DNF: Decouple and Feedback Network for Seeing in the Dark

Xin Jin[1*] Ling-Hao Han[1*] Zhen Li[1] Chun-Le Guo[1†] Zhi Chai[2] Chongyi Li[3]
[1]VCIP, CS, Nankai University [2]Hisilicon Technologies Co. Ltd.
[3]S-Lab, Nanyang Technological University
{xjin, lhhan}@mail.nankai.edu.cn, zhenli1031@gmail.com, guochunle@nankai.edu.cn,
chaizhi2@huawei.com, chongyi.li@ntu.edu.sg
https://github.com/Srameo/DNF

Abstract

The exclusive properties of RAW data have shown great potential for low-light image enhancement. Nevertheless, the performance is bottlenecked by the inherent limitations of existing architectures in both single-stage and multi-stage methods. Mixed mapping across two different domains, noise-to-clean and RAW-to-sRGB, misleads the single-stage methods due to the domain ambiguity. The multi-stage methods propagate the information merely through the resulting image of each stage, neglecting the abundant features in the lossy image-level dataflow. In this paper, we probe a generalized solution to these bottlenecks and propose a ***D****ecouple a****N****d* ***F****eedback framework, abbreviated as* ***DNF****. To mitigate the domain ambiguity, domain-specific subtasks are decoupled, along with fully utilizing the unique properties in RAW and sRGB domains. The feature propagation across stages with a feedback mechanism avoids the information loss caused by image-level dataflow. The two key insights of our method resolve the inherent limitations of RAW data-based low-light image enhancement satisfactorily, empowering our method to outperform the previous state-of-the-art method by a large margin with only 19% parameters, achieving 0.97dB and 1.30dB PSNR improvements on the Sony and Fuji subsets of SID.*

1. Introduction

Imaging in low-light scenarios attracts increasing attention, especially with the popularity of the night sight mode on smartphones and surveillance systems. However, low-light image enhancement (LLIE) is a challenging task due to the exceptionally low signal-to-noise ratio. Recently, deep learning solutions have been widely studied to tackle this task in diverse data domains, ranging from sRGB-based methods [14,15,21,40] to RAW-based methods [2,7,35,47]. Compared with sRGB data, RAW data with the unprocessed signal is unique in three aspects that benefit LLIE: 1) the signal is linearly correlated with the photon counts in the RAW domain, 2) the noise distributions on RAW images are tractable before the image signal processing (ISP) pipeline [33], and 3) the higher bit depth of RAW format records more distinguishable low-intensity signals.

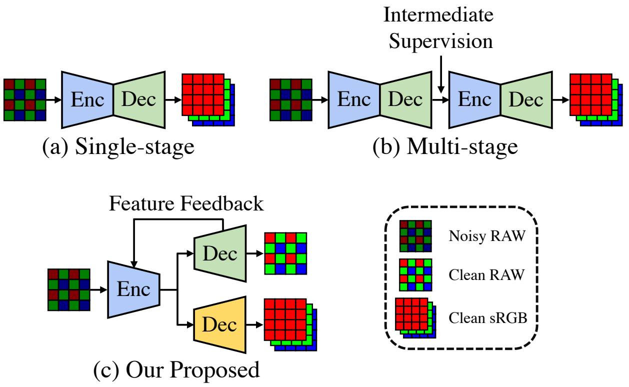


Figure 1. Thumbnail of different RAW-based low-light image enhancement methods. (a) Single-stage method. (b) Multi-stage method with intermediate supervision. (c) The proposed DNF.

The pioneering work SID [2] proposed a large-scale paired dataset for RAW-based LLIE, igniting a renewed interest in data-driven approaches. As shown in Fig. 1, one line of work [2, 5, 12, 13, 22, 42] focuses on designing single-stage network architectures, and another [4, 7, 35, 47] exploits the multi-stage networks for progressive enhancement. Despite the great performance improvement, both architectures are still bottlenecked by inherent limitations. First, current single-stage methods force neural networks to learn a direct mapping from noisy RAW domain to clean sRGB domain. The mixed mapping across two different domains, noisy-to-clean and RAW-to-sRGB, would mislead the holistic enhancement process, leading to the *domain ambiguity* issue. For example, the tractable noise in RAW images would be mapped to an unpredictable distribution during color space transformation. Therefore, shifting colors and unprocessed noises inevitably appear in the

[*] Equal contribution.
[†] C. L. Guo is the corresponding author.

final results. Second, existing multi-stage methods compose the pipeline by cascaded subnetworks, each of which is responsible for gradual enhancement based on the output image of the last stage. Under their designs with the image-level dataflow, only images are propagated forward across multiple stages, and the later stage only obtains information from the result of the previous stage. Meanwhile, every subnetwork in each stage may incur information loss due to the downsampling operation or the separate objective function [41]. Consequently, the suboptimal performance is bound up with the *lossy image-level dataflow*. The error is propagated, accumulated, and magnified along with stages, finally failing to reconstruct the texture details.

To exploit the potential of RAW images for LLIE, a generalized pipeline is expected, which transcends the above two limitations. Specifically, neural networks ought to utilize the aforementioned merits in different domains [7], rather than being confused by the domain ambiguity. According to the unique properties of the RAW and sRGB domains, it is essential to decouple the enhancement into domain-specific subtasks. After exploring the linearity and tractable noise in the RAW domain, the color space transformation from RAW domain to sRGB domain can be performed deliberately without noise interference. Besides, the pipeline cannot hinder communication across stages, instead of the image-level dataflow that merely allows a small portion of lossy information to pass through. Due to the diverse subtasks, the intermediate feature of each level tends to be complementary to each other [20, 46]. Meanwhile, multi-scale features preserve texture and context information, providing additional guidance for later stages [41]. Hence, the features in different stages are required to propagate across dataflow, aggregating the enriched features and sustaining the intact information. The domain-specific decoupling, together with the feature-level dataflow, facilitates the learnability for better enhancement performance and retains the method's interpretability.

Based on these principles, we propose a **D**ecouple a**N**d **F**eedback (**DNF**) framework, with the following designs tailored for RAW-based LLIE. The enhancement process is decoupled into two domain-specific subtasks: denoising in the RAW domain [30, 33, 45, 48] and the color restoration into the sRGB domain [8, 28, 39], as shown in Fig. 1(c). Under the encoder-decoder architecture commonly used in previous works [27], each module in the subnetwork is derived from the exclusive properties of each domain: the Channel Independent Denoising (CID) block for RAW denoising, and the Matrixed Color Correction (MCC) block for color rendering. Besides, instead of using the inaccurate denoised RAW image, we resort to the multi-scale features from the RAW decoder as denoising prior. Then, the features are flowed into the shared RAW encoder by proposed Gated Fusion Modules (GFM), adaptively distinguishing the texture details and remaining noise. After the Denoising Prior Feedback, signals are further distinguished from noises, yielding intact and enriched features in the RAW domain. Benefiting from the feature-level dataflow, a decoder of MCC blocks could efficiently deal with the remaining enhancement and color transformation to sRGB domain.

The main contributions are summarized as follows:

- The domain-specific task decoupling extends the utilization of the unique properties in both RAW and sRGB domains, avoiding domain ambiguity.
- The feature-level dataflow empowered by the Denoising Prior Feedback reduces the error accumulation and aggregates complementary features across stages.
- Compared with the previous state-of-the-art method, the proposed method gains a significant margin improvement with only 19% parameters and 63% FLOPs, *e.g.* 0.97dB PSNR improvement on the Sony dataset of SID and 1.30dB PSNR improvement on the Fuji dataset of SID.

2. Related Work

2.1. RAW-based Low-Light Image Enhancement

RAW images have been widely explored for image enhancement under extremely low-light conditions, owing to their unique properties, as we mentioned in Sec. 1. As shown in Fig. 1, RAW-based methods generally involve two categories by whether there is intermediate supervision: single-stage and multi-stage. **Single-stage** methods [2, 5, 12, 13, 22, 42] intend to force the deep neural network to learn a direct mapping from noisy RAW domain to clean sRGB domain. Multiple attempts have been adopted for better performance, including similarity and perceptual loss [42], residual learning [22], multi-scale features [5], and lightweight [12, 13]. However, the above single-stage methods often fail to recover texture details due to the domain ambiguity. **Multi-stage** methods [4, 7, 35, 47] are raised for solving the limitation of the single-stage method. With intermediate supervision on the sRGB domain, EEMEFN [47] and LDC [35] reconstructed detail in the second stage. Intermediate supervision on different domains is used for different purposes, *e.g.* RAW domain for decoupling [7], monochrome domain for low-light information complementation [4]. However, all the existing multi-stage methods share the same architecture, *i.e.* cascaded encoder-decoders [27]. Their image-level dataflow induces an error accumulation across stages. Our proposed method differs from existing methods in two aspects. 1) A domain-specific decoupled architecture is employed for fully utilizing the properties of RAW and sRGB format. 2) A feature-level feedback architecture is employed to handle the error accumulation of the image-level dataflow.

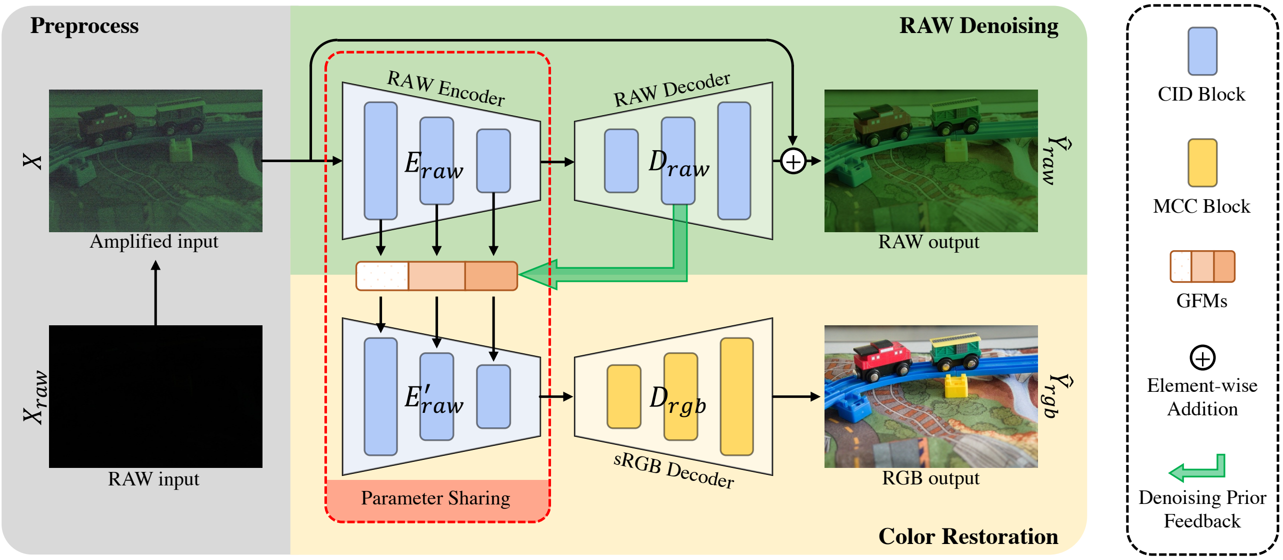


Figure 2. Overview of the proposed DNF. Our DNF contains three main components: 1) an RAW encoder and an RAW decoder that performs the denoising in the RAW domain with auxiliary supervision on the RAW image output, 2) gated fusion modules (GFMs) that handle the feedback features, and 3) an sRGB decoder that performs color space transformation and outputs the final result.

2.2. Decouple Mechanism

Decouple mechanism aims to divide the original task into several simpler subtasks, then conquer them explicitly. With proper decoupling, neural network can easier converge, resulting in better performance. Li *et al.* [18] decoupled the extrapolation task into bounding-box layout generation, segmentation layout generation, and image generation. Reasonable decoupling reduced the domain gap between text and image by two footholds, bounding-box and segmentation layout. Recent works on inpainting [16, 25, 26] aim to decouple the inpainting task into structure and texture reconstruction for a better result. In high-level task [11], decoupling the domain adaptation task into feature distribution alignment and segmentation gains a performance improvement. We implement this mechanism through domain-specific task decoupling, which decouples the LLIE task into RAW denoising and color restoration.

2.3. Feedback Mechanism

The feedback mechanism enables the network to access the grasp from previous states. This idea has been applied in many tasks, *e.g*. classification [37], super-resolution [17, 19], and point cloud completion [36]. With the feedback mechanism involved, Li *et al.* [19] employed a curriculum learning strategy for gradual restoration. Yan *et al.* [36] intended to enrich the low-resolution features with the high-resolution ones using feedback mechanism. All the existing methods applied the feedback mechanism for progressively fulfilling a sole task, which differs from our method. Our feedback mechanism enables our network to communicate between two different subtasks, also in diverse domains.

3. Methodology

As shown in Fig. 2, the proposed decouple and feedback framework consists of two stages, RAW denoising and color rendering, to progressively enhance the low-light RAW images. Given an input image X_{raw}, after multiplying the pre-defined amplification ratio [2], the amplified image X is first denoised by the encoder E_{raw} and decoder D_{raw} in the RAW domain. Then, instead of using the inaccurate $\hat{Y}_{raw}$ for color rendering, we feed the denoising features $\mathbf{F}_{dn}$ from D_{raw} back to E'_{raw}, further distinguishing signals with denoising priors, and composing enriched features in RAW domain. Finally, the sRGB decoder D_{rgb} takes the multi-scale features in RAW domain to render the final output $\hat{Y}_{rgb}$ in the sRGB domain.

Specifically, a shared encoder E_{raw} and two decoders (D_{raw} and D_{rgb}) are specially designed for the subtasks decoupled by *Domain-Specific Task Decoupling* with task-specific blocks (Sec. 3.1). The Channel Independent Denoising (CID) block is introduced to learn the tractable and independent noise distribution in different color channels in the RAW domain. In accordance with the definition of the color space, a Matrixed Color Correction (MCC) block accomplishes the remaining enhancement into the sRGB domain using the global matrix transformation. Besides, we incorporate a *Denoising Prior Feedback* mechanism to avoid error accumulation across stages. With the denoising features $\mathbf{F}_{dn}$ extracted from the RAW decoder, the RAW encoder enriches the shallow features with high-frequency information. Furthermore, a Gated Fusion Module (GFM) is proposed with gated mechanism [17] for adaptively exploring the details buried in the noise (Sec. 3.2).

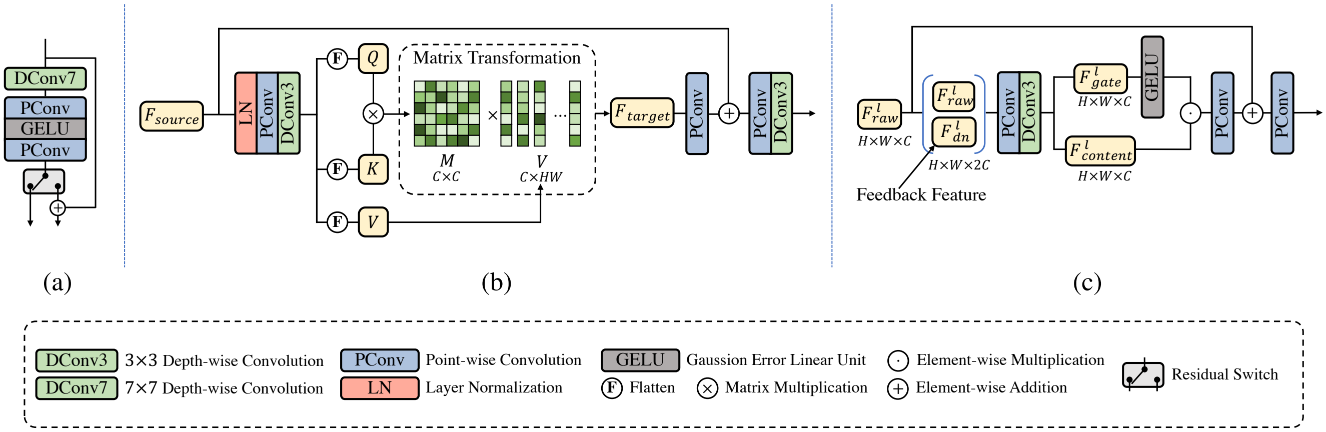


Figure 3. Detailed architectures of proposed task-specific blocks and the fusion module: (a) Channel-Independent Denoising (CID) Block, (b) Matrixed Color Correction (MCC) Block, and (c) Gated Fusion Module (GFM).

3.1. Domain-Specific Task Decoupling

We propose Domain-Specific Task Decoupling for handling the domain ambiguity caused by the mixed mapping of noisy-to-clean and RAW-to-sRGB. The chasm between noisy RAW and clean sRGB domain makes it difficult for the network to learn the direct mapping across the two domains. Thus, we propose to involve intermediate supervision on clean RAW domain between the chasm for easing the difficulty in directly learning the mixed mapping. By the intermediate supervision on clean RAW domain, we can: 1) decouple the enhancement into RAW denoising and color restoration, 2) fully utilize the property, that is, noise distribution is tractable on RAW images for denoising, and 3) reduce the noise interference during color restoration, resulting in less color shifting.

Denoising in RAW Domain. As shown in Fig. 2, we stack multiple channel-independent denoising (CID) blocks to implement the RAW encoder E_{raw} and the RAW decoder D_{raw}. The design of a CID block is based on the following two prior knowledge: 1) low-light images with RAW format suffer from the signal-independent noise which follows a zero-mean distribution [9, 33], and 2) the noise distribution tends to be independent across channels as the signals of different channels are inherently less correlated in RAW domain [24, 34]. Thus, we need burst observations on nearly the same signals (adjacent pixels) to remove the interference of the zero-mean noise. Also, preventing channel-wise information exchange during denoising is indispensable for handling channel-independent noise distribution. According to the above discussions, we introduce depth-wise convolution with a large kernel for denoising in the CID block. The detailed structure of the CID block is shown in Fig. 3 (a). To be specific, for the input feature F_{in}, the output feature F_{out} after the channel-independent denoising block can be formulated as:

$$F_{out} = \mathtt{MLP}(\mathtt{DConv7}(F_{in})) + F_{in}, \tag{1}$$

where $\mathtt{DConv7}$ is a depth-wise convolution with 7×7 kernels. $\mathtt{MLP}$ is implemented by two point-wise convolutional layers and a GELU [6] non-linearity function. Also, a residual switch is set to perform two different functionalities with a weight-sharing CID block, detailed in Sec. 3.2.

Color Correction for RAW-to-sRGB. Matrix transformation is commonly employed in the canonical ISP pipelines [23]. Due to the globally shared settings, such as environmental illumination and color space specifications, the colors of an image are mainly enhanced or converted to another color space through a channel-wise matrix transformation. Following this principle, we introduce a Matrixed Color Correction (MCC) block to perform global color enhancement as well as local refinement, as shown in Fig. 3 (b). For the sRGB decoder D_{rgb}, we stack multiple MCC blocks for color correction. The design of this block benefits from recent advances in transposed self-attention [38]. The global receptive field and channel-wise operation of it fits color correction in canonical ISP well. Given the input source feature $F_{source} \in \mathbb{R}^{C\times H\times W}$, the vectors of query $Q \in \mathbb{R}^{C\times HW}$, key $K \in \mathbb{R}^{C\times HW}$, and value $V \in \mathbb{R}^{C\times HW}$ are first generated through the projection with a 1×1 convolutional layer followed by a 3×3 depth-wise one and a flatten operation. Then, the transformation matrix $M \in \mathbb{R}^{C\times C}$ is obtained by matrix multiplication. This procedure can be formulated as:

$$Q, K, V = \mathtt{Flatten}(\mathtt{DConv3}(\mathtt{PConv}(F_{source}))), \tag{2}$$

$$M = \mathtt{Softmax}(Q \cdot K^T/\lambda), \tag{3}$$

where a scaling coefficient λ is applied for numerical stability. Then, the color vector V is transformed by the matrix M, performing color space conversion in feature-level. The target feature after color transformation can be obtained by $F_{target} = M \cdot V$. As a complement to the global matrix transformation, we use a depth-wise convolution and a point-wise convolution to refine the local details further.

3.2. Denoising Prior Feedback

In previous RAW-based methods [4, 7, 35, 47], a portion of high-frequency content is erroneously identified as noises in the process of enhancement, severely deteriorating the final results with detail loss and resulting in a lossy dataflow. To avoid the lossy image-level dataflow of existing multi-stage methods, we propose a Denoising Prior Feedback mechanism with feature-level information propagating. We denote $\mathbf{F}_{dn} = \{F^1_{dn}, F^2_{dn}, ..., F^L_{dn}\}$ as a set of denoising features extracted from the RAW decoder D_{raw}, where L denotes the number of stages. Each element of $\mathbf{F}_{dn}$ mainly contains the information of the final noise estimation at different scales in the RAW domain. Specifically, these features make noises more distinguished and serve as a guidance for further denoising. Through rerouting the set of denoising features $\mathbf{F}_{dn}$ to the corresponding stages of the RAW encoder with multiple feedback connections [1, 19, 29], the encoder gradually generates better denoising features with the last estimation for further enhancement. Thus, the sRGB decoder D_{rgb} can concentrate more on color correction. The feedback pipeline is shown in Fig. 2 and can be formulated as:

$$\mathbf{F}_{dn} = D_{raw}(E_{raw}(X)), \quad F_{rdn} = E'_{raw}(X, \mathbf{F}_{dn}), \quad (4)$$

where F_{rdn} denotes the refined denoising feature that will be forwarded to the sRGB decoder. E'_{raw} denotes the RAW encoder that not only contains the weights of E_{raw} but is equipped with L gated fusion modules (GFMs). Each GFM is responsible for handling one feedback feature from $\mathbf{F}_{dn}$.

Gated Fusion Modules. The GFM is designed to adaptively fuse the feedback noise estimation with initial denoising features with a gated mechanism [17]. During feature gating, we expect that the helpful information is adaptively selected and merged along both spatial and channel dimensions. For efficiency, we use a point-wise convolution and a depth-wise convolution [3] to aggregate the channel and local content information, respectively. Then, we split the mixed feature along the channel dimension into two chunks, *i.e.*, F^l_{gate} and F^l_{con}. After activated by a GELU non-linearity function, F^l_{gate} gates F^l_{con} through point-wise multiplication. We achieve both spatial and channel adaptability by this gating mechanism. The detailed structure of the GFM is shown in Fig. 3 (c). The operations at the l-th ($l \in \{1, 2, ..., L\}$) stage can be formulated as:

$$F^l_{gate}, F^l_{con} = \texttt{DConv3}(\texttt{PConv}([F^l_{raw}, F^l_{dn}])), \quad (5)$$

$$F^l_{fuse} = \texttt{PConv}(F^l_{con} \odot \texttt{GELU}(F^l_{gate})) + F^l_{raw}, \quad (6)$$

where DConv3 and PConv represent depth-wise convolution with the kernel of 3×3 and a point-wise convolution, respectively. $\odot$ denotes the hadamard product. F^l_{raw} is the feature obtained after the l-th upscaling layer in the original RAW encoder. F^l_f is the corresponding fused feature. A point-wise convolution performs channel mixing on this fused feature. The mixed feature is fed to the next CID block in the RAW encoder for further refinement.

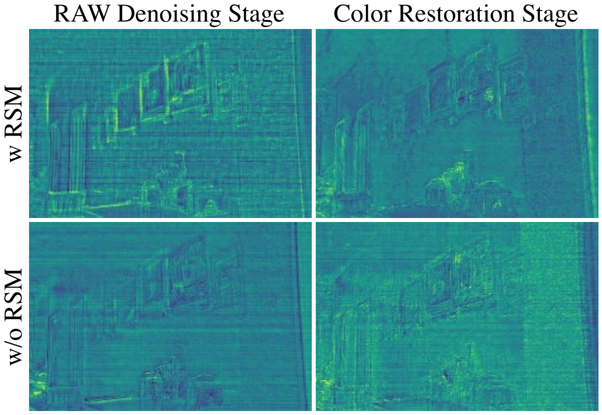


Figure 4. Feature visualization of the shared RAW encoder with or without RSM (*Zoom-in for best view*).

Residual Switch Mechanism. We only keep the global shortcut at the denoising stage in RAW domain for better denoising [22, 43], while removing it at the color restoration stage to avoid the ambiguous connection between noisy RAW domain and clean sRGB domain, as shown in Fig. 2. Thus, the encoder is required to perform noise estimation when denoising, however, to reconstruct the signal during color restoration. Toward the two contradicting functionalities in a single encoder, we propose a simple yet effective Residual Switch Mechanism (RSM), as shown in Fig. 3 (a), empowering the CID blocks in the shared RAW encoder to yield two contradict features: noise and signal. At the denoising stage with global residual connection, the local residual shortcuts are switched off to estimate the noise. On the contrary, the local residuals are triggered on at the rendering stage, counteracting the noise with the original feature on the shortcut, and finally reconstructing the signal. As shown in Fig. 4, the CID block of the shared RAW encoder is able to yield two different features at different stages with RSM. However, without RSM, the weight sharing CID block failed to distinguish noise and signal at the color restoration stage, resulting in ambiguous features. The remaining noise bottlenecks the color correction procedure and introduce the domain ambiguity again.

3.3. Training Objectives

To sequentially fulfill the RAW denoising and color restoration subtasks decoupled by the domain-specific task decoupling, we introduce two different supervision on different domains, *i.e.* clean RAW and clean sRGB. The ground truth is the clear RAW image Y_{raw}. We denote the output RAW image of the denoising decoder as $\hat{Y}_{raw}$. The loss function of our network is:

$$L = \left\| Y_{raw} - \hat{Y}_{raw} \right\|_1 + \left\| Y_{rgb} - \hat{Y}_{rgb} \right\|_1, \quad (7)$$

Table 1. Quantitative results of RAW-based LLIE methods on the Sony and Fuji subsets of SID [2]. The best result is in **bold** whereas the second best one is in underlined. Metrics with ↑ and ↓ denote higher better and lower better, respectively. Methods with * indicate that the model is trained and inference with a downsampled resolution, and we manually upsample the results to the original resolution during testing. Methods with $^{\#}$ indicate that the model is trained and inferenced with only the images of small digital gains ($\times$100) on the SID datasets. "-" indicates the result is not available.

Category	Method	Params.	FLOPs	Sony			Fuji		
				PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Single-Stage	SID [2]	7.7 M	48.5 G	28.96	0.787	0.356	26.66	0.709	0.432
	DID [22]	2.5 M	669.2 G	29.16	0.785	0.368	-	-	-
	SGN [5]	19.2 M	75.5 G	29.28	0.790	0.370	27.41	0.720	0.430
	LLPackNet [12]	1.2 M	7.2 G	27.83	0.755	0.541	-	-	-
	RRT [13]	0.8 M	5.2 G	28.66	0.790	0.397	26.94	0.712	0.446
Multi-Stage	EEMEFN [47]	40.7 M	715.6 G	29.60	0.795	0.350	27.38	0.723	0.414
	LDC* [35]	8.6 M	124.1 G	29.56	**0.799**	0.359	27.18	0.703	0.446
	MCR$^{\#}$ [4]	15.0 M	90.5G	29.65	0.797	0.348	-	-	-
	RRENet [7]	15.5 M	96.8 G	29.17	0.792	0.360	27.29	0.720	0.421
	Ours	2.8 M	57.0 G	**30.62**	0.797	**0.343**	**28.71**	**0.726**	**0.391**

where Y_{rgb} is the ground truth sRGB image. It is worth noticing that only L_1 loss is employed for both RAW supervision and sRGB supervision in our method, instead of blending complex loss functions like previous methods [7, 30, 32, 42, 47]. Training details and detailed network architectures can be found in the supplementary material.

4. Experiments and Analysis

4.1. Datasets and Evaluation Metrics

We have benchmarked our proposed DNF on two different RAW-based LLIE datasets, *i.e.* the See-In-the-Dark (SID) [2] dataset and Mono-Colored Raw Paired (MCR) [4] dataset. The SID [2] dataset contains 5094 extremely low-light RAW images with corresponding normal-light reference images taken by two cameras: Sony A7S2 with Bayer sensor and a resolution of 4240 × 2832, and Fuji X-T2 with X-Trans sensor and a resolution of 6000 × 4000. The exposure time of the low-light image varies from 0.1s to 0.033s, and the reference images are captured 100 to 300 times longer than the exposure time of the low-light images. Noted that the long-short pairs of three scenes are misaligned in the test set of Sony subsets, so we discard these images during the testing stage following previous methods [22, 47]. For fair comparisons, all the compared methods are evaluated under the same settings. The MCR [4] dataset contains 4980 images with a resolution of 1280 × 1024 for training and testing, including 3984 low-light RAW images, 498 monochrome images, and 498 sRGB images. With two different kinds of scenes, indoor and outdoor, different exposure times are set, 1/256s to 3/8s for indoor scenes and 1/4096s to 1/32s for outdoor scenes. However, no ground truth in RAW format is provided, which is indispensable for training our method. Thus, we select images with the longest exposure time of each scene as the RAW ground truth. Also, the monochrome images are not taken into account in our DNF. We regard PSNR, SSIM [31], and LPIPS [44] as the quantitative evaluation metrics for pixel-wise, structural, and perceptual assessment, respectively.

Table 2. Quantitative results of RAW-based LLIE methods on the MCR dataset [4]. The best result is in **bold** whereas the second best one is in underlined. Metrics with ↑ and ↓ denote higher better and lower better, respectively.

Category	Method	PSNR↑	SSIM↓
Single-Stage	RRT [13]	25.74	0.851
	SGN [5]	26.29	0.882
	DID [22]	26.16	0.888
	SID [2]	29.00	0.906
Multi-Stage	LDC [35]	29.36	0.904
	MCR [4]	31.69	0.908
	Ours	**32.00**	**0.915**

4.2. Comparison with State-of-the-Art Methods

We evaluate our DNF on two subsets, Sony and Fuji, of the SID [2] and MCR [4] datasets, and compare it with state-of-the-art RAW-based LLIE methods, including the single-stage methods, SID [2], DID [22], SGN [5], LLPackNet [12], and RRT [13], as well as the multi-stage methods, EEMEFN [47], LDC [35], RRENet [7], and MCR [4].

Quantitative Evaluation. As shown in Tab. 1 and Tab. 2, our method outperforms the previous state-of-the-art method by a large margin. On the SID dataset, our DNF yields the best PSNR and LPIPS scores, achieving

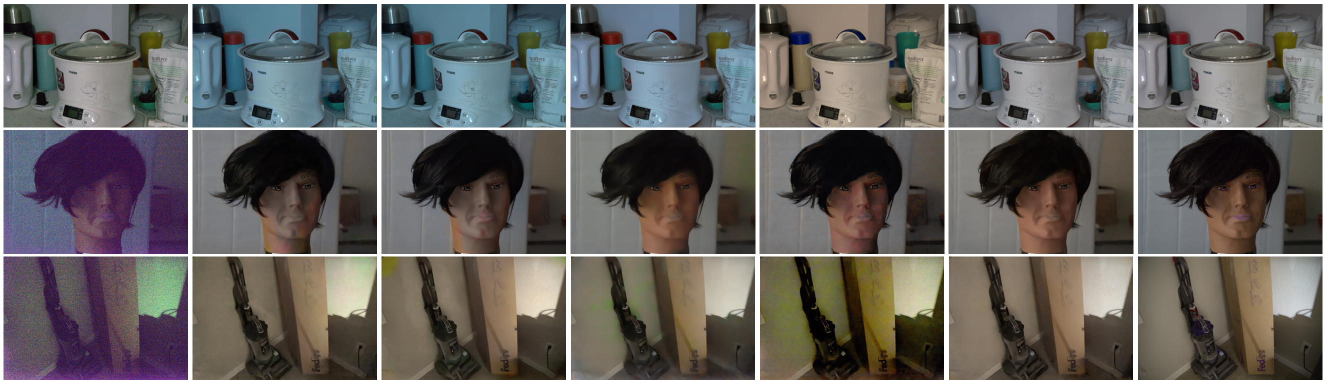


Figure 5. Visual comparisons between our DNF and the state-of-the-art methods (*Zoom-in for best view*). We amplified and post-processed the input images with an ISP for visualization [2].

0.97 dB and 1.30 dB improvements in PSNR, as well as 0.005 and 0.023 improvements in LPIPS than the second-best method on the Sony and Fuji subsets, respectively. Note that LDC [35] and MCR [4] are trained and tested in different schemes [1] which might lead to a better performance. Regarding the complexity, our DNF has notably fewer parameters and FLOPs than the current best methods (*i.e.*, MCR and EEMEFN). Our network uses 1/5 and 1/15 fewer parameters, as well as 3/5 and 1/13 fewer FLOPs than MCR and EEMEFN, respectively. On the MCR dataset, our method achieves the best PSNR and SSIM scores as shown in Tab. 2, exceeding the previous state-of-the-art method 0.31dB and 0.07 with fewer parameters and FLOPs.

Qualitative Evaluation. Fig. 5 and Fig. 6 show the qualitative results on the SID [2] dataset. It can be seen that the results enhanced by the compared methods suffer from severe content distortions and artifacts due to their limited denoising capability. In addition, benefiting from the decouple and feedback architecture, colors are transformed and enhanced more accurately without noise interference, therefore exhibiting better color consistency together with more realistic and vivid color rendering. Our method succeeds in suppressing the intensive noises, as well as preserving rich texture details. Qualitative comparisons on the MCR dataset can be found in the supplementary material.

4.3. Ablation Studies

We conduct extensive ablation studies on proposed DNF. All experiments are performed on the SID [2] Sony subset.

Domain-Specific Task Decoupling. To better evaluate the impact of our Domain-Specific Task Decoupling, we involve different kinds of intermediate supervision on our denoising decoder, as shown in Tab. 3. 1) Without supervision (w/o Sup.) fails to feedback the denoising prior back to the shared RAW encoder with a 0.14dB drop on PSNR. 2) sRGB supervision (sRGB Sup.) decouples the main task into first-stage enhancement and detail reconstruction, like [35, 47]. The first-stage enhancement suffers from the domain ambiguity caused by directly learning from noisy RAW to clean sRGB domain, resulting in a 0.42dB drop on PSNR. The comparison between the sRGB Sup. and w/o Sup. (0.28dB↓) denotes that the domain ambiguity severely bottlenecks the performance of the network.

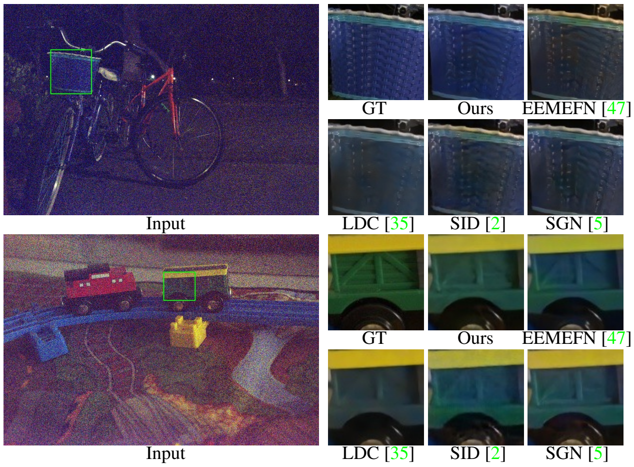


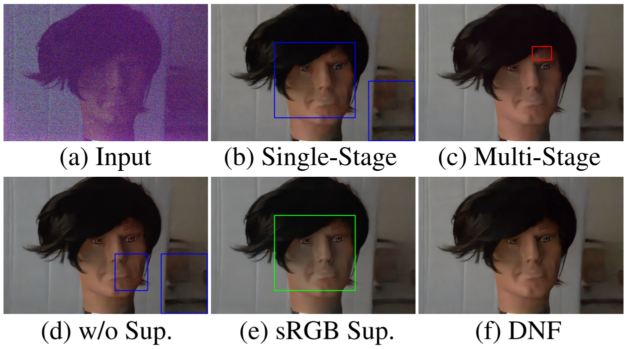
Figure 6. Visual comparisons between our DNF and the state-of-the-art methods. In comparison to the state-of-the-art methods, our method achieves better texture preservation and color recovery.

Denoising Prior Feedback. To validate the effectiveness of our framework based on the feedback mechanism, we first examine the single-stage and multi-stage (two-stage like most of the existing methods [4, 7, 35, 47]) variants of our framework in Tab. 3. 1) Single-stage variant by directly cascading RAW encoder and sRGB decoder results in 0.46dB drop on PSNR. 2) Multi-stage variant simply cascades two UNets [27] equipped with CID blocks and MCC blocks, respectively. The lossy image-level dataflow deteriorates the performance severely, with a 0.30dB drop on PSNR.

[1] The LDC is trained and tested on downsampled images, as well as the MCR is trained and tested only on the images with small digital gains (×100). We keep the settings exactly the same with their implementations.

Table 3. Ablation study on the deouple and feedback framework. Sup. denotes the supervision of the denoising decoder.

Module	Replacement	PSNR	SSIM
RAW Sup.	w/o Sup.	30.48	0.795
	sRGB Sup.	30.20	0.796
Feedback	Single-Stage	30.16	0.792
	Mulit-Stage	30.32	0.795
GFM	Conv	30.40	0.795
	w/o Gate	30.35	0.794
	SKFF [40]	30.37	0.795
Original		**30.62**	**0.797**



(a) Input (b) Single-Stage (c) Multi-Stage

(d) w/o Sup. (e) sRGB Sup. (f) DNF

Figure 7. Visual comparisons between our DNF and ablated models (*Zoom-in for best view*). blue, red, green boxes represent remaining noise, detail loss, and color shifts, respectively.

Gated Fusion Module. Three other fusion modules are adopted: traditional convolutional layer (0.22dB↓), depth-wise convolution without a gating mechanism (0.27dB↓), and SKFF [10, 40] module (0.25dB↓), specialized for feature fusion. Our method enjoys the best performance due to the pixel-wise selection provided by the gated mechanism.

Residual Switch Mechanism. As shown in Tab. 4, comparing without any residual shortcuts at all, leveraging a global residual shortcut increases the performance (0.06dB↑). However, the global shortcut at the color restoration stage would limit the performance by introducing the domain ambiguity (0.03dB↓). The experiments with or without all of the local shortcuts introduce functional contradiction, thus resulting in varying degrees of performance degradation (0.33dB↓ and 0.30dB↓, comparing with DNF). Compared with another approach to implement the residual switch mechanism: switch on when denoising or switch off during color restoration, the local shortcut of the CID block during the color restoration provides more information about the image content, thus resulting in higher performance.

Comparison with Other Feature-level Dataflow. As shown in Tab. 5, our model yields the best performance compared with a feature-level dataflow multi-stage framework, which validates the effectiveness of residual switch mechanism (RSM). The feature-level multi-stage framework preserves the gated fusion modules but involves two different RAW encoders. The results show that a weight-sharing encoder can perform two different functionalities with our proposed RSM. Also, the two complementary functionalities, noise estimation and signal reconstruction, complement each other for achieving a better performance.

Table 4. Ablation study on the residual mechanism of CID block. Global and Local represent the global residual shortcut between encoder and decoder and the local shortcut in the CID block, respectively. RAW and sRGB represent the RAW denoising stage and color restoration stage, respectively. The last row represents the implementation and performance of our proposed DNF.

Global		Local		PSNR	SSIM
RAW	sRGB	RAW	sRGB		
				30.26	0.794
✓				30.32	0.795
✓	✓			30.29	0.794
✓		✓		30.48	0.794
✓		✓	✓	30.29	0.794
✓			✓	**30.62**	**0.797**

Table 5. Comparison with other feature-level dataflow. Multi-Stage* represents a feature-level multi-stage framework.

Method	w/o RSM	DNF	Multi-Stage*
PSNR	30.32	**30.62**	30.46
SSIM	0.794	**0.797**	0.796

5. Conclusion

In light of the exclusive properties in RAW format, we propose a decouple and feedback framework for the RAW-based LLIE. As a generalized pipeline, the proposed DNF overcomes the inherent limitations of previous methods. The Domain-Specific Task Decoupling eliminates the domain ambiguity incurred by the single-stage methods, and the Denoising Prior Feedback supersedes the multi-stage methods that are with the lossy image-level dataflow. Significant performance and extensive experiments show the superiority of the proposed framework as well as the great potential for low-light image enhancement of RAW images.

Limitations. One remaining limitation in the proposed framework, also shared with most of the existing methods, is that the amplification ratios of the input images are pre-defined according to the exposure time. Under the extremely low-light condition, estimating the normal illumination is essential and difficult in the real-world scenarios.

Acknowledgement. This work is funded by the National Key Research and Development Program of China (NO. 2018AAA0100400), NSFC (NO. 62225604), China Postdoctoral Science Foundation (NO.2021M701780). We are also sponsored by CAAI-Huawei MindSpore Open Fund.

References

[1] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *ICCV*, 2015. 5

[2] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018. 1, 2, 3, 6, 7

[3] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017. 5

[4] Xingbo Dong, Wanyan Xu, Zhihui Miao, Lan Ma, Chao Zhang, Jiewen Yang, Zhe Jin, Andrew Beng Jin Teoh, and Jiajun Shen. Abandoning the bayer-filter to see in the dark. In *CVPR*, 2022. 1, 2, 5, 6, 7

[5] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *ICCV*, 2019. 1, 2, 6, 7

[6] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv:1606.08415*, 2016. 4

[7] Haofeng Huang, Wenhan Yang, Yueyu Hu, Jiaying Liu, and Ling-Yu Duan. Towards low light enhancement with raw images. *IEEE TIP*, 2022. 1, 2, 5, 6, 7

[8] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *CVPR Workshops*, 2020. 2

[9] Geonwoon Jang, Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. C2n: Practical generative noise modeling for real-world denoising. In *ICCV*, 2021. 4

[10] Aupendu Kar, Sobhan Kanti Dhara, Debashis Sen, and Prabir Kumar Biswas. Zero-shot single image restoration through controlled perturbation of koschmieder's model. In *CVPR*, 2021. 8

[11] Xin Lai, Zhuotao Tian, Xiaogang Xu, Yingcong Chen, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Decouplenet: Decoupled network for domain adaptive semantic segmentation. *arXiv:2207.09988*, 2022. 3

[12] Mohit Lamba, Atul Balaji, and Kaushik Mitra. Towards fast and light-weight restoration of dark images. *arXiv:2011.14133*, 2020. 1, 2, 6

[13] Mohit Lamba and Kaushik Mitra. Restoring extremely dark images in real time. In *CVPR*, 2021. 1, 2, 6

[14] Chongyi Li, Chunle Guo, and Change Loy Chen. Learning to enhance low-light image via zero-reference deep curve estimation. *TPAMI*, 2021. 1

[15] Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Mingming Cheng, Jinwei Gu, and Chen Change Loy. Low-Light Image and Video Enhancement Using Deep Learning: A Survey. *TPAMI*, 2021. 1

[16] Jingyuan Li, Fengxiang He, Lefei Zhang, Bo Du, and Dacheng Tao. Progressive reconstruction of visual structure for image inpainting. In *ICCV*, 2019. 3

[17] Qilei Li, Zhen Li, Lu Lu, Gwanggil Jeon, Kai Liu, and Xiaomin Yang. Gated multiple feedback network for image super-resolution. *BMVC*, 2019. 3, 5

[18] Yijun Li, Lu Jiang, and Ming-Hsuan Yang. Controllable and progressive image extrapolation. In *WACV*, 2021. 3

[19] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *CVPR*, 2019. 3, 5

[20] Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Wassell. Multi-task adversarial network for disentangled feature learning. In *CVPR*, 2018. 2

[21] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *BMVC*, 2018. 1

[22] Paras Maharjan, Li Li, Zhu Li, Ning Xu, Chongyang Ma, and Yue Li. Improving extreme low-light image denoising via residual learning. In *ICME*, 2019. 1, 2, 5, 6

[23] Junichi Nakamura. *Image sensors and signal processing for digital still cameras*. CRC press, 2017. 4

[24] Seonghyeon Nam, Youngbae Hwang, Yasuyuki Matsushita, and Seon Joo Kim. A holistic approach to cross-channel image noise modeling and its application to image denoising. In *CVPR*, 2016. 4

[25] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *ICCV Workshops*, 2019. 3

[26] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *ICCV*, 2019. 3

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *ICML*, 2015. 2, 7

[28] Eli Schwartz, Raja Giryes, and Alex M Bronstein. Deepisp: Toward learning an end-to-end image processing pipeline. *IEEE TIP*, 2018. 2

[29] Abhinav Shrivastava, Rahul Sukthankar, Jitendra Malik, and Abhinav Gupta. Beyond skip connections: Top-down modulation for object detection. *arXiv:1612.06851*, 2016. 5

[30] Yuzhi Wang, Haibin Huang, Qin Xu, Jiaming Liu, Yiqun Liu, and Jue Wang. Practical deep raw image denoising on mobile devices. In *ECCV*, 2020. 2, 6

[31] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 6

[32] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *CVPR*, 2020. 6

[33] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *TPAMI*, 2021. 1, 2, 4

[34] Jun Xu, Lei Zhang, David Zhang, and Xiangchu Feng. Multi-channel weighted nuclear norm minimization for real color image denoising. In *ICCV*, 2017. 4

[35] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *CVPR*, 2020. 1, 2, 5, 6, 7

[36] Xuejun Yan, Hongyu Yan, Jingjing Wang, Hang Du, Zhihong Wu, Di Xie, Shiliang Pu, and Li Lu. Fbnet: Feedback network for point cloud completion. In *ECCV*, 2022. 3

[37] Amir R Zamir, Te-Lin Wu, Lin Sun, William B Shen, Bertram E Shi, Jitendra Malik, and Silvio Savarese. Feedback networks. In *CVPR*, 2017. 3

[38] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *arXiv:2111.09881*, 2021. 4

[39] 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 *CVPR*, 2020. 2

[40] 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 *ECCV*, 2020. 1, 8

[41] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 2

[42] Syed Waqas Zamir, Aditya Arora, Salman Khan, Fahad Shahbaz Khan, and Ling Shao. Learning digital camera pipeline for extreme low-light imaging. *Neurocomputing*, 2021. 1, 2, 6

[43] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE TIP*, 2017. 5

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

[45] Yi Zhang, Hongwei Qin, Xiaogang Wang, and Hongsheng Li. Rethinking noise synthesis and modeling in raw denoising. In *ICCV*, 2021. 2

[46] Muming Zhao, Jian Zhang, Chongyang Zhang, and Wenjun Zhang. Leveraging heterogeneous auxiliary tasks to assist crowd counting. In *CVPR*, 2019. 2

[47] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. EEMEFN: Low-light image enhancement via edge-enhanced multi-exposure fusion network. In *AAAI*, 2020. 1, 2, 5, 6, 7

[48] Yunhao Zou and Ying Fu. Estimating fine-grained noise model via contrastive learning. In *CVPR*, 2022. 2