 / 0.8502 / 0.164
MW-ISPNet [20]	29.2	110.5	21.42 / 0.7544 / 0.213	23.07 / 0.8479 / 0.165	23.31 / 0.8578 / 0.164
Ours (LiteISPNet)	11.9	23.3	21.55 / 0.7487 / 0.187	23.76 / 0.8730 / 0.133	23.87 / 0.8737 / 0.133
MW-ISPNet (GAN) [20]	29.2	110.5	21.16 / 0.7317 / 0.159	22.80 / 0.8285 / 0.134	23.38 / 0.8513 / 0.131
Ours (LiteISPGAN)	11.9	23.3	21.28 / 0.7387 / 0.159	23.47 / 0.8642 / 0.120	23.56 / 0.8670 / 0.119

4. Experiments

4.1. Implementation Details

Datasets. We conduct experiments on two datasets, *i.e.*, Zurich RAW to RGB (ZRR) [22] and SR-RAW [62].

In the ZRR dataset, 20 thousand image pairs are collected and roughly aligned via SIFT keypoints [34] and the RANSAC algorithm [49], and the cropped patches with cross-correlation < 0.9 are discarded, resulting in 48,043 raw-sRGB pairs of size 448×448 . We follow the official division to train our LiteISPNet with 46.8k pairs, and report the quantitative results on the remaining 1.2k pairs.

In the SR-RAW dataset, there are 500 scenes of images collected. In each scene, the authors take 7 photos with various focal length (24, 35, 50, 70, 100, 150 and 240 mm), where the 24/100, 35/150 and 50/240 pairs form a $4 \times$ super-resolution dataset (*i.e.*, with 1,500 pairs in total). We use 400 scenes for training, 50 scenes for validation, and report the performance on 35/150 mm pairs of rest 50 scenes. For a fair comparison, we replace LiteISPNet by SRResNet [27] used in Zhang *et al.* [62] on the SR-RAW dataset.

Training Details. During training, data augmentation is applied on training images, including random horizontal

flip, vertical flip and 90° rotation. In each iteration, 16 packed raw patches with 4 channels are extracted as inputs. Our framework is optimized by ADAM algorithm [26] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for 100 epochs. The learning rate is initially set to 1×10^{-4} and is decayed to half after 50 epochs. The experiments are conducted with PyTorch [39] framework on an Nvidia GeForce RTX 2080Ti GPU.

Evaluation Configurations. To evaluate the performance quantitatively, we compute three metrics on RGB channels, *i.e.*, Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM) [51] and Learned Perceptual Image Patch Similarity (LPIPS) [61]. Note that in this paper, we use the version 0.1 of LPIPS trained on the AlexNet network. All results of the competing methods are generated via the officially released model. In addition, we also count the inference time on ZRR dataset to evaluate model efficiency.

Besides providing the metrics computed with *Original GT*, we additionally provide two sets of metrics for a comprehensive and more accurate comparison by considering the alignment. Specifically, we align y with GCM output \tilde{y} by PWC-Net [46], and the metrics computed with such warped y are denoted by *Align GT with raw*. In addition, considering that previous models trained with misaligned

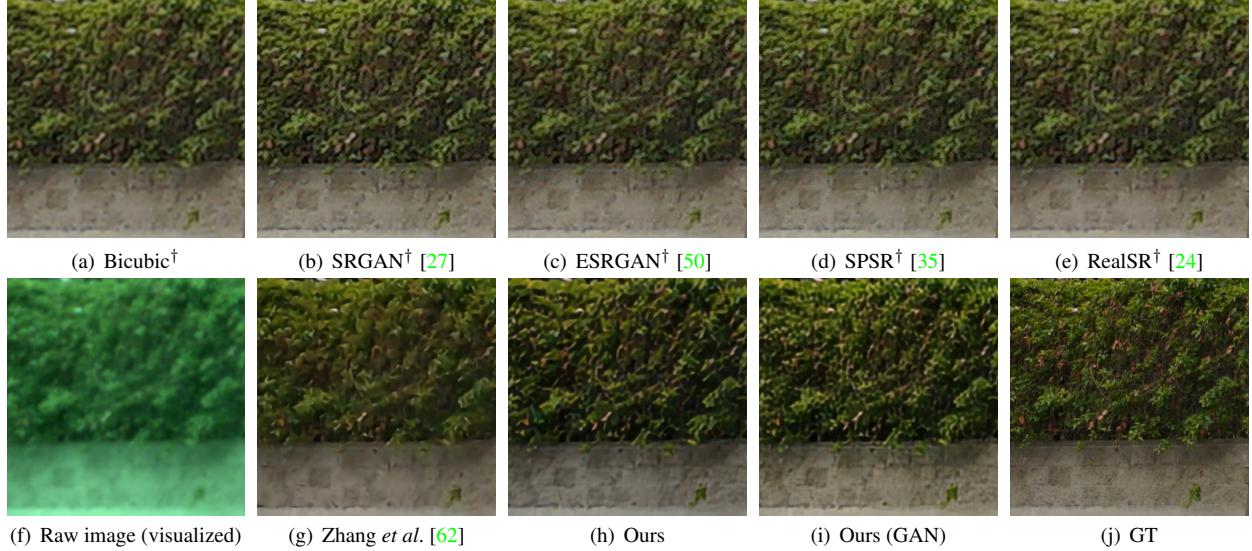


Figure 4: Visual comparison on SR-RAW dataset. \dagger means that the result is obtained given LR sRGB image as input.

Table 2: Quantitative results on SR-RAW dataset. Methods taking LR sRGB image as input are marked with \dagger .

Method	<i>Original GT</i>	<i>Align GT with raw</i>	<i>Align GT with result</i>
	PSNR \uparrow / SSIM \uparrow / LPIPS \downarrow	PSNR \uparrow / SSIM \uparrow / LPIPS \downarrow	PSNR \uparrow / SSIM \uparrow / LPIPS \downarrow
SRGAN \dagger [27]	18.42 / 0.5534 / 0.456	19.32 / 0.5999 / 0.419	21.89 / 0.6832 / 0.398
ESRGAN \dagger [50]	18.66 / 0.5563 / 0.435	19.55 / 0.6018 / 0.411	21.99 / 0.6785 / 0.393
SPSR \dagger [35]	18.64 / 0.5428 / 0.454	19.50 / 0.5854 / 0.441	21.90 / 0.6603 / 0.425
RealSR \dagger [24]	18.69 / 0.5570 / 0.435	19.58 / 0.6026 / 0.412	22.03 / 0.6796 / 0.394
Zhang et al. [62]	16.03 / 0.5184 / 0.517	17.43 / 0.5745 / 0.440	22.26 / 0.7205 / 0.372
Ours	17.74 / 0.5572 / 0.430	22.00 / 0.7049 / 0.346	22.50 / 0.7205 / 0.342
Ours (GAN)	17.71 / 0.5501 / 0.422	22.10 / 0.6996 / 0.340	22.59 / 0.7156 / 0.336

data may cause pixel shift in the result, we further align the ground-truth with the output of each method, and the metrics are denoted by *Align GT with result*.

4.2. Experimental Results on ZRR Dataset

To evaluate the effectiveness of the proposed method on ZRR dataset, we compare our model with three state-of-the-art methods, *i.e.*, PyNet [22], AWNet [9] and MW-ISPNNet [20]. Note that AWNet (raw) and AWNet (demosaicked) denote two models proposed in AWNet, which take 4-channel raw image and 3-channel demosaicked image as input respectively, and MW-ISPNNet (GAN) denotes MW-ISPNNet trained with adversarial loss.

As shown in Table 1, LiteISPGAN exceeds the competing methods on all metrics in *Align GT with raw* and *Align GT with result*. Furthermore, when training with adversarial loss, our LiteISPGAN achieves better LPIPS score (as well as the PSNR and SSIM indices) than MW-ISPNNet (GAN), which is the winner of perceptual track in the AIM 2020 Learned Smartphone ISP challenge [20]. It is worth mentioning that our model achieves the superior performance with a lightweight structure (the number of parameters is $\sim 25\%$ and $\sim 40\%$ of AWNet and MW-ISPNNet), and the inference time is only ~ 23 ms for a 448×448 input ($\sim 40\%$ and $\sim 20\%$ of AWNet and MW-ISPNNet).

Besides, we show the qualitative results in Fig. G. It can be seen that the results of PyNet, AWNet (demosaicked) and MW-ISPNNet (GAN) are blurry. AWNet (raw) is able to retain more details, however, it may generate artifacts and the result is less satisfactory. In contrast, our results are visually more pleasant while preserving finer details. Please refer to the suppl. for more results.

4.3. Experimental Results on SR-RAW Dataset

The proposed method is also evaluated on SR-RAW dataset for $4\times$ SR. Apart from the raw image SR method (Zhang et al. [62]), we also compare with state-of-the-art sRGB image SR methods (*i.e.*, SRGAN [27], ESRGAN [50], SPSR [35] and RealSR [24]), where low-resolution sRGB images provided in SR-RAW dataset are taken as input. The quantitative and qualitative results are given in Table 2 and Fig. 4, respectively.

We can see that SRGAN [27], ESRGAN [50], SPSR [35] and RealSR [24] tend to generate noisy results with undesired textures, and show unsatisfactory quantitative performance. Although trained with mild misalignment robust contextual bilateral (CoBi) loss, Zhang et al. [62] is unable to recover fine details and resulting in blurry results. Our method, with the same SRResNet [27] backbone as Zhang et al. [62], can preserve more textures and generate

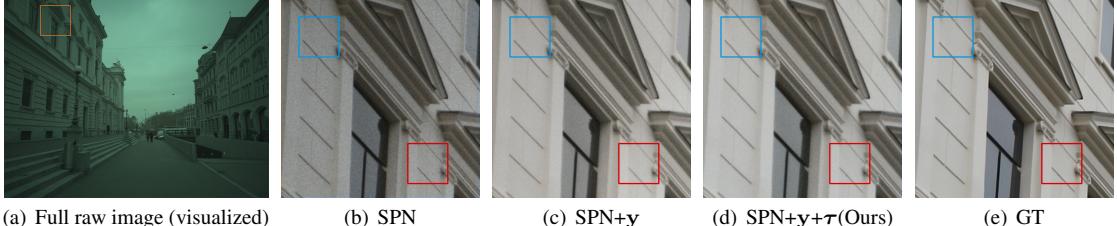


Figure 5: Visual results of GCM output (\hat{y}). With the guidance of y , the color of (c)(d) is closer to GT. Dark corner can also be observed, as the patch in blue box is darker in (b)(c) but has similar illumination with the patch in red box in (d)(e).

Table 3: Ablation study on alignment strategies.

Method	PSNR	LPIPS
SIFT (baseline)	23.49	0.148
Align y with \hat{y}	23.33	0.136
Align y with \hat{x}	23.52	0.135
Align y with \hat{y} (Ours)	23.76	0.133

Table 4: Ablation study on GCM. PSNR is calculated with GCM output (\hat{y}) and final result (\hat{y}), respectively.

GCM Components	PSNR (\hat{y})	PSNR (\hat{y})
N/A	-	23.49
SPN	20.67	23.61
SPN + y	26.33	23.69
SPN + y + τ (Ours)	26.93	23.76

Table 5: Ablation study on LiteISPNet.

# RCAB	Time (ms)	PSNR
2	15.3	23.54
4	23.3	23.76
8	38.0	23.74
20	91.5	23.79

neat results. More qualitative results are given in the suppl.

5. Ablation Study

In this section, we conduct extensive ablation studies on the proposed joint learning framework, and report the PSNR metric of *Align GT with raw* on the ZRR dataset.

5.1. Alignment Strategies

To solve the misalignment issue between input raw and target sRGB image, an intuitive way is to align them. For example, image pairs in ZRR dataset are roughly aligned by SIFT [34] algorithm. We train a LiteISPNet with such image pairs as a baseline (see Table 3), and evaluate several potential strategies for better alignment. (1) Learning image alignment jointly with RAW-to-sRGB mapping may lead to an iterative optimization process, so we align y with the output of LiteISPNet (\hat{y}) during training. Unfortunately, RAW-to-sRGB mapping is not a pixel-wise operation, resulting in more freedom for optical flow estimation and may cause more severe pixel shift. (2) We also estimate an optical flow between the demosaicked image \hat{x} and y , and align y with \hat{x} . Pixel shift is alleviated to some extent, yet the quality is still limited due to the color inconsistency. (3) With proposed GCM module, we can obtain \hat{y} whose color is consistent with y while the pixel positions are same with \hat{x} . Thus, our method performs favorably against other alignment strategies. Please refer to the suppl. for visual results.

5.2. GCM Module

To evaluate each individual component of GCM module, we further perform experiments as shown in Table 5. The baseline is a LiteISPNet trained with mildly aligned image pairs in the dataset. When adding an SPN to the baseline, a 0.12 dB PSNR gain is attained owing to a rough color correction. By further introducing the guidance provided by target sRGB image y and concatenating the coordinate map τ , the quality of GCM output \hat{y} is effectively improved (see Fig. C), which helps in optical flow estimation and provid-

ing better aligned supervision. As a result, the performance of the LiteISPNet is also promoted. Note that the ground-truth y is applied to generate the GCM output, leading to a higher PSNR metric than the LiteISPNet output. However, it is unavailable during inference of the LiteISPNet.

5.3. Structure of LiteISPNet

To explore the structure of LiteISPNet, we also conducted experiments on the number of residual channel attention blocks (RCABs) in each residual group (RG). As shown in Table 5, using 4 RCABs is sufficient with the well aligned training data, and deeper networks do not bring noticeable performance improvements. Therefore, we apply 4 RCABs to achieve a better efficiency-performance tradeoff in our LiteISPNet.

6. Conclusion

Learning with inaccurately aligned supervision is prone to causing pixel shift and generating blurry results, but existing methods usually fail to solve the inherent misalignment problem in many RAW-to-sRGB tasks due to the severe color inconsistency between raw and sRGB pairs. To diminish the effect of color inconsistency, we presented a global color mapping (GCM) module, where SPN is leveraged to avoid spatial shift of pixels, and the target sRGB image serves as a guidance to convert the color of the raw data. Then, a pre-trained optical flow estimation model (e.g., PWC-Net) is deployed for obtaining well aligned supervision, which is used to train the RAW-to-sRGB mapping in a joint learning manner. Extensive experiments on ZRR and SR-Raw datasets show that our proposed method can achieve better performance against state-of-the-art methods both quantitatively and qualitatively.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No.s U19A2073 and 62006064.

References

- [1] Jonathan T Barron. Convolutional color constancy. In *IEEE International Conference on Computer Vision (ICCV)*, pages 379–387, 2015. 3
- [2] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 60–65. IEEE, 2005. 3
- [3] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing (TIP)*, 27(4):2049–2062, 2018. 3
- [4] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3086–3095, 2019. 3
- [5] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Genet: Non-local networks meet squeeze-excitation networks and beyond. In *International Conference on Computer Vision Workshops (ICCVW)*, pages 0–0, 2019. 1, 3
- [6] Prashant Chaudhari, Franziska Schirrmacher, Andreas Maier, Christian Riess, and Thomas Köhler. Merging-isp: Multi-exposure high dynamic range image signal processing. *arXiv preprint arXiv:1911.04762*, 2019. 1, 3, 5
- [7] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3185–3194, 2019. 3
- [8] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3291–3300, 2018. 1, 3
- [9] Linhui Dai, Xiaohong Liu, Chengqi Li, and Jun Chen. Awnet: Attentive wavelet network for image isp. In *European Conference on Computer Vision Workshops (ECCVW)*, 2020. 1, 3, 5, 6, 7, 12, 13, 17, 18
- [10] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015. 2, 3
- [11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3
- [12] Michaël Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frédéric Durand. Deep joint demosaicking and denoising. *ACM Transactions on Graphics (TOG)*, 35(6):1–12, 2016. 3
- [13] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédéric Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 3
- [14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 5
- [15] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 3
- [16] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. Conditional sequential modulation for efficient global image retouching. In *European Conference on Computer Vision (ECCV)*, pages 679–695. Springer, 2020. 3, 5
- [17] Keigo Hirakawa and Thomas W Parks. Adaptive homogeneity-directed demosaicing algorithm. *IEEE Transactions on Image Processing (TIP)*, 14(3):360–369, 2005. 3
- [18] Yuanming Hu, Baoyuan Wang, and Stephen Lin. Fc4: Fully convolutional color constancy with confidence-weighted pooling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4085–4094, 2017. 3
- [19] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3277–3285, 2017. 3
- [20] Andrey Ignatov, Radu Timofte, et al. Aim 2020 challenge on learned image signal processing pipeline. In *European Conference on Computer Vision Workshops (ECCVW)*, 2020. 2, 3, 5, 6, 7, 12, 13, 17, 18
- [21] Andrey Ignatov, Radu Timofte, Sung-Jea Ko, Seung-Wook Kim, Kwang-Hyun Uhm, Seo-Won Ji, Sung-Jin Cho, Jun-Pyo Hong, Kangfu Mei, Juncheng Li, et al. Aim 2019 challenge on raw to rgb mapping: Methods and results. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 3584–3590. IEEE, 2019. 3
- [22] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 536–537, 2020. 1, 2, 3, 4, 5, 6, 7, 12, 13, 17, 18
- [23] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2462–2470, 2017. 2, 3
- [24] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 466–467, 2020. 7, 12, 13, 16, 17
- [25] Hamid Reza Vaezi Joze, Ilya Zharkov, Karlton Powell, Carl Ringler, Luming Liang, Andy Roulston, Moshe Lutz, and Vivek Pradeep. Imagepairs: Realistic super resolution dataset via beam splitter camera rig. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 518–519, 2020. 3
- [26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [27] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision*

- and Pattern Recognition (CVPR)*, pages 4681–4690, 2017. 3, 6, 7, 12, 13, 16, 17
- [28] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *European Conference on Computer Vision (ECCV)*, pages 272–289, 2018. 3
- [29] Zhetong Liang, Jianrui Cai, Zisheng Cao, and Lei Zhang. Cameranet: A two-stage framework for effective camera isp learning. *IEEE Transactions on Image Processing (TIP)*, 30:2248–2262, 2021. 1, 3
- [30] Lin Liu, Xu Jia, Jianzhuang Liu, and Qi Tian. Joint demosaicing and denoising with self guidance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2240–2249, 2020. 3
- [31] Pengju Liu, Hongzhi Zhang, Wei Lian, and Wangmeng Zuo. Multi-level wavelet convolutional neural networks. *IEEE Access*, 7:74973–74985, 2019. 5
- [32] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 773–782, 2018. 3
- [33] Zhongyu Lou, Theo Gevers, Ninghang Hu, Marcel P Lucassen, et al. Color constancy by deep learning. In *British Machine Vision Conference (BMVC)*, pages 76–1, 2015. 3
- [34] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004. 1, 3, 6, 8
- [35] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7769–7778, 2020. 7, 12, 13, 16, 17
- [36] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2017. 5
- [37] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2017. 5
- [38] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. In *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015. 3
- [39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019. 6
- [40] Rajeev Ramanath, Wesley E Snyder, Youngjun Yoo, and Mark S Drew. Color image processing pipeline. *IEEE Signal Processing Magazine*, 22(1):34–43, 2005. 1, 2
- [41] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking de-blurring algorithms. In *European Conference on Computer Vision (ECCV)*, pages 184–201. Springer, 2020. 3
- [42] Alessandro Rizzi, Carlo Gatta, and Daniele Marini. A new algorithm for unsupervised global and local color correction. *Pattern Recognition Letters*, 24(11):1663–1677, 2003. 3
- [43] 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. 5, 12
- [44] Eli Schwartz, Alex Bronstein, and Raja Giryes. Isp distillation. *arXiv preprint arXiv:2101.10203*, 2021. 3
- [45] Eli Schwartz, Raja Giryes, and Alex M Bronstein. Deepisp: Toward learning an end-to-end image processing pipeline. *IEEE Transactions on Image Processing (TIP)*, 28(2):912–923, 2018. 1, 3
- [46] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8934–8943, 2018. 2, 3, 4, 5, 6, 12
- [47] Runjie Tan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Color image demosaicking via deep residual learning. In *IEEE International Conference on Multimedia and Expo (ICME)*, volume 2, page 6, 2017. 3
- [48] Joost Van De Weijer, Theo Gevers, and Arjan Gijsenij. Edge-based color constancy. *IEEE Transactions on Image Processing (TIP)*, 16(9):2207–2214, 2007. 3
- [49] Andrea Vedaldi and Brian Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *ACM International Conference on Multimedia*, pages 1469–1472, 2010. 1, 3, 6
- [50] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 0–0, 2018. 7, 12, 13, 16, 17
- [51] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4):600–612, 2004. 6
- [52] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 101–117. Springer, 2020. 3
- [53] Chyuan-Tyng Wu, Leo F Isikdogan, Sushma Rao, Bhavin Nayak, Timo Gerasimow, Aleksandar Sutic, Liron Ainkerdem, and Gilad Michael. Visionisp: Repurposing the image signal processor for computer vision applications. In *IEEE International Conference on Image Processing (ICIP)*, pages 4624–4628. IEEE, 2019. 3
- [54] Jin Xiao, Shuhang Gu, and Lei Zhang. Multi-domain learning for accurate and few-shot color constancy. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3258–3267, 2020. 3
- [55] Xiangyu Xu, Yongrui Ma, and Wenxiu Sun. Towards real scene super-resolution with raw images. In *IEEE Conference*

- on Computer Vision and Pattern Recognition (CVPR)*, pages 1723–1731, 2019. 3, 5
- [56] Wonpil Yu. Practical anti-vignetting methods for digital cameras. *IEEE Transactions on Consumer Electronics*, 50(4):975–983, 2004. 5
- [57] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. CycleISP: Real image restoration via improved data synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2696–2705, 2020. 3
- [58] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 3
- [59] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing (TIP)*, 26(7):3142–3155, 2017. 3
- [60] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing (TIP)*, 27(9):4608–4622, 2018. 3
- [61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 6
- [62] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3762–3770, 2019. 1, 2, 3, 5, 6, 7, 12, 13, 16, 17
- [63] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 3, 5
- [64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 12

A. Content

The content of this supplementary material involves:

- Network structure of GCM, LiteISPNet and the discriminator of LiteISPGAN in Sec. B.
- Visual results of alignment in Sec. C.
- Qualitative results of ablation study in Sec. D
- Implementation details on SR-RAW dataset in Sec. E.
- Quantitative results for re-splitting the train/test set on ZRR dataset in Sec. F.
- Additional visual comparison results on SR-RAW and ZRR dataset in Sec. G.

B. Network Structure

Global color mapping (GCM) module involves two components: spatially preserving network (SPN) and GuideNet. SPN stacks 1×1 convolutional layers to guarantee spatial independence of the mapping and GuideNet generates a conditional guidance vector from the target sRGB to modulate SPN features. The detailed structure of GCM are shown in Table A.

The structure configuration of LiteISPNet are shown in Table B. LiteISPNet is a U-Net [43] based multi-level wavelet ISP network. In each residual group, we only apply 4 residual channel attention blocks (RCABs).

The discriminator structure of LiteISPGAN are shown in Table C. We apply 54×54 PatchGAN [64], which distinguishes whether the image patch is real or fake.

C. Visual Results of Alignment

We show the demosaicked raw image (\hat{x}), GCM output (\tilde{y}), LiteISPNet output (\hat{y}), warped target sRGB image (y^w) and the original target sRGB image (y) in Fig. A.

It can be seen that the color of \tilde{y} is consistent with y . Although GCM model cannot perform local operations (e.g., denoising), benefiting from PWC-Net [46], we can still align y with \tilde{y} robustly. Under the supervision of well aligned training data, LiteISPNet output has almost no pixel shift. In short, our method achieves the joint of image alignment and RAW-to-sRGB mapping.

D. Qualitative Results on Ablation Study

We show more qualitative results of different alignment strategies in Fig. B.

Due to the limitation of space in the submitted manuscript, we only showed the outputs of GCM model in Sec. 5.2. Here, by visualizing the illuminance ratio between the output of GCM and ground-truth (GT), we show the influence of each component and the dark corner phenomenon more clearly in Fig. C.

E. Implementation Details on SR-RAW Dataset

In each image pair of the SR-RAW dataset, the short focal-length raw image is used as input, while the long focal-length sRGB image is adopted as the ground-truth. In order to align high-resolution (HR) sRGB images with low-resolution (LR) raw images, we adopt downsampled HR sRGB image y_\downarrow to generate the conditional guidance vector in the GCM model. Then the optical flow between the GCM output \tilde{y} and y_\downarrow is estimated. Note that the size of optical flow is a quarter of the HR sRGB image. Thus, we upsample the optical flow to get the warped HR sRGB image. Finally, the warped HR sRGB image is utilized to supervise the learning of the backbone (SRResNet [27]).

Following [62], we use 400 scenes of images for training, 50 for validation, and the rest 50 for testing, and report the performance on 35/150 mm pairs in Table 2 of the main text. For a comprehensive comparison, we further show the performance on all 24/100, 35/150 and 50/240 test pairs in Table D. It can be seen that our method can still achieve better quantitative performance against all competing methods.

F. Quantitative Results for Re-splitting the Train/Test Set on ZRR Dataset

For the ZRR dataset, we follow the official division to train our LiteISPNet with 46,839 pairs, and report the quantitative results on the remaining 1,204 pairs in the main text. Here we conducted an experiment by re-splitting the dataset at approximately 9 : 1, i.e., 43,200 pairs for training and the rest 4,843 pairs for testing. Table E shows the quantitative results, and it can be seen that our LiteISPNet also exceeds the competing methods.

G. Additional Visual Comparison Results on SR-RAW and ZRR Dataset

In Fig. D~F, we show more qualitative comparison results generated by SRGAN [27], ESRGAN [50], SPSR [35], RealSR [24], Zhang *et al.* [62] and our models on the SR-RAW dataset.

In Fig. G~I, we show more qualitative results generated by PyNet [22], AWNet [9], MWISPNet [20] and our models on the ZRR dataset.

Table A: Structure configuration of GCM model. GCM involves two components: SPN (left column) and GuideNet (right column). Except for the stride of the first convolutional layer in GuideNet is 2, the stride of other convolutional layers is 1.

Spatially Preserving Network (SPN)				GuideNet			
Layer	Output size	Kernel size	Filter	Layer	Output size	Kernel size	Filter
Conv, ReLU	448 × 448	1 × 1	5 → 64	Conv	222 × 222	7 × 7	8 → 32
[Conv, ReLU] × 3	448 × 448	1 × 1	64 → 64	Conv, ReLU, Conv	222 × 222	3 × 3	32 → 32
Conv	448 × 448	1 × 1	64 → 3	Global Average Pooling	1 × 1	-	-
				Conv	1 × 1	1 × 1	32 → 64

Table B: Structure configuration of LiteISPNet. DWT and IWT denote discrete wavelet transform and inverse wavelet transform, respectively. RG denotes the residual group containing 4 residual channel attention blocks (RCABs).

LiteISPNet		
Layer	Output size	Filter
Conv	224 × 224	4 → 64
RG	224 × 224	64 → 64
DWT	112 × 112	64 → 256
Conv	112 × 112	256 → 64
RG	112 × 112	64 → 64
DWT	56 × 56	64 → 256
Conv	56 × 56	256 → 128
RG	56 × 56	128 → 128
DWT	28 × 28	128 → 512
Conv	28 × 28	512 → 128
RG	28 × 28	128 → 128
RG	28 × 28	128 → 128
Conv	28 × 28	128 → 512
IWT	56 × 56	512 → 128
RG	56 × 56	128 → 128
Conv	56 × 56	128 → 256
IWT	112 × 112	256 → 64
RG	112 × 112	64 → 64
Conv	112 × 112	64 → 256
IWT	224 × 224	256 → 64
RG	224 × 224	64 → 64
Conv	224 × 224	64 → 64
Conv	224 × 224	64 → 256
PixelShuffle	448 × 448	256 → 64
Conv	448 × 448	64 → 3

Table C: Structure configuration of the discriminator. The kernel size of all convolutional layers is 4×4 . The stride of the first three convolutional layers is 2, while the stride of the last two convolutional layers is 1.

Discriminator		
Layer	Output size	Filter
Conv, LeakyReLU	224 × 224	3 → 64
Conv, BatchNorm, LeakyReLU	112 × 224	64 → 128
Conv, BatchNorm, LeakyReLU	56 × 56	128 → 256
Conv, BatchNorm, LeakyReLU	55 × 55	256 → 512
Conv, BatchNorm, LeakyReLU	54 × 54	512 → 1

Table D: Average results on all 24/100, 35/150 and 50/240 test pairs of SR-RAW dataset. Methods taking LR sRGB image as input are marked with \dagger . The metrics are computed by Align GT with result.

Method	PSNR \uparrow / SSIM \uparrow / LPIPS \downarrow
SRGAN † [27]	21.72 / 0.6917 / 0.394
ESRGAN † [50]	21.85 / 0.6904 / 0.393
SPSR † [35]	21.75 / 0.6692 / 0.427
RealSR † [24]	21.89 / 0.6918 / 0.388
Zhang <i>et al.</i> [62]	21.97 / 0.7360 / 0.357
Ours	22.50 / 0.7369 / 0.329
Ours (GAN)	22.56 / 0.7341 / 0.323

Table E: Quantitative results for re-splitting the train/test set of ZRR dataset. The metrics are computed by Align GT with result.

Method	PSNR \uparrow / SSIM \uparrow / LPIPS \downarrow
PyNet [22]	22.67 / 0.8535 / 0.149
AWNet (raw) [9]	22.83 / 0.8513 / 0.160
AWNet (demosaicked) [9]	22.68 / 0.8447 / 0.173
MWISPNet [20]	23.00 / 0.8530 / 0.166
Ours (LiteISPNet)	23.31 / 0.8747 / 0.131

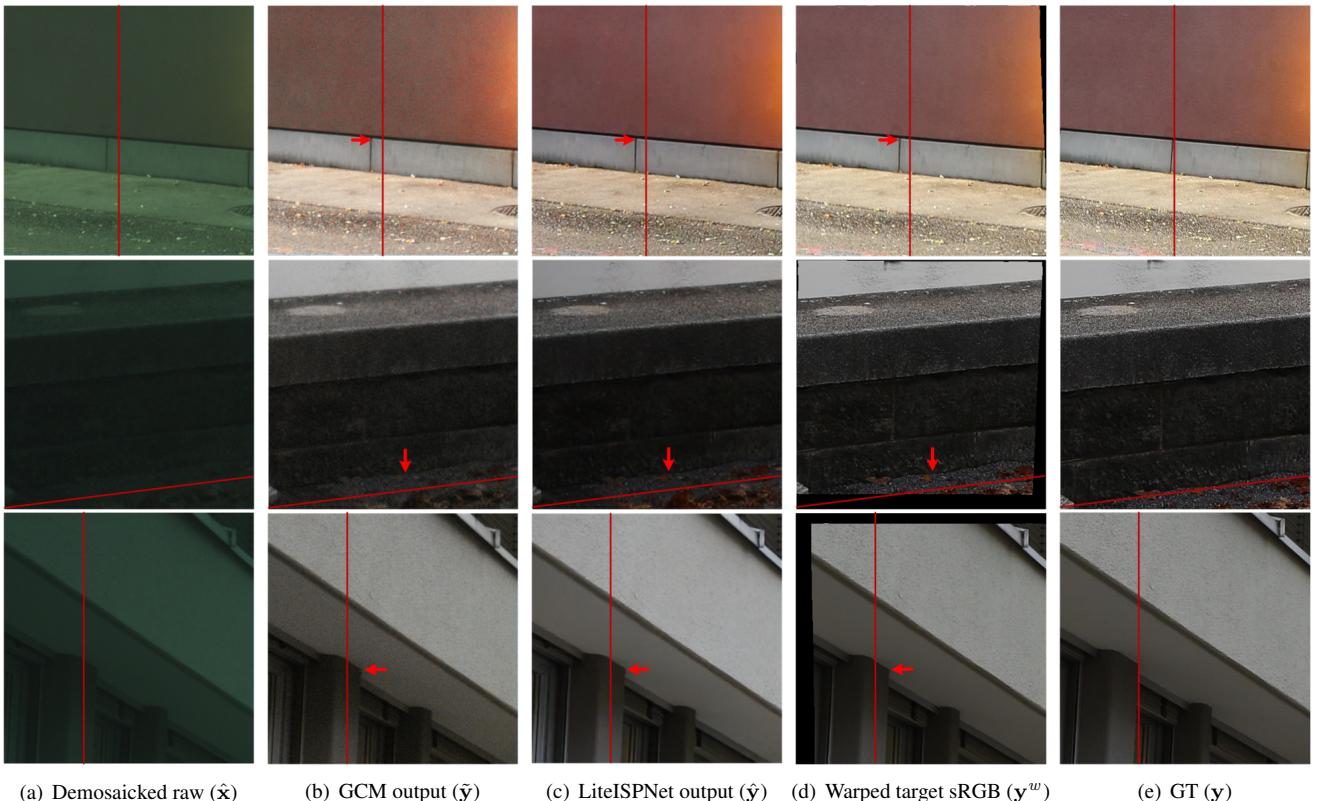


Figure A: Alignment visual results obtained by our joint learning framework. With the reference line, it can be observed our method obtains the well aligned data pairs while the demosaicked raw is not aligned with GT.

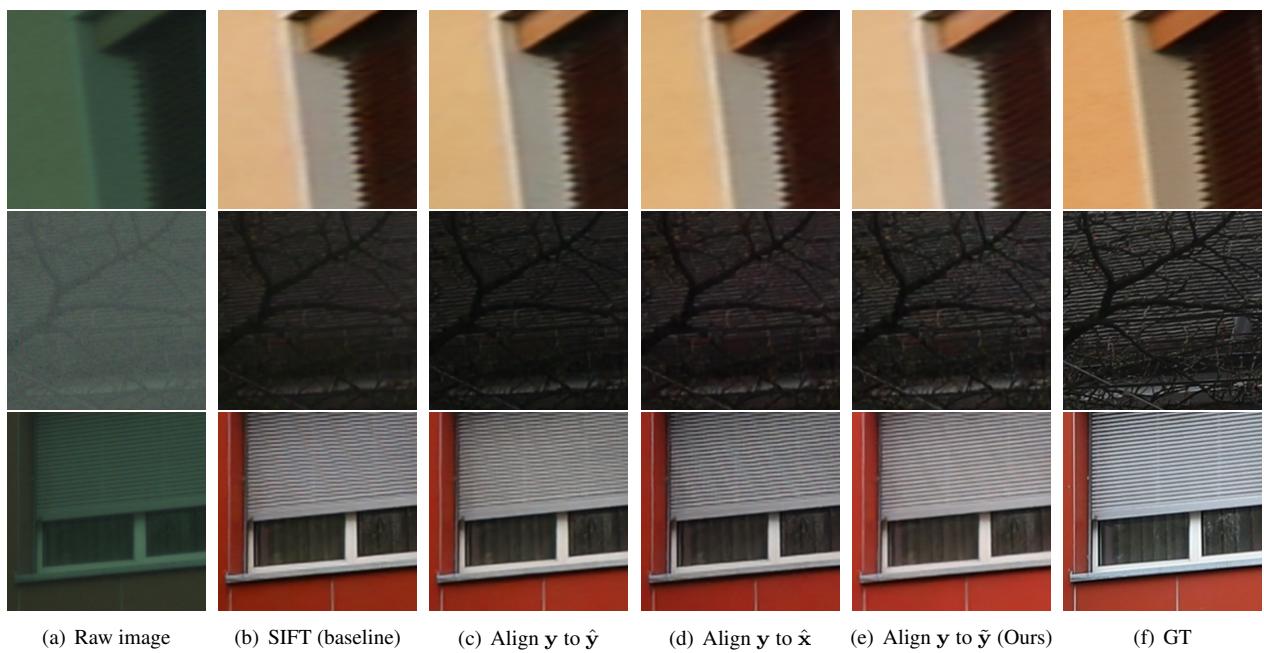


Figure B: Visual results of LiteISPNet output \hat{y} . (b)~(e) denote different alignment strategies. Our method (e) performs favorably against other alignment strategies.

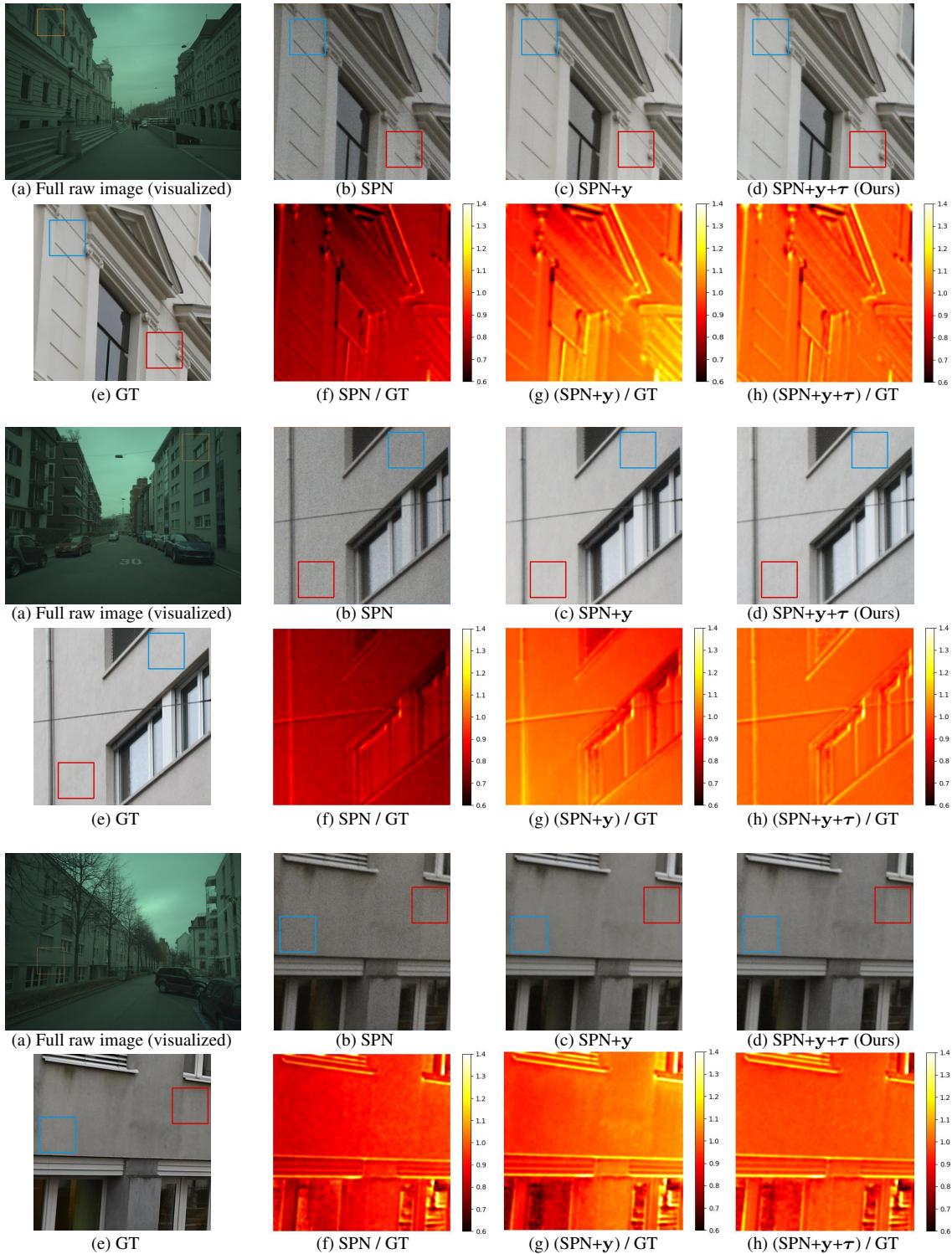


Figure C: Visual results of GCM output \hat{y} . (b)~(d) denote the results using different GCM components. (f)~(h) denote the illumination ratio between (b)~(d) and GT, respectively. With the guidance of y , the color of GCM output \hat{y} in (c) and (d) is closer to the target sRGB image. Dark corner can be observed in (b) and (c). In (b) and (c), the patch in the blue box is darker than the patch in the red box. But in (d) and (e), the patch in different boxes has similar illumination. The phenomenon can be seen more clearly in (f)~(h).

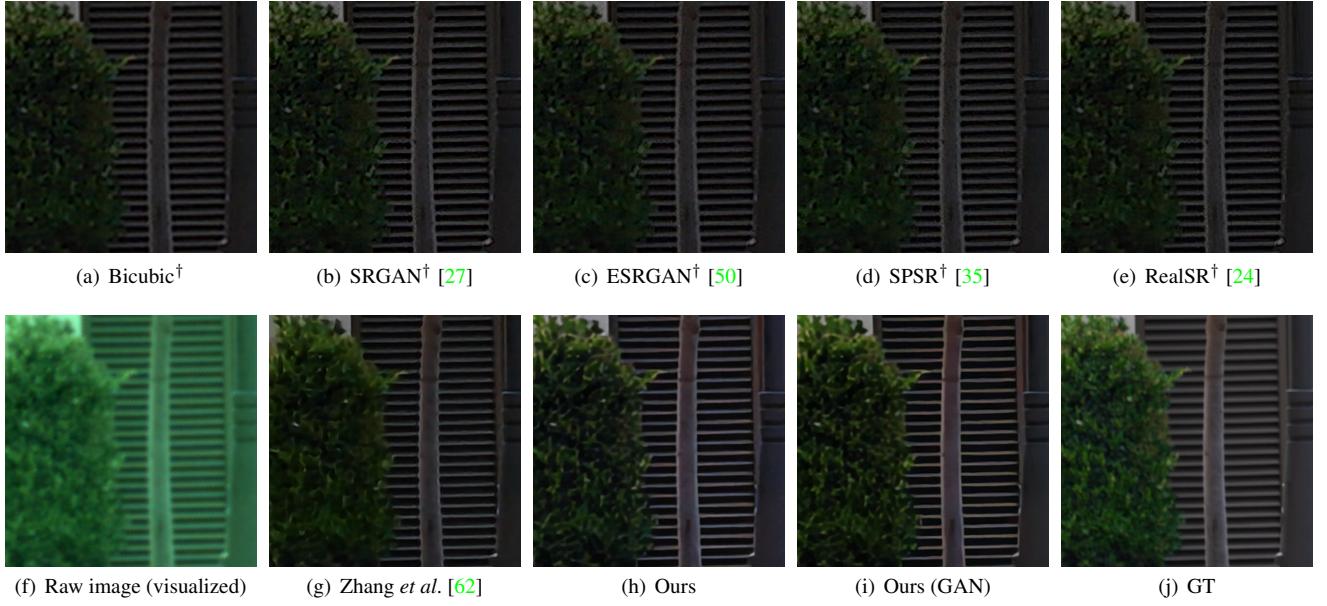


Figure D: Visual comparison on SR-RAW dataset. \dagger means that the result is obtained given LR sRGB image as input. Our results have more textures on the leaves.

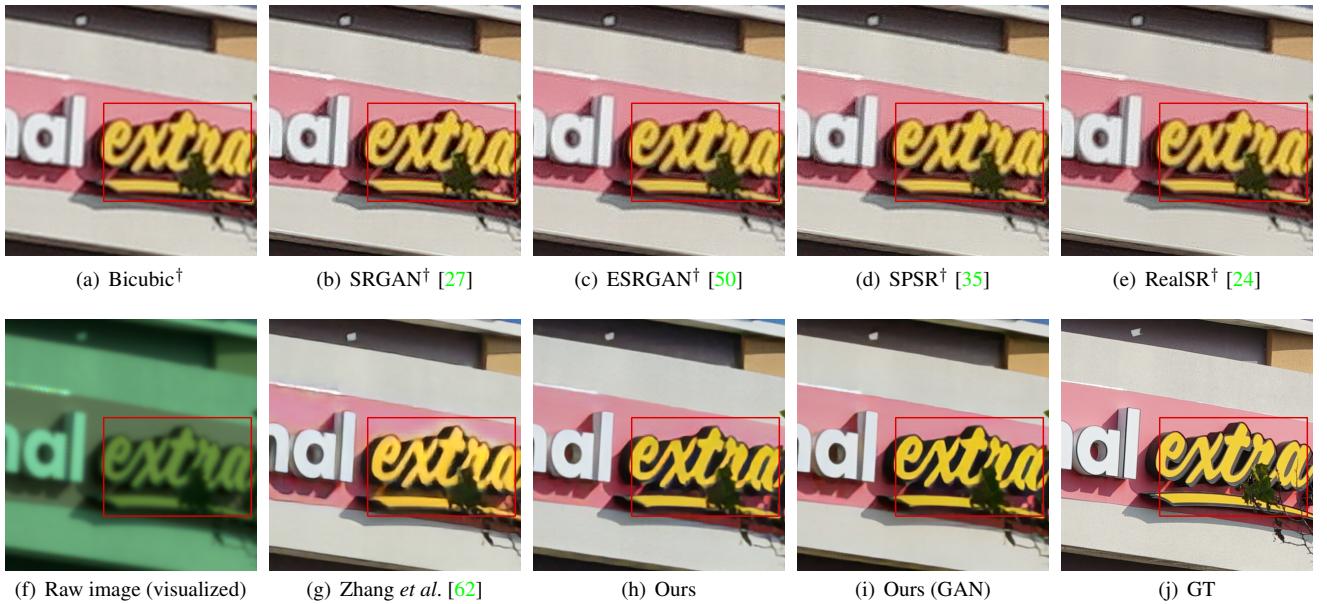


Figure E: Visual comparison on SR-RAW dataset. \dagger means that the result is obtained given LR sRGB image as input. The edges of our results are sharper. It can be clearly observed in the red box.

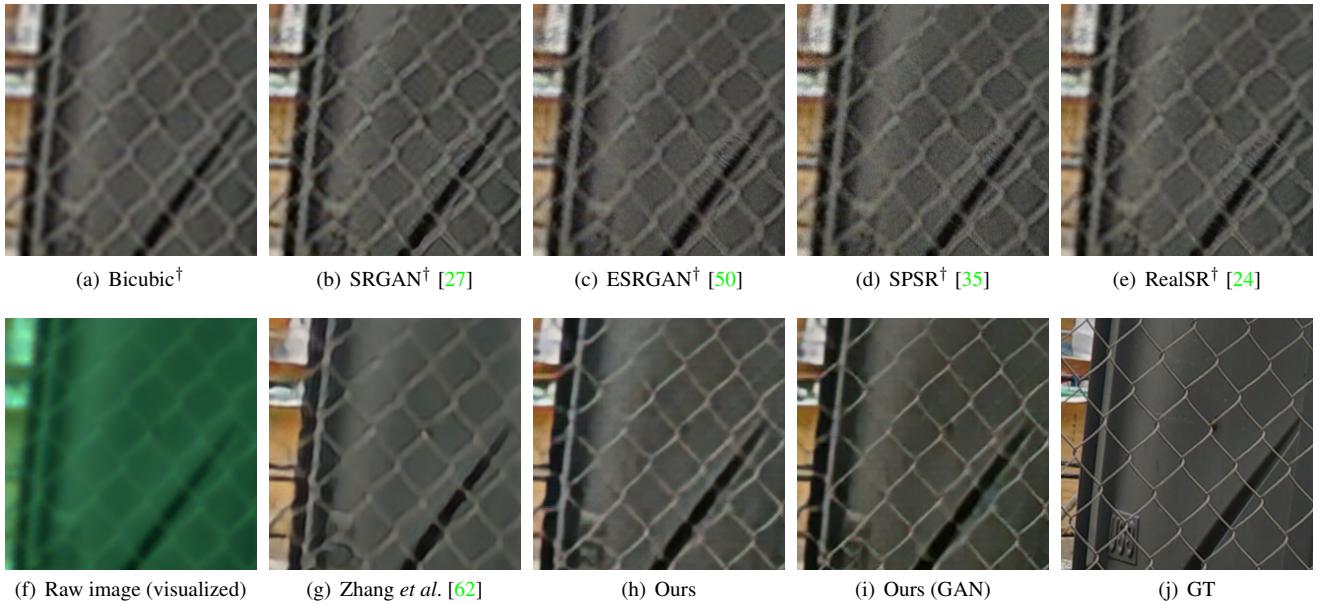


Figure F: Visual comparison on SR-RAW dataset. \dagger means that the result is obtained given LR sRGB image as input. The edges of our results are sharper.

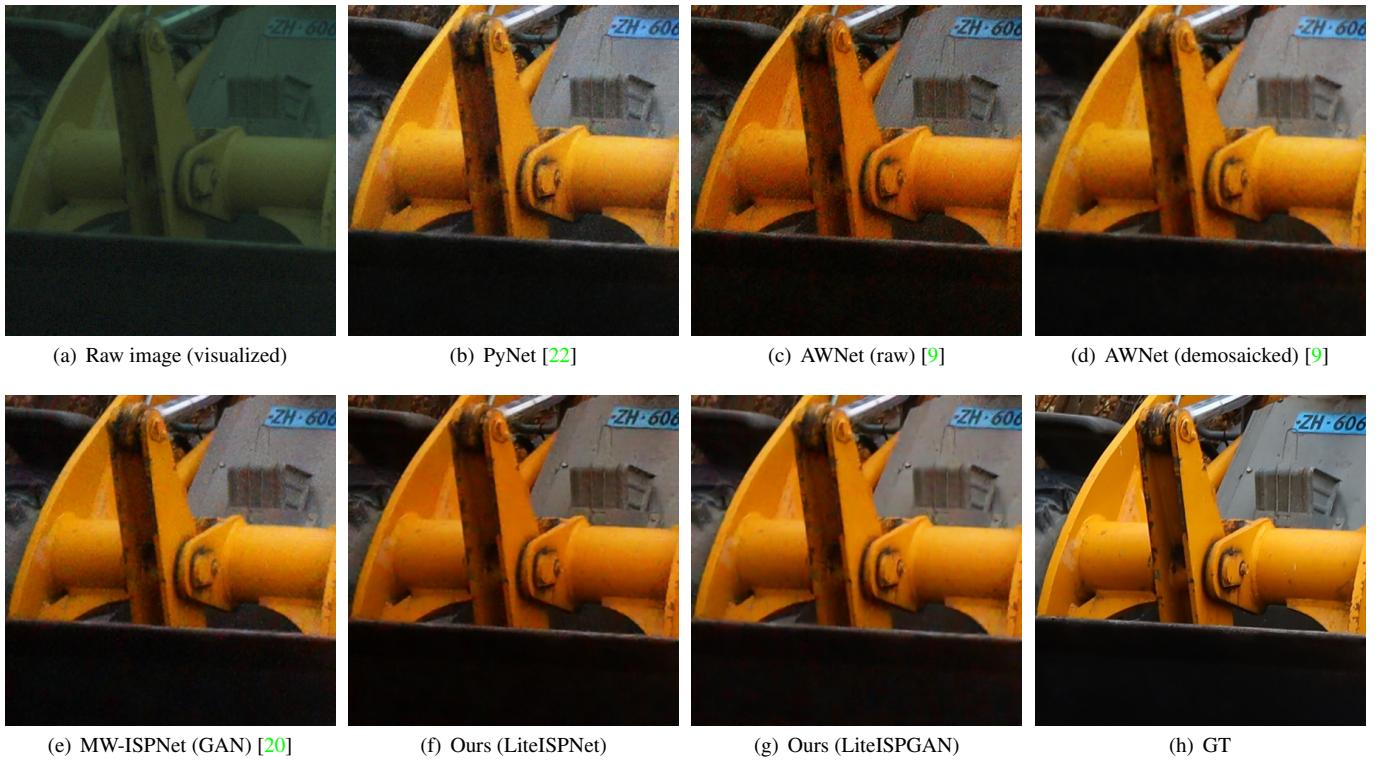


Figure G: Visual comparisons on ZRR dataset. Our results have less noise.



Figure H: Visual comparisons on ZRR dataset. Our results have richer textures on the grass.

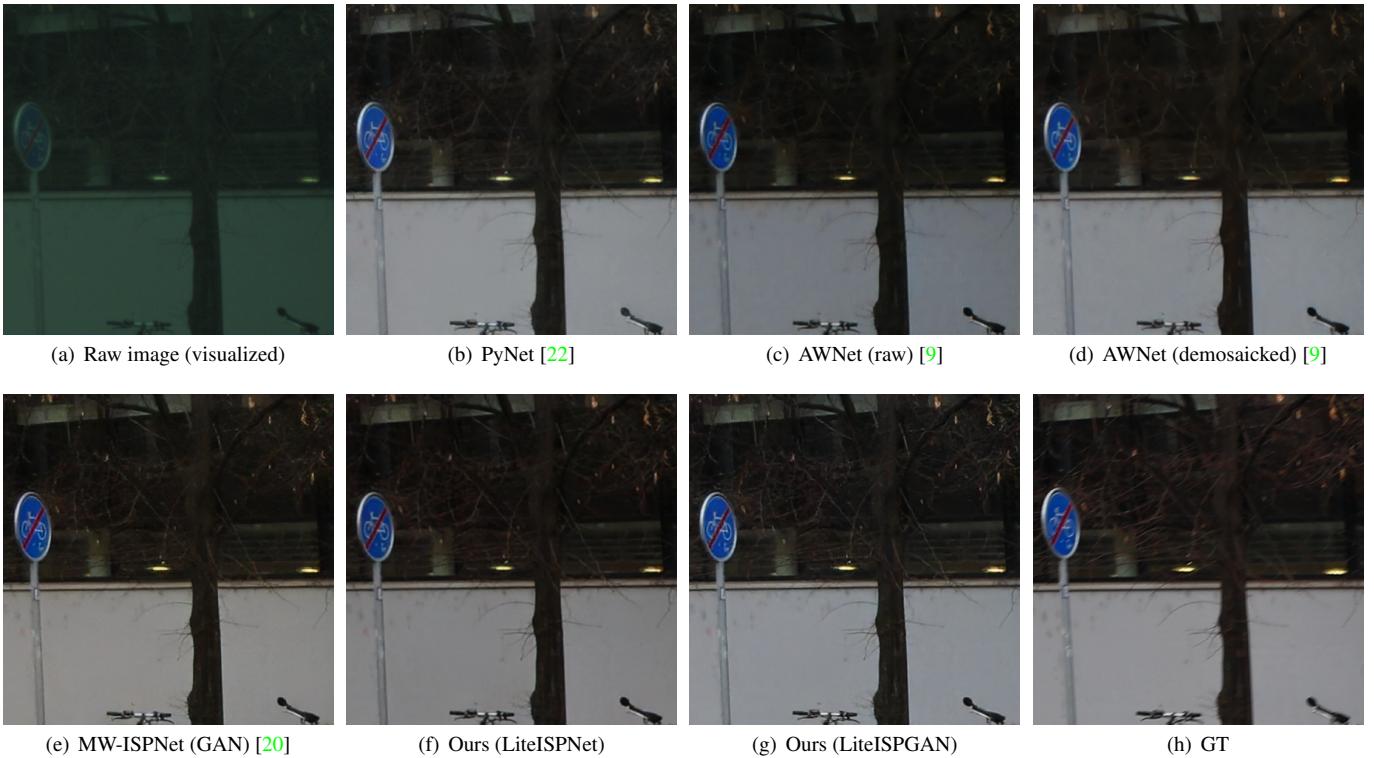


Figure I: Visual comparisons on ZRR dataset. The tree branches in our results are clearer.