From Synthetic to Real: A Calibration-free Pipeline for Few-shot Raw Image Denoising

Ruoqi Li, Chang Liu, Ziyi Wang, Yao Du, Jingjing Yang, Long Bao, Heng Sun
Video Algorithm Group, Camera Department, Xiaomi Inc.
Beijing, China
{liruoqi,liuchang58,wangziyi5,duyao3,yangjingjing3,baolong,sunheng3}@xiaomi.com

Abstract

*Calibration-based and paired data-based methods have achieved significant developments in the RAW image denoising field. However, the former requires accurate noise modeling to synthesize training data, which is laborious owing to the specificity across different camera sensors. Meanwhile, the latter relies on the large quantity and high quality of real paired datasets, which are difficult to collect in real-world scenarios. To overcome these limitations, we propose a simple pipeline termed as **S2R** to efficiently adapt **Synthetic** noise **to Real** noise. S2R contains i) a calibration-free synthetic pre-training stage to teach the network to recognize a variety of noise patterns and intensities and ii) a few-shot real fine-tuning stage for quickly adapting to target camera sensors. Moreover, a multi-perspective feature ensemble strategy is applied to enhance the network with stronger generalization ability and further boost the performance. We achieve a competitive score of 30.97 with PSNR 31.23dB and SSIM 0.95 on MultiRAW test set, ranking 1st place in the MIPI2024 Few-shot RAW Image Denoising Challenge.*

1. Introduction

Noise Reduction (NR) plays an important role in the image restoration field, since the image capturing process of the camera inevitably introduces different types of noises. The original RAW images produced by camera sensors possess the most primitive noise distribution for better distinguishing real signals, making RAW image denoising a popular topic.

Benefiting from the rapid development of deep learning techniques, utilizing real noisy-clean datasets [1, 2, 7, 22] for network training have made a breakthrough in RAW image denoising. However, these paired data-based methods [8, 9, 15, 23–25] highly rely on the quality and quantity of real paired datasets, which are extremely difficult to acquire for every specific camera sensor. Another mainstream approach is to train the network using synthetic noise data. While saving efforts in real data collection, the design of the physics noise model requires accurate alignment with the electronic imaging pipeline and careful calibration of noise distribution parameters, which suffer from disparity not only among different camera sensors but also between the simulation and real imaging processes.

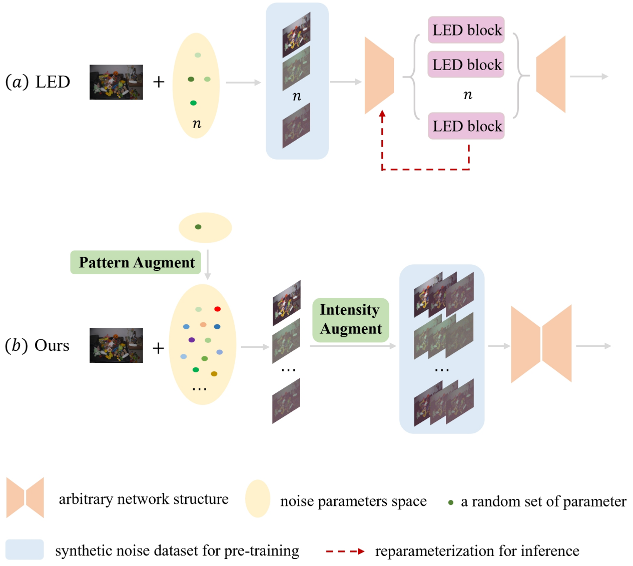

(a) LED
+
n
n
LED block
LED block
n
LED block
Pattern Augment
Intensity Augment
(b) Ours
+
...
...
...
arbitrary network structure
noise parameters space
a random set of parameter
synthetic noise dataset for pre-training
reparameterization for inference

Figure 1. The primary differences between (a) LED and (b) our S2R pipeline. On the one hand, given a clean image, LED produces only a fixed number of n virtual noise distributions for network pre-training, whereas our formation can enrich the variety of noise patterns and intensities to a large extend. On the other hand, the designed blocks of LED used for learning camera-specific features require further combination and reparameterization during inference, while our pipeline is more flexible for direct application to arbitrary existing networks without any modifications of the architectures.

To address the aforementioned limitations, a groundbreaking pipeline LED [13] is proposed to eliminate the need for noise model calibration and allow quick deployment to different camera sensors with minimal real data. It offers the innovative idea of implicitly calibrating the denoiser instead of explicitly constructing noise model through few-shot learning, giving rise to the Few-shot RAW Image Denoising track on Mobile Intelligent Photography and Imaging (MIPI) 2024 Workshop. In this challenge, with the purpose of performing raw image denoising in scenarios where only few-shot paired data are accessible, participants are asked to propose a denoiser with strong generalization ability to remove various levels of noise originating from distinct cameras.

As a solution to this challenge, we propose a simple pipeline S2R for efficiently adapting Synthetic noise to Real noise, which contains a calibration-free synthetic pre-training stage and a few-shot real fine-tuning stage for network learning. During pre-training, unlike LED which only samples a fixed number of virtual camera parameters with limited noise diversity (See Figure 1(a)), we build an enormous noise parameter set with the designed Pattern-Augment (PAug) module and Intensity-Augment (IAug) module (See Figure 1(b)), endowing the network with the power to recognize as many noise forms and levels as possible. Then, we perform fine-tuning using the accessed minimal real paired data to efficiently adapt to specific camera. The whole S2R pipeline can be directly applied to arbitrary network without any inner modifications of the structures, while LED requires additional combination and reparameterization step to embed their pre-trained blocks (See Figure 1(a)).

To further enhance the generalization ability of the network under real-world scenarios, we propose a Multi-Perspective Feature Ensemble (MPFE) strategy which is able to identify and strengthen noise features from different dimensions for better real signal extraction.

Extensive experiments have demonstrated the effectiveness of our solution. We summarize the contributions of this paper as follows.

- We aim to form an extremely wide range of random noise set with abundant noise patterns and intensities and propose a calibration-free two-stage training pipeline for efficiently adapting synthetic noise to real noise.
- We mine the characteristics of noise and propose a multi-perspective feature ensemble strategy to combine useful information for better network learning, which further boosts the generalization ability under complicated real-world scenarios.
- We rank 1st place in the MIPI2024 Few-shot RAW Image Denoising Challenge with our simple yet effective pipeline, which can be easily applied to many learning-based methods without any inner modification of the network structures.

2. Related Works

2.1. Real Data-based RAW Image Denoising

Training an end-to-end neural network using real noise-clean data pairs is the most intuitive idea for applying deep learning technique to RAW image denoising. Early researchers struggle with large-scale real RAW dataset collection, giving rise to a series of popular benchmarks [1, 2, 7, 22]. With the support of real noise database, an increasing number of elegant networks [8, 9, 15, 23–25] have been carefully designed to perform image denoising. While achieving encouraging results, difficulty of these methods lies in the acquisition of clean images, whose qualities are commonly impractical to guarantee when used as training labels. To address this problem, some works [12, 14, 17] adopt the idea of self-supervised learning using noise-noise data pairs, which liberate the strong dependence on clean images. However, these methods are built upon the statistical characteristics of noise, which may degenerate under challenging scenarios.

2.2. Physics-based RAW Image Modeling

To alleviate the burden of founding large-scale real paired datasets, another mainstream line is synthesizing noise data based on physics noise model. For the real-world RAW image noise, the most fundamental and popular noise model is Poisson-Gaussian distribution [11, 16], which assumes the photon shot noise to be Poisson and the remaining noise components to be Gaussian. For simplification, some works [3, 18] approximate the Poisson distribution to Gaussian and build the heteroscedastic Gaussian model with zero mean and signal-dependent variance. To improve the accuracy of noise modeling, ELD [22] disassembles the row noise and quantization noise out of Gaussian distribution for more concrete modeling. Yi *et al.*[27] further proposes to directly sample the signal-independent noise from dark frames with pattern-aligned and high-bit reconstruction strategies. Despite their significant contributions to noise modeling, discrepancies between synthetic and real noise are still unavoidable.

2.3. Few-shot RAW Image Denoising.

Owing to the different characteristics of cameras, methods based on noise modeling [6, 10, 20, 22, 26, 27] require careful parameter calibration process, which is laborious and easily influenced by the lighting environment. Recently, the groundbreaking method LED [13] brings the idea of few-shot learning to eliminate the needs for noise model calibration with a two-stage training pipeline. During pre-training, LED simulates several virtual cameras with randomly sampled noise parameters and subsequently designs camera-specific alignment block to align different features. For the fine-tuning, it inserts an out-of-model noise removal

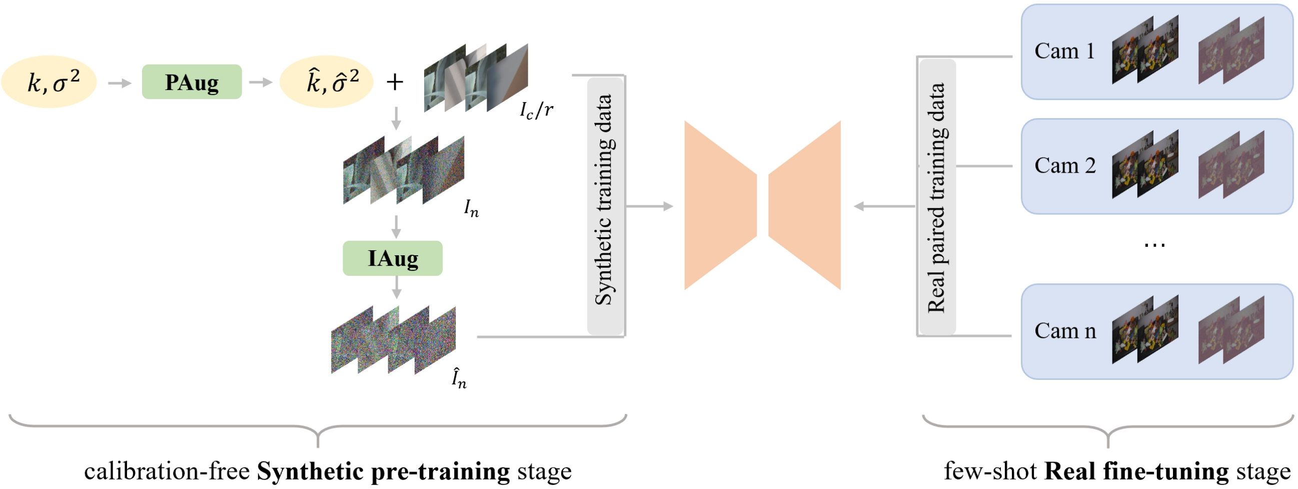

k, σ^2
PAug
$\hat{k}, \hat{\sigma}^2$
+
I_c/r
I_n
IAug
$\hat{I}_n$
Synthetic training data
Real paired training data
Cam 1
Cam 2
...
Cam n
calibration-free **Synthetic pre-training** stage
few-shot **Real fine-tuning** stage

Figure 2. Overview of the proposed two-stage Synthetic to Real (S2R) pipeline. During the pre-training stage, PAug module augments the initially sampled parameters k, σ^2 with abundant noise patterns and IAug module varies the synthetic data I_n with noise intensities. Then the network is quickly adapted to a specific camera sensor through fine-tuning using minimal real paired data.

branch to learn the gap between noise model and real noise using minimal real paired data.

Behind the enormous success of LED, some limitations still exist in terms of the lack of synthetic noise diversity and the complexity of network deployment. Inspired by the aforementioned pros and cons, our framework in this paper follows the two-stage few-shot pipeline with an enlarged synthetic noise set augmented by abundant noise diversities and the simplification for fast adapting to real noise, which shows strong generalization ability and further boosts the performance of RAW image denoising.

3. Method

Under the few-shot RAW image denoising framework, only a small amount of real-paired data is available given a specific camera sensor. We address this problem through a simple two-stage Synthetic to Real (S2R) pipeline (Section 3.2), which contains a calibration-free pre-training stage using synthetic noise data with extremely abundant noise patterns and intensities and an easily applied fine-tuning stage using the accessed few-shot real-paired data. To further strengthen the generalization ability and boost the performance in real-world scenarios, we design a Multi-Perspective Feature Ensemble (MPFE) strategy (Section 3.3) to selectively combine noise features from different dimensions. In the following, we first provide the preliminaries of physics noise model (Section 3.1) and then describe the proposed method and implementation in detail (Section 3.4).

3.1. Preliminaries of Noise Model Formation

When we look into the inner raw imaging process of camera sensors, the captured signals I_n are generally produced by converting the incident photons to digital values, which can be formulated as an ISO-related combination of the clean image I_c and various noise components N.

$$I_n = K_d(K_a(I_c + N_d + N_{i1}) + N_{i2}), \tag{1}$$

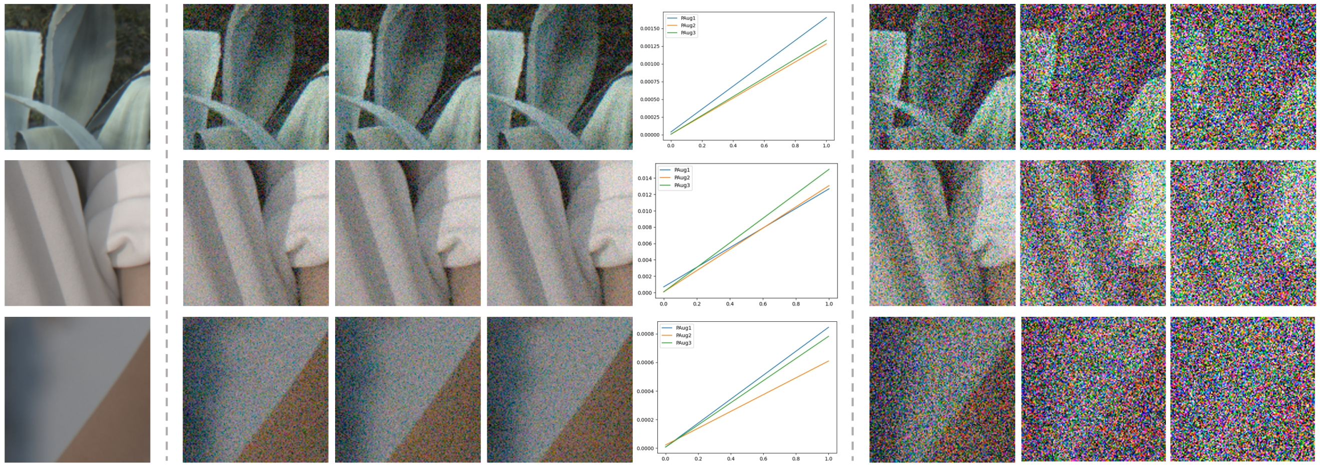
where K_a, K_d denote system analog gain and digital gain varied by ISO settings. N_d is the signal-dependent noises (*i.e.* photon shot noise) and N_i is the remaining signal-independent noises (*i.e.* dark noise, fix pattern noise, quantization noise, *etc.*). Eqn. 1 can be simplified as follows.

$$I_n = K(I_c + N_d) + B, \tag{2}$$

with $K = K_d K_a$, $B = K_d K_a N_{i1} + K_d N_{i2}$.

Considering the enormous randomness of real-world noise and the different characteristics of various camera sensors, a more fundamental and simpler noise model can possess a stronger generalization ability. Therefore, we choose the most representative Poisson-Gaussian distribution that has been widely explored in plenty of previous works [11, 16], which samples the signal-dependent and -independent components from Poisson $\mathcal{P}(\cdot)$ and Gaussian $\mathcal{N}(\cdot)$ distributions, respectively.

$$\begin{aligned} K(I_c + N_d) &\sim k\mathcal{P}(\frac{I_c}{k}), \\ B &\sim \mathcal{N}(0, \sigma^2). \end{aligned} \tag{3}$$



(a) Clean image (b) PAug1 (c) PAug2 (d) PAug3 (e) (f) IAug1 (g) IAug2 (h) IAug3

Figure 3. Visualization examples of the synthetic noise data generated by PAug (column (b)(c)(d)) and IAug (column (f)(g)(h)), respectively. Column (a) is the corresponding original clean images and column (e) plots the relations between real signal values in clean image and the corresponding noise variances to better illustrate different noise patterns.

Based on Eqn. 2 and Eqn. 3, our final noise model used for synthesizing noise data I_n from clean image I_c can be expressed as follows.

$$I_n = k\mathcal{P}(\frac{I_c}{k}) + \mathcal{N}(0, \sigma^2). \tag{4}$$

Moreover, we can build the relation between the parameters k and σ^2 through the mean and variance over I_n.

$$\begin{aligned} &E(I_n) = I_c, \\ &Var(I_n) = kI_c + \sigma^2. \end{aligned} \tag{5}$$

According to previous works [3], the parameters k and σ^2 follow linear and quadratic distributions with ISO, respectively, which can be fitted as follows.

$$\begin{aligned} &k = k_a * ISO + k_b, \\ &\sigma^2 = \sigma_a^2 * ISO^2 + \sigma_b^2 * ISO + \sigma_c^2. \end{aligned} \tag{6}$$

Therefore, for a clean image I_c, once we obtain the parameter set $\{k_a, k_b, \sigma_a^2, \sigma_b^2, \sigma_c^2\}$, we can calculate k, σ^2 with Eqn. 6 given an ISO value and then generate a noise distribution through Eqn. 4 to form the noise data I_n for training.

3.2. Synthetic to Real (S2R) Pipeline

As illustrated in Figure 2, the S2R pipeline contains calibration-free pre-training with synthetic noise data and few-shot fine-tuning with paired real data. We describe the two stages as follows.

3.2.1 Calibration-free Synthetic Pre-training

The aim of pre-training is to endow the network with strong generalization ability to rapidly adapt to specific camera without careful noise calibration. We achieve this goal by focusing on building an abundant synthetic noise set for the network to recognize enormous noise forms and levels.

Given the noise model in Eqn. 4 and the parameter formulations in Eqn. 6, we can add different noises on the clean images I_c by varying parameters under different ISO settings to form noise data I_n. Specifically, we first empirically generate a random parameter set of $\{k_a, k_b, \sigma_a^2, \sigma_b^2, \sigma_c^2\}$ to simulate an average noise level of different sensor types, then randomly choose an ISO value between ISO_{min} and ISO_{max} for calculating k and σ^2.

To increase the diversity of the synthetic noise set, we propose a Pattern-Augment (PAug) module for disturbing noise parameters and an Intensity-Augment (IAug) module for simulating various digital gains, giving rise to the final noise data $\hat{I}_n$ used in our pre-training stage.

$$\begin{aligned} &\hat{I}_n = \text{IAug}(I_n|_{\hat{k}, \hat{\sigma^2}}), \\ &\hat{k} = \text{PAug}(k), \hat{\sigma}^2 = \text{PAug}(\sigma^2). \end{aligned} \tag{7}$$

PAug module. Inspired by Eqn. 5, a set of parameter values of k, σ^2 can determine a specific noise form given a clean image under a constant ISO setting. Therefore, we propose pattern-augment module to simulate various noise forms derived from different types of cameras, which randomly disturbs the initially sampled parameters k and σ^2 to generate sensor-diversified parameters $\hat{k}$ and $\hat{\sigma}^2$.

$$\hat{k} = k + kp_1,$$
$$\hat{\sigma}^2 = \sigma^2 + \sigma^2 p_2, \tag{8}$$

where $p_1, p_2 \in [0, p], p \in (0, 1)$. The pattern augmented $\hat{k}$ and $\hat{\sigma}^2$ are used as the final parameters for a clean image to construct the corresponding noise distribution.

In Figure 3, we provide several visualization examples of different noise data generated by PAug for the same clean image under a same ISO setting. Since the differences in noise patterns are difficult to distinguish visually, we plot the relations of the real signal values in clean image I_c and their corresponding noise variances $Var(I_n)$ for better illustration. The slope and intercept of each curve indicate a set of k and σ^2 (See Eqn. 5), showing the diversity of the noise pattern simulated by PAug.

IAug module. Considering the different noise levels affected by additional digital gain under diverse luminance environments, we selectively apply intensity augmentation to the synthetic noise data I_n after PAug by multiplying a random ratio $r \in [1, 300]$ to amplify noise.

$$\hat{I}_n = rI_n,$$
$$I_n = \hat{k}\mathcal{P}(\frac{I_c/r}{\hat{k}}) + \mathcal{N}(0, \hat{\sigma}^2). \tag{9}$$

Note that the clean image is firstly divided by r to ensure normal luminance of the amplified noise data. We use the intensity-augmented synthetic noise $\hat{I}_n$ as the training data during the pre-training stage. Some visually examples are also depicted in Figure 3, which intuitively illustrates the various noise intensities produced by IAug.

Parameter settings. We experimentally set $ISO_{min} = 50$, $ISO_{max} = 6400$, $p = 0.3$ in our method. During the training process, p_1, p_2 in PAug and r in IAug are randomly regenerated within the corresponding ranges for each training sample, which can enlarge the randomness and diversity of the noise set without extensive efforts on hyper-parameter tuning.

3.2.2 Few-shot real fine-tuning

After the pre-training stage, the model is ready to transfer denoising ability from synthetic space to real cameras, bringing in the fine-tuning stage. Given a target camera, only pairs of real data captured by the sensor are required for few-shot learning. While saving huge effort on data collection, it can also reach the superior performance within a short period, enjoying both high efficiency and effectiveness. No special designs such as network modifications and training tricks are applied since we want the fine-tuning stage to be as easy to deploy as possible.

3.3. Multi-Perspective Feature Ensemble (MPFE) strategy

We further dive into the noise characteristics in real scenarios with the purpose of inferring the important ones that benefit network learning. Firstly, we consider that the spatial resolution of the noise pattern varies with noise intensity. The stronger the noise, the larger the receptive field required to distinguish the original signals from noise. Therefore, the network can be enhanced with the ability to recognize real signals from various scales. Secondly, we observe negative values for the noise data owing to the distribution characteristics of noise. While maintaining an accurate and complete noise form, this may also disrupt the stability of the learning process.

Therefore, we propose a Multi-Perspective Feature Ensemble (MPFE) strategy to combine the advantages of noise features from different dimensions. As shown in Figure 4, we construct the training noise data following two principles: i) different input patch sizes during the pre-training stage to enrich multi-scale information and ii) whether to maintain the negative values for both accurate noise model learning and better network convergence. We separately train four models with different entries and integrate the results through average summation, which further strengthens the generalization ability and boost the performance.

Note that the proposed MPFE strategy is an additional bonus for our model to further reach to better behavior, while our original S2R pipeline has already achieved an outstanding performance (See Section 4.3). However, the averaged ensemble operation may introduce over-smoothing problem on the denoising results, leaving us future explorations on extracting fine-grained detail features and weighted combination strategy.

3.4. Implementation Details

Network Architecture. We choose the original NAFNet [15] as our main network structure, which is a recently proposed simple baseline for image restoration task building upon UNet [19] architecture with couples of simple attention modules. We also apply our method on other popular structures and provide the experimental results in Section 4.3.

Training Data. In the pre-training stage, clean raw images are required to generate synthetic noise data. In order to obtain a strong pretrained model which can effectively adapt to different sensors under various real-world scenarios, we take the scene diversity and image quality into consideration and carefully select $\sim 3.5k$ high-quality clean images from two public datasets Fivek [4] and RealSR [5] captured by DSLR. In the fine-tuning stage, we only use the paired real data from two unknown types of cameras released by the challenge organizer.

Training Details. As illustrated in Section 3.2, the model is

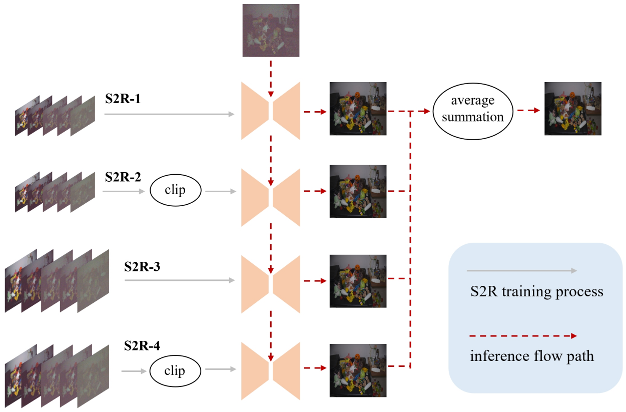


Figure 4. Overview of the Multi-Perspective Feature Ensemble (MPFE) strategy. Four entries varied by input resolution and negative value retention are provided to extract different noise characteristics, which is further combined to generate more robust results.

trained in two stages. We detail the training configurations as follows.

Synthetic pre-training stage. We pretrain the model from scratch using the synthetic noise data aforementioned, with the batch size of 8 and patch size of 224 for around 300k iterations. The model is optimized by AdamW optimizer using L1 loss, with the initial learning rate of $3e-4$, which decreases by 0.6 in 100k and 200k iterations. This process lasts for around three days on a single A100 GPU.

Real fine-tuning stage. We further finetune the model using the provided real paired data for each specific camera, with the batch size of 4 and patch size of 640 for around 5k iterations. Data augmentations include random flip and rotation are applied. The model is optimized by AdamW optimizer using L1 loss, with the initial learning rate of $1e-4$ and decreases by 0.6 in 1k and 3k iterations. It only takes less than 5 hours to achieve the optimal results.

Note that given a new camera sensor, we only need to perform few-shot fine-tuning without any inner modification of the network, which can be completed within a few hours to achieve optimal results, showing strong generality and efficiency.

4. Experiments

We first provide the ablation studies to verify our main contributions, followed by the comparison of our pipeline against other methods.

4.1. Datasets and Evaluation Metrics

Datasets. We evaluate our method on part of the MultiRAW dataset officially provided by MIPI2024 challenge, which is captured from two unknown types of cameras. For the main results in comparison with other methods (Section 4.3), we use the label-inaccessible test set with 120 noise data for each camera and run the results on the online server. For the ablation studies (Section 4.2), in order to conveniently perform offline evaluation on our own, we randomly crop 120 patches from the provided training set with ground truth for each camera under the same resolution as the test set.

PAug	IAug	pre-training			fine-tuning		
		Score	PSNR	SSIM	Score	PSNR	SSIM
-	-	18.23	19.37	0.812	-	-	-
✓	-	20.43	21.24	0.863	28.97	29.23	0.953
-	✓	21.31	22.17	0.855	27.79	28.07	0.951
✓	✓	**23.44**	**24.23**	**0.866**	**30.98**	**31.24**	**0.954**

Table 1. Ablations of PAug module and IAug module on the sampled training set.

Variants	size128	size224	clip	woclip	Score	PSNR	SSIM
S2R-1	✓			✓	30.89	31.15	0.953
S2R-2	✓		✓		30.90	31.16	0.954
S2R-3		✓		✓	30.83	31.10	0.952
S2R-4		✓	✓		**30.98**	**31.24**	**0.954**
S2R+MPFE	✓	✓	✓	✓	**31.22**	**31.47**	**0.956**

Table 2. Ablations of the MPFE strategy on the sampled training set. The top four rows indicate the single models with different input characteristics and the last row is the ensemble results of them.

Evaluation metrics. The evaluation metrics include the standard Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index [21] (SSIM) in grayscale. The overall score is formulated as follows.

$$Score = PSNR - \log_k(SSIM), \tag{10}$$

where $k = 1.2$ in the official implementation of the challenge. We use the average results of all predictions regardless of camera as the final evaluation score.

4.2. Ablation Studies

To verify the effectiveness our main contributions, we conduct evaluations on different variants of our method using the sampled dataset described in Section 4.1.

Pattern-Augment (PAug) and Intensity-Augment (IAug) module. One of our main contributions is building an enormous synthetic noise set for the network to cover abundant noise patterns and intensities, which is achieved using the proposed PAug and IAug. We evaluate the two modules in Table 1, where the first row denotes a baseline method which simply samples a set of noise parameters without any augmentation. Then we add PAug and IAug separately upon the baseline. The last row is our final S2R pipeline without MPFE strategy. Results demonstrate that the proposed modules can bring considerable performance boosts.

Ranking	Score	PSNR	SSIM
1 (S2R+MPFE)	**30.97**	**31.23**	**0.95**
1 (S2R-4)	**30.72**	**30.99**	**0.95**
2	30.71	30.96	0.95
3	29.69	29.98	0.95
4	29.63	29.93	0.95
5	29.61	29.93	0.94

Table 3. The released top-five results of MIPI2024 challenge on the provided test set. Our original pipeline without MPFE strategy (the second row) can still reach the top performance.

Method	Network	Score	PSNR	SSIM
		pre-training		
LED	UNet	19.40	20.42	**0.83**
S2R-4	UNet	**19.58**	**20.67**	0.82
LED	Restormer	21.36	22.35	0.84
S2R-4	Restormer	**22.32**	**23.28**	**0.84**
LED	NAFNet	21.32	22.25	0.86
S2R-4	NAFNet	**23.38**	**24.19**	**0.87**
		fine-tuning		
LED	NAFNet	28.24	28.53	0.95
S2R-4	NAFNet	**30.72**	**30.99**	**0.95**
S2R+MPFE	NAFNet	**30.97**	**31.23**	**0.95**

Table 4. Comparison results of our S2R with the state-of-the-art LED for both pre-training and fine-tuning stage on the provided test set.

Multi-Perspective Feature Ensemble (MPFE) strategy. The MPFE strategy is designed to combine diverse noise information to augment our S2R pipeline with more robust network learning. As illustrated in Figure 4, we differ the entries of the network into four variants (denoted as S2R-1,2,3,4) with different patch size and whether to apply clip operation to the noise data. The results are summarized in Table 2, where 'size128' and 'size224' denote cropping the inputs to 128×128 and 224×224, respectively. 'clip' means the noise data is clamped to $[0, 1]$ before feeding into the network while 'woclip' means maintaining the original data unchanged (*i.e.* keep the negative values of the noise). The last row shows the score of applying MPFE strategy, which integrates the predictions of the above four single models through average summation. The results indicate an obvious performance boost of the MPFE strategy compared to all the ensemble models, verifying its strong ability to better distinguish real signals from noise.

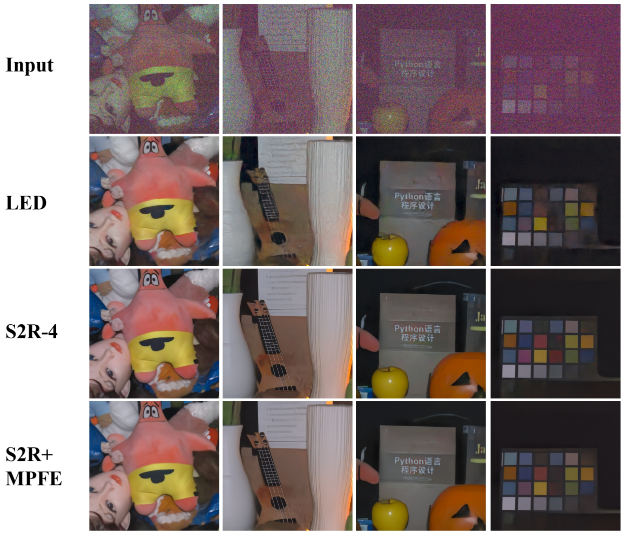


Figure 5. Qualitative comparisons on the provided test set. The first row is the noise inputs, followed by the predictions of LED, our single model S2R-4 and MPFE augmented model, respectively.

4.3. Comparison with State-of-the-Arts

Results on MIPI2024 challenge. We reach 30.97 Score with PSNR 31.23 and SSIM 0.95 on the real test set, ranking 1st on the MIPI2024 Few-shot RAW Image Denoising Challenge. As illustrated in Table 3, we lead the second and third place by 0.26 and 1.28 Score, respectively. Note that even the single model S2R-4 without MPFE strategy can still achieve superior results, verifying the effectiveness of the proposed method.

Comparison with LED. We compare our S2R pipeline with the state-of-the-art few-shot method LED in Table 4. Evaluation results of LED for pre-training are directly inferred using the pre-trained model released by the authors. For fine-tuning, we reproduce the results using the official code and configuration with the same training set as ours. We can see that our S2R outperforms LED on both stages, showing the stronger generalization ability of our pipeline.

We also provide the qualitative comparison in Figure 5. The visualization results show that the model trained with S2R can better distinguish real signals with clearer details than LED, while MPFE can further improve the accuracy of noise removal.

Apply S2R to different network architecture. We further equip our pipeline on other network architectures including UNet [19] and Restormer [25]. The results are also displayed on Table 4, which shows that our S2R pipeline can achieve competitive performance on arbitrary networks.

5. Conclusion

In this paper, we propose a simple yet effective two-stage pipeline S2R for learning from synthetic noise to real noise. In the pre-training stage, we synthesize an enormous noise set with abundant noise forms and levels through the designed Pattern-Augment (PAug) module and Intensity-Augment (IAug) module without parameter calibration, which enhances the network with strong generalization ability for efficiently adapting to specific camera sensor fine-tuned with few-shot paired real data. We further propose a Multi-Perspective Feature Ensemble (MPFE) strategy to utilize different characteristics of noise features to better extract the real signals, which significantly boosts the performance. We achieve 1st place in the MIPI2024 Few-shot RAW Image Denoising Challenge, offering an easily-plugged pipeline for future research.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[2] Josue Anaya and Adrian Barbu. Renoir - a dataset for real low-light noise image reduction. *arXiv preprint arXiv:1409.8230*, 2014. 1, 2

[3] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. Unprocessing images for learned raw denoising. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11028–11037, 2019. 2, 4

[4] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 5

[5] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 5

[6] Yue Cao, Ming Liu, Shuai Liu, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Physics-guided iso-dependent sensor noise modeling for extreme low-light photography. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5744–5753, 2023. 2

[7] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3291–3300, 2018. 1, 2

[8] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 182–192, 2021. 1, 2

[9] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22367–22377, 2023. 1, 2

[10] Hansen Feng, Lizhi Wang, Yuzhi Wang, Haoqiang Fan, and Hua Huang. Learnability enhancement for low-light raw image denoising: A data perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):370–387, 2024. 2

[11] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008. 2, 3

[12] Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14781–14790, 2021. 2

[13] Xin Jin, Jia-Wen Xiao, Ling-Hao Han, Chunle Guo, Ruixun Zhang, Xialei Liu, and Chongyi Li. Lighting every darkness in two pairs: A calibration-free pipeline for raw denoising. 2023. 2

[14] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018. 2

[15] Chen Liangyu, Chu Xiaojie, Zhang Xiangyu, and Sun Jian. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33, 2022. 1, 2, 5

[16] Xinhao Liu, Masayuki Tanaka, and Masatoshi Okutomi. Practical signal-dependent noise parameter estimation from a single noisy image. *IEEE Transactions on Image Processing*, 23(10):4361–4371, 2014. 2, 3

[17] Ali Maleky, Shayan Kousha, Michael S. Brown, and Marcus A. Brubaker. Noise2noiseflow: Realistic camera noise modeling without clean images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17611–17620, 2022. 2

[18] Ben Mildenhall, Jonathan T. Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2502–2510, 2018. 2

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015*, pages 234–241, 2015. 5, 7

[20] Yuzhi Wang, Haibin Huang, Qin Xu, Jiaming Liu, Yiqun Liu, and Jue Wang. Practical deep raw image denoising on mobile devices. In *European Conference on Computer Vision (ECCV)*, pages 1–16, 2020. 2

[21] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 6

[22] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8520–8537, 2021. 1, 2

[23] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *European conference on computer vision*, 2020. 1, 2

[24] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[25] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 7

[26] Feng Zhang, Bin Xu, Zhiqiang Li, Xinran Liu, Qingbo Lu, Changxin Gao, and Nong Sang. Towards general low-light raw noise synthesis and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10820–10830, 2023. 2

[27] Yi Zhang, Hongwei Qin, Xiaogang Wang, and Hongsheng Li. Rethinking noise synthesis and modeling in raw denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4593–4601, 2021. 2