DarkFeat: Noise-Robust Feature Detector and Descriptor for Extremely Low-Light RAW Images

Yuze He,[1] Yubin Hu,[1] Wang Zhao,[1] Jisheng Li,[1] Yong-Jin Liu,[1*] Yuxing Han,[2] Jiangtao Wen[3]

[1]Tsinghua University
[2]Research Institute of Tsinghua University in Shenzhen
[3]Eastern Institute for Advanced Study

Abstract

Low-light visual perception, such as SLAM or SfM at night, has received increasing attention, in which keypoint detection and local feature description play an important role. Both handcraft designs and machine learning methods have been widely studied for local feature detection and description, however, the performance of existing methods degrades in the extreme low-light scenarios in a certain degree, due to the low signal-to-noise ratio in images. To address this challenge, images in RAW format that retain more raw sensing information have been considered in recent works with a denoise-then-detect scheme. However, existing denoising methods are still insufficient for RAW images and heavily time-consuming, which limits the practical applications of such scheme. In this paper, we propose DarkFeat, a deep learning model which directly detects and describes local features from extreme low-light RAW images in an end-to-end manner. A novel noise robustness map and selective suppression constraints are proposed to effectively mitigate the influence of noise and extract more reliable keypoints. Furthermore, a customized pipeline of synthesizing dataset containing low-light RAW image matching pairs is proposed to extend end-to-end training. Experimental results show that DarkFeat achieves state-of-the-art performance on both indoor and outdoor parts of the challenging MID benchmark, outperforms the denoise-then-detect methods and significantly reduces computational costs up to 70%.

Introduction

Keypoint detection and local feature description are fundamental operations in computer vision. Recently, simultaneous localization and mapping (SLAM) and structure-from-motion (SfM) in extremely low-light scenes, e.g., capture under moonlight with a video-rate exposure, has received increasing attention with the rise of applications such as autonomous driving, in which low-light feature detection and description play an important role. However, different from common daylight images, strong noises exist in the captured low-light images and it remains a great challenge to effectively detect and describe reliable keypoints on these images.

Despite existing local feature detection and description methods function well, it is difficult to adapt them to extremely low-light scenarios. Classical methods (Bay, Tuytelaars, and Gool 2006; Lowe 1999; Rublee et al. 2011; Shi et al. 1994) designed hand-crafted features based on local gradients or color distributions, which can be seriously affected by strong noise in dark images. Neural network based methods learn to detect and describe keypoints in the feature space with different strategies, which can effectively overcome lighting changes, blurs and low-textures (DeTone, Malisiewicz, and Rabinovich 2018; Revaud et al. 2019; Luo et al. 2020). However, none of these methods explicitly consider strong noise, which is an crucial factor cannot be neglected in extremely low-light scenarios.

To improve performance in dark environments, recent works utilize images in RAW format instead of the common RGB images. Although a considerable measure of noise still exists, RAW images have larger bit widths and retain richer original information than RGB images, which are processed by the image signal processing (ISP) module within cameras. Recently, based on RAW images, a benchmark called MID (Song et al. 2021) with a *denoise-then-detect* pipeline is proposed, which first denoises RAW images using BM3D (Dabov et al. 2007) or SID (Chen et al. 2018) and then applies feature detection, description and matching algorithms. Nevertheless, existing denoising algorithms are usually very time-consuming and sometimes introduce artifacts that hinder accurate keypoint detection, which limit the application of this *denoise-then-detect* pipeline.

In order to effectively detect and describe features in extremely low-light scenarios, we present an end-to-end framework called DarkFeat with noise-resistant oriented training constraints. To identify and exclude areas that are not salient or susceptible to noise, we propose *Noise Robustness Map*, which is learned by the average matching accuracy of pairs of noisy and noise-free images. To extract sufficient and reliable keypoints under noisy conditions, we further introduce selective suppression constraints to reduce keypoint scores in noisy or information-poor area. To train the DarkFeat, we also propose a simulation pipeline that can generate high-quality low-light RAW format image dataset with corresponding JPEG images.

In summary, our contributions are three-fold:

- Trained on our synthesized dataset, DarkFeat is the first end-to-end framework for RAW-format low-light image keypoint detection and local feature description;

*Corresponding author.
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

- We design a Noise Robustness Map that learns to mask less salient or noise-susceptible regions in the image;
- Selective suppression constraints are introduced that help to extract as many reliable keypoints as possible under noisy conditions.

Experimental results demonstrate that our DarkFeat framework can achieve state-of-the-art performance on MID benchmark with much less computation time, and is robust to different noisy conditions.

Related Works

Local feature detection and description. Early methods for keypoint and descriptor extraction use handcrafted features (Bay, Tuytelaars, and Gool 2006; Lowe 1999; Tardos 2016; Calonder et al. 2010; Rublee et al. 2011; Shi et al. 1994). They are generally efficient, but have a very limited receptive field and are highly susceptible to noise. For instance, small changes in local pixels can severely alter the results of the classic SIFT method (which makes use of local gradients to locate keypoints) and the classic FAST method (which detects corners through local color distributions). HarrisZ+ (Bellavia and Mishkin 2022) and KAZE (Alcantarilla, Bartoli, and Davison 2012) use edge enhancement algorithms to improve performance, but are still limited by a small perceptual field.

Learning-based approaches have received considerable attention recently due to the capacity of neural networks that can automatically extract good features and attenuate the effects of noise. Many methods adopt the *detect-then-describe* strategy (Yi et al. 2016; Tian, Fan, and Wu 2017; Simo-Serra et al. 2015). This strategy first extracts keypoints, then forms a small image patch centered at each keypoint, and finally passes each patch to the neural network to obtain the descriptor of the keypoint. As the number of keypoints increases, the efficiency of this scheme will be significantly reduced, and under the condition of strong noise, the small size of image patches prevents the network to restore the original distribution of the image with surrounding information, resulting in inaccurate descriptors.

Jointly learning feature detectors and descriptors have also been widely studied. R2D2 (Revaud et al. 2019) acquires keypoints by learning repeatability and reliability separately, where reliability is implicitly learned through differentiable average precision (AP). But the performance of R2D2 degrades when repeatability is perturbed by noise or wrong textures. SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018) creates pseudo-labels for keypoints using synthetic shapes and homographic adaption, and converts keypoint extraction into a classification problem for self-supervised learning. This method is helpful to resist noise, but at the same time leads to a reduction in the number of keypoints, which is detrimental to subsequent tasks such as pose estimation.

D2-Net (Dusmanu et al. 2019) adopts a *describe-and-detect* strategy to learn dense descriptors and extract keypoints through a handcrafted selection rule. ASLFeat (Luo et al. 2020) inherits this strategy and extends it to multiple levels of the network. This hierarchical structure allows the network to take into account features at different scales at the same time and has the potential to resist noise. Our work follows this structure to design appropriate modules and constraints, demonstrating the ability to resist strong noise.

Low-light vision tasks. Vision tasks in the low-light scenarios are gaining increasing attraction recently (Cui et al. 2021; Mildenhall et al. 2022; Wang, Yang, and Liu 2021; Wang et al. 2022). In the few existing works, most are about object detection. In order to achieve the object recognition of low-light images in RAW format, YOLO in the dark (Sasagawa and Nagahara 2020) uses the generative model and glues layers to connect the pre-trained networks of two independent tasks together. This method restricts the original size of the RAW image as input, and it is difficult to cope with extreme low light conditions. The method (Li et al. 2021) uses a knowledge distillation strategy, and maintains consistency by imposing a Euclidean distance loss between normal and noisy image models. The method (Hnewa and Radha 2021) designs a network that can learn domain invariant features through adversarial learning by adding a domain adaptive network. However, feature detectors and descriptors in low-light condition are not simple domain-invariant problems. As a low-level task, the detection score of certain regions in the image should be reduced after noise occurs, and there should be enough high-score regions to ensure a sufficient number of keypoints. To address this issue, our work proposes selective suppression constraints.

Dataset Construction

In order to provide a deep learning solution that can train feature detector and descriptor for low-light RAW images end-to-end, a large-scale dataset containing (RAW, RAW) image pairs with ground truth pixel-to-pixel correspondence is needed. Furthermore, these image pairs should contain sufficient variations of viewpoints and illuminations. However, to the best of our knowledge, there are no available datasets that meet all these requirements.

Existing image matching datasets are all in RGB format and are mainly constructed following two approaches. The first method warps monocular images with given homography matrices (DeTone, Malisiewicz, and Rabinovich 2018; Geiger, Lenz, and Urtasun 2012; Lin et al. 2014). Groundtruth pixel correspondences are naturally provided in this way. The second method directly selects image pairs from captured video sequences (Shen et al. 2018; Balntas et al. 2017; Luo et al. 2018; Sattler et al. 2018, 2012; Brown, Hua, and Winder 2010). These methods utilize Structure-from-Motion (SfM) and Multi-view Stereo (MVS) algorithms (Schonberger and Frahm 2016) to estimate camera poses and dense 3d maps. Pixel correspondences are then derived from map projections.

However, neither of the above two approaches is feasible to construct an image pair in RAW format, despite there exists low-light RAW image and video sequence datasets such as SID (Chen et al. 2018) and SMID (Chen et al. 2019). Synthesizing RAW image pairs by homography transformation results in distortions, where the noise in the warped RAW images severely shifts from the actual distribution, as shown

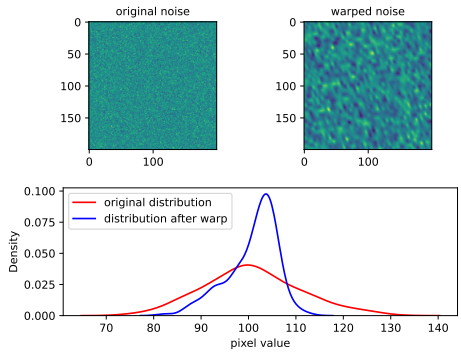

Figure 1: Warping the noise will change the original distribution (Poisson noise is used as an example in the figure).

in Fig.1, thus can not appropriately simulate the natural low-light RAW images. Meanwhile, the low signal-to-noise ratio in low-light RAW images makes it impossible to estimate accurate pose ground-truth by SfM. Furthermore, all existing RAW image datasets are in a small scale and not sufficient for training feature detectors and descriptors.

To address this, we propose a high-quality low-light RAW image generation pipeline to synthesize low-light RAW image pairs from the current large-scale RGB image matching dataset. Specifically, we first generate RAW images from RGB images by the reverse ISP, and then add the simulated low-light noise to the RAW images. Finally, we deal with image data augmentation for training deep models.

Reverse ISP. The ISP process is not fully reversible due to the information loss during bit width reduction and JPEG compression. We adopt the Invertible-ISP (Xing, Qian, and Chen 2021) to approximate the reverse ISP process with an invertible convolutional network and a differentiable JPEG module. Since it only requires RAW images for training, it is very easy to retrain with the new camera configuration.

Noise simulation. Each step of the sensor imaging in low-light environment generates a different noise distribution. We follow the noise model, distribution parameters, and calibration methods in ELD (Wei et al. 2020) to simulate the noise of extremely low-light raw images.

Data augmentation. Data augmentation of RAW-format images is under-explored and directly transferring the same augmentation operators and parameters from RGB leads to poor training performance. Thanks to our customized pipeline, we instead apply data augmentation on the original RGB image before reverse ISP, and then generate the noisy RAW images. This allows us to conduct a number of augmentations including motion blur, which is important for training deep matching models.

Method

Inspired by ASLFeat (Luo et al. 2020), our proposed system DarkFeat jointly learns descriptors and detection scores from the input image. However, directly training the network by the descriptor loss of noisy image pairs results in unstable and degraded performances. Instead, we propose to guide the training with both noise-free (note as normal image) and noisy RAW images, as shown in Fig. 2. More importantly, two novel noise-resistant oriented constraints, namely noise robustness map and selective suppression constraint, are introduced to cope with the noise influence. With these carefully designed constraints and customized large-scale RAW-format low-light matching dataset, DarkFeat could effectively detect and describe the keypoints under extremely low-light environments with considerable noise in RAW images. Next, we will introduce the noise robustness map and selective suppression constraint in detail.

Noise Robustness Map

Some image regions are unreliable for keypoint extraction and matching, such as low-textured or repeated-textured regions, as studied in R2D2 (Revaud et al. 2019), but these regions may contain random structural information under severe noise and be identified as reliable regions. Meanwhile, regions with complex textures or low contrast are also vulnerable to noise. Thus, it is beneficial to effectively identify and exclude these regions under the noise condition.

Global metrics like the average precision (AP) have been introduced to detect unreliable image regions. (He, Lu, and Sclaroff 2018) proposed a differentiable average precision (AP) approximation, which measures how well a particular descriptor matches. Given a patch descriptor as a query, the Euclidean distances between the query descriptor and the target descriptors of all matching candidates are first calculated and ranked. The AP achieves the optimal (largest) value if the correct matching target is ranked above all other candidates. R2D2 (Revaud et al. 2019) followed this idea and replaced patch descriptors with dense descriptors, so that AP can be calculated for each pixel (i, j) with the aid of ground-truth pixel-wise correspondences. For low-textured or repeated-textured image regions such as flat wall or grass, AP is always low. R2D2 then proposed the reliability map R_{ij} to identify these regions, and trained the learning of reliability map through the AP loss function:

$$\mathcal{L}_{AP_\kappa}(i,j) = 1 - [AP(i,j)R_{ij} + \kappa(1 - R_{ij})], \quad (1)$$

where $\kappa \in [0, 1]$ is a threshold hyperparameter. The AP loss makes the network tend to predict R_{ij} as 1 when $AP(i, j)$ is larger than κ, and 0 otherwise.

This mechanism works well on normal-light images, but suffers when directly applied to low-light images. The calculation of AP loss at vulnerable image regions may be severely affected by noise and not able to guide the learning of reliability map. From this, we propose a novel noise robustness map $\mathcal{R}$ to address the above problem. Note that during the training, the network takes both normal image pair I_{normal}, I'_{normal} and noisy image pair I_{noise}, I'_{noise} as input, so it gets both $AP_{normal}, \mathcal{R}_{normal}$ and $AP_{noise}, \mathcal{R}_{noise}$.

To mask out unreliable regions where accurate matching is not possible even in the absence of noise, we supervise the $\mathcal{R}_{noise}$ with AP_{normal}:

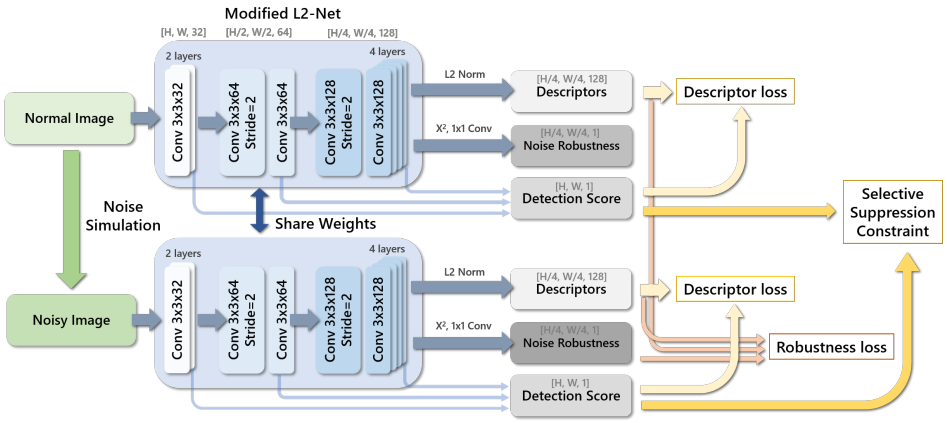

Figure 2: The overall training pipeline of DarkFeat. During training, both noisy image pair and normal image pair are input and we use the combination of descriptor loss, noise robustness map loss and selective suppression constraint to train the network. At inference time, the network takes a RAW image as input and predicts the noise robustness map, detection score and dense descriptors. We further extract the top scoring keypoints regarding to the noise robustness map and detection scores.

$$\mathcal{L}_{AP}(i,j) = 1 - [AP_{normal}(i,j)\mathcal{R}_{noise,ij} + \kappa(1 - \mathcal{R}_{noise,ij})]. \tag{2}$$

To identify vulnerable regions where matching accuracy drops significantly under the influence of noise, we use the difference of AP before and after adding noise as the supervision:

$$\begin{aligned} \Delta AP(i,j) &= \max(AP_{normal}(i,j) - AP_{noise}(i,j), 0), \\ \mathcal{L}_{\Delta AP}(i,j) &= 1 - [\Delta AP(i,j)(1 - \mathcal{R}_{noise,ij}) + \kappa' \mathcal{R}_{noise,ij}]. \end{aligned} \tag{3}$$

The final loss function is obtained by superimposing the above two items:

$$\mathcal{L}_{Rob}(i,j) = \lambda \mathcal{L}_{AP}(i,j) + \lambda' \mathcal{L}_{\Delta AP}(i,j), \tag{4}$$

where κ and κ' are expectation hyperparameters of AP and ΔAP set to 0.25 and 0.2, λ and λ' are scale factors set to 0.5. In this way, our noise robustness map $\mathcal{R}$ learns to identify and exclude both non-discriminative regions and noise-vulnerable regions, by effectively supervising the noisy image pairs with the normal image outputs, thus greatly relieve the training difficulty.

Selective suppression constraint

Noise robustness map helps to exclude the unreliable image regions, preventing sampling non-discriminative keypoints. Nevertheless, as discussed in ASLFeat (Luo et al. 2020), salient keypoints are further supposed to be localized accurately to satisfy the requirements of camera geometry recovery, and multi-level spatial details are beneficial for accurate keypoint localization. Therefore they introduced detection score map to indicate the keypoint localization accuracy and injected it into the descriptor loss for training. However, with strong noise in exhibition, training the keypoint detector with vanilla detection score map leads to inferior accuracy, since the noise seriously disturbs the local feature distribution. Our method inherits the design of detection score map and makes key contributions to deal with noise.

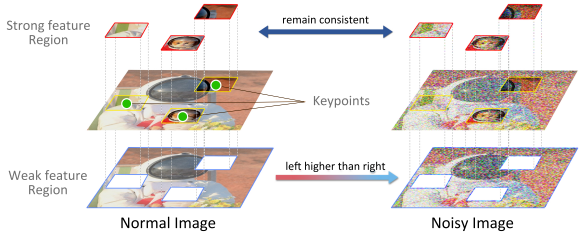

Figure 3: Illustration of selective suppression constraint. In strong feature regions, the detection score of noisy image is supposed to match up with the score of normal image, while in weak feature regions the detection score of noisy image should be lower than the score of normal image.

Specifically, we propose a selective suppression constraint on the keypoint detection score map. We treat the detection scores of noise-free and noisy image as two domains, and manage to regularize the noise one using the clean noise-free detection scores. The intuition is that for regions with the rich useful features, the detection scores of the noisy image should be as close as possible to the noise-free image; conversely, for the regions with poor features, the score of the noise image should be smaller than the normal image, avoiding identifying the noise as structures.

To calculate the selective suppression constraint, we first define the strong feature regions by the keypoints of noise-free image. For a noise-free image input I_{normal}, a standard keypoint extraction process is conducted according to the score map s_{normal} and the keypoints with top k scores are acquired to get a point set $\{p_s\}$. Then we define a strong feature region mask $\mathcal{M}$ in the neighborhood of each keypoint in $\{p_s\}$ as follows:

$$M_{ij} = \begin{cases} 0, & \forall p_l \in \{p_s\},\ (i,j) \notin \mathcal{N}(p_l) \\ 1, & else. \end{cases} \tag{5}$$

where i, j are the horizontal and vertical coordinates of the pixel, and $\mathcal{N}(\cdot)$ is the 3x3 neighborhood centered on the point. In practice we set k as 512. After that, the selective suppression constraint is defined as:

$$\begin{aligned} l_{1,ij} &= M_{ij} \cdot |s_{normal,ij} - s_{noise,ij}|, \\ l_{2,ij} &= (1 - M_{ij}) \cdot \max(s_{noise,ij} - s_{normal,ij} - \theta, 0), \end{aligned} \tag{6}$$

where s_{normal}, s_{noise} are the keypoint detection scores of the noise-free image and the noisy image, respectively, and θ is a threshold set to 0.1 that regulates the gap between the score map of noise image and noise-free image in the weak feature regions. To avoid trivial collapse of $\mathcal{M}$, we add the normalization weights into the final loss:

$$\begin{aligned} \mathcal{L}_{ss}(I_{normal}, I_{noise}) &= \sum_{i=1}^{H}\sum_{j=1}^{W}(w_1 l_{1,ij} + w_2 l_{2,ij}), \\ w_1 &= \frac{1}{\sum_{i=1}^{H}\sum_{j=1}^{W} M_{ij}}, \\ w_2 &= \frac{1}{\sum_{i=1}^{H}\sum_{j=1}^{W}(1 - M_{ij})}. \end{aligned} \tag{7}$$

We also use the detection score based descriptor loss to jointly train the detector and descriptor, following the design of ASLFeat. The descriptor loss is calculated for both normal and noisy image pairs:

$$\mathcal{L}_{feat}(I, I') = \frac{1}{\mathcal{C}} \sum_{c \in \mathcal{C}} \frac{\hat{s}_c \hat{s}'_c}{\sum_{q \in \mathcal{C}} \hat{s}_q \hat{s}'_q} \mathcal{M}(f_c, f'_c), \tag{8}$$

where $\hat{s}_k$ and $\hat{s}'_k$ are keypoint detection scores; f_k and f'_k are feature descriptors.The ranking loss $\mathcal{M}(\cdot, \cdot)$ is defined as

$$\begin{aligned} \mathcal{M}(f_c, f'_c) &= [D(f_c, f'_c) - m_p]_+ \\ &+ [m_n - \min(\min_{k \neq c} D(f_c, f'_k), \min_{k \neq c} D(f_k, f'_c))]_+, \end{aligned} \tag{9}$$

where $D(\cdot, \cdot)$ is the Euclidean distance and m_p, m_n are the thresholds respectively set to 0.2 and 1.0.

Implementations

Network architecture. The network backbone adopts L2-Net (Tian, Fan, and Wu 2017). Following R2D2 (Revaud et al. 2019) and ASLFeat (Luo et al. 2020), we replace the last 8x8 convolutional layer with three 3x3 convolutional layers. The dense descriptors and noise robustness map are obtained by l2-normalization in the channel dimension and 1x1 convolution after element-wise squaring on the output of the last layer of the network, respectively. Keypoint detection score maps are obtained by using the same MulDet strategy as ASLFeat on three convolutional layers with different resolutions (conv1, conv3 and conv8).

Dataset. We use GL3D (Luo et al. 2018) as the source RGB image matching dataset to generate our RAW image matching dataset. The noise model parameters and reverse ISP network are determined based on Canon EOS 5D4 camera. During training, the original JPEG images in GL3D are subjected to data augmentation, simulated into RAW format and added with noise, and then bilinearly demosaiced and sent to the network.

Training. In order to effectively train the network, the training process is divided into three stages: the first stage only uses noise-free demosaic-raw image pairs as input, and trains 160k iters with only the descriptor loss; the second stage uses the same input but adds noise robustness map to the output, and train 60k iters with the descriptor loss and vanilla reliability map loss ($\kappa = 0.3$) as defined in Eq. 1; the final stage takes both noise-free and noisy demosaic-raw image pairs as input, and uses all three loss functions $\mathcal{L}_{feat}, \mathcal{L}_{rob}, \mathcal{L}_{ss}$ to train 50k iters. The three losses are superimposed by coefficients 1.0, 0.5, and 1.0 respectively as the final loss. Images are resized to 480x480 during training, and the network is optimized using a SGD optimizer with a batch size of 2, a weight decay of 0.0001, the initial learning rate is 0.1 and drops to 0.01 in the final stage of training.

Testing. To extract the keypoints, we first multiply the noise robustness map element-wise with the detection score map to get the final score map, then apply a non-maximum suppression with a radius of 3 and an edge removal with a width of 5. Similar to ASLFeat, we postprocess with a SIFT-like edge elimination with a size of 10. Finally we select the top 5000 keypoints regarding to the final scores, and discard the keypoints whose scores are lower than 0.50.

Experiments

Dataset and metrics

We evaluate the proposed method on the challenging Matching In the Dark (MID) benchmark (Song et al. 2021), a large-scale low-light stereo RAW image dataset, which is also the only currently available dataset for low-light image matching evaluation. The MID dataset contains 54 indoor scenes and 54 outdoor scenes, and each scene contains 48 stereo image pairs in RAW format taken with 6 different exposure times and 8 different ISOs. The ground truth relative camera pose between each stereo image pair is provided.

Following the metrics of MID benchmark, we evaluate the estimated relative camera pose accuracy for all methods. Given estimated pixel matches, we first estimate the essential matrix using the OpenCV library API, and then calculate the angular difference of rotation ΔR and translation Δt with the ground truth pose, then the maximum value of these two is defined as the angular error, same as (Sarlin et al. 2020; Yi et al. 2018). Given the threshold τ of 5°and 10°, we compute the ratio N_τ between the image pairs whose angular error is less than τ and all image pairs.

Comparisons with the state-of-the-art

Baselines. To the best of our knowledge, DarkFeat is the first end-to-end learning method for low-light keypoint de-

Enhancer	Method	Indoor		Outdoor		Time(s)
		$N\tau$@5	$N\tau$@10	$N\tau$@5	$N\tau$@10	
ISP	ORB+NN	0.019	0.027	0.030	0.057	4.282
	SIFT+NN	0.022	0.027	0.040	0.063	5.494
	SP+NN	0.247	0.288	0.237	0.311	4.267
	R2D2+NN	0.115	0.146	0.091	0.143	4.741
	ASLFeat+NN	0.354	0.412	0.338	0.430	4.379
	SP+SG	0.267	0.296	0.226	0.308	4.296
BM3D	ORB+NN	0.205	0.293	0.134	0.211	60.11
	SIFT+NN	0.276	0.367	0.296	0.400	61.33
	SP+NN	0.529	0.586	0.419	0.533	60.10
	R2D2+NN	0.404	0.487	0.307	0.414	60.57
	ASLFeat+NN	0.556	0.640	0.508	0.610	60.21
	SP+SG	0.520	0.592	0.461	0.564	60.13
SID	ORB+NN	0.151	0.238	0.110	0.175	0.665
	SIFT+NN	0.315	0.399	0.308	0.400	1.877
	SP+NN	0.562	0.636	0.447	0.562	0.650
	R2D2+NN	0.414	0.487	0.312	0.406	1.124
	ASLFeat+NN	0.514	0.599	0.456	0.558	0.763
	SP+SG	0.616	0.697	0.507	0.627	0.679
HistEQ	ORB+NN	0.204	0.304	0.152	0.235	0.298
	SIFT+NN	0.288	0.375	0.276	0.371	1.510
	SP+NN	0.528	0.584	0.422	0.537	0.284
	R2D2+NN	0.404	0.474	0.286	0.393	0.757
	ASLFeat+NN	0.562	0.634	0.503	0.598	0.396
	SP+SG	0.527	0.595	0.445	0.552	0.312
	DarkFeat (ours)	**0.618**	**0.705**	**0.519**	**0.633**	0.299
None	**DarkFeat (ours)**	0.617	0.701	0.500	0.629	**0.077**

Table 1: Evaluation Results on MID dataset. SP represents for SuperPoint and SG represents for SuperGlue.

tection and local feature description based on RAW images. Current state-of-the-art systems follow the *denoise-then-detect* strategy, as studied in the MID benchmark. By combining RAW image denoising methods with RGB based keypoint detection and description methods, these systems achieve reasonable performance. We follow MID and compare with these strong *denoise-then-detect* baselines.

For the RAW-format denoising methods, we compare three representative methods HistEQ (Pizer et al. 1987), BM3D (Dabov et al. 2007) and SID (Chen et al. 2018). We also include the standard camera ISP as the pre-processing for comparison. For HistEQ, we demosaic the original RAW image, perform channel-wise histogram equalization and then map the brightness in the range $[m - 2d, m + 2d]$ to the range $[0, 255]$, where m is the average brightness and d is the mean absolute difference from m to each pixel value, and finally resize it to 960x640. For BM3D, we use the same brightness mapping as above, then resize to 960x640, and employ BM3D with a noise PSD ratio of 0.08. For SID, we directly input the RAW image to the pretrained SID network to obtain the denoised RGB image, and resize it to 960x640.

As for the local feature detection and description, we select ORB (Rublee et al. 2011) and SIFT (Lowe 1999) as the representative handcraft method, and compare with state-of-the-art learning methods ASLFeat (Luo et al. 2020), SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018) and R2D2 (Revaud et al. 2019). For the keypoint matching, we test the nearest neighbor search (NN) for all methods and additionally compare with the recently popular SuperGlue (Sarlin et al. 2020) (only pretrained for SuperPoint).

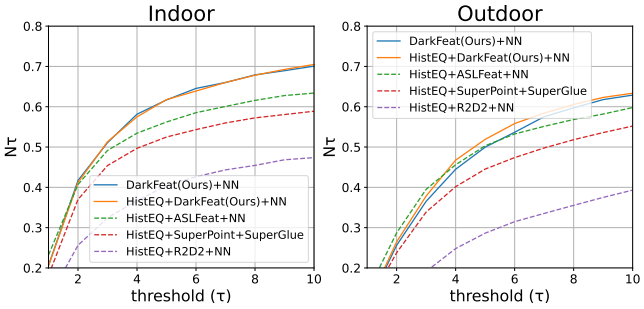

Figure 4: Comparisons on MID dataset with angular error evaluated at different error thresholds.

More baseline comparisons can be found in the Appendix.

Quantitative results. We summarize the quantitative results in Table 1. Our DarkFeat outperforms all baseline methods in both indoor and outdoor splits, thanks to the large-scale end-to-end training and noise-resistant constraint designs. Explicit noise reduction is overall effective compared to the standard camera ISP, however in some cases it introduces wrong non-existing textures, and can not actively identify and exclude regions that are not conducive to keypoint extraction. In contrast, DarkFeat utilizes noise robustness map and selective suppression constraint to effectively handle noise, and the end-to-end training avoids the modular error accumulation. We also show the complete curve with respect to different angular error thresholds in Figure 4.

Inference time. In Table 1 we also show the inference running time of each method. Effective noise reduction methods such as BM3D and SID are heavily time-consuming, limiting their deployment in practical applications. As a comparison, DarkFeat without HistEQ pre-processing can achieve around 13fps speed and still remain superior performance, which makes it possible for practical real-time low-light systems. All timing tests are executed on a single Nvidia RTX 2080Ti GPU and Intel Xeon Gold 6126 CPU @ 2.60GHz.

Visualization. Qualitative example of low-light keypoint detection and matching is shown in Figure 5. The original RAW images are captured with ISO=800 and 1/100 shutter speed. HistEQ+ASLFeat+NN provides noisy pixel matches, while SID+SuperPoint+SuperGlue only gives a small set of confident matches, both resulting in poor camera pose estimation. Our DarkFeat can extract dense and accurate keypoints and descriptors for accurate pose estimation.

Ablation study

Selective suppression constraints. We study the effectiveness of each proposed contribution by detailed ablations. In Table 2, we test different design choices of regularize the detection score maps of noisy image with noise-free image. We note the baseline without selective suppression constraint as Finetune. Furthermore, we replace the selective suppression constraint with several commonly used domain-invariant constraints. MSE refers to applying MSE loss to the detection score map of noise image and noise-free im-

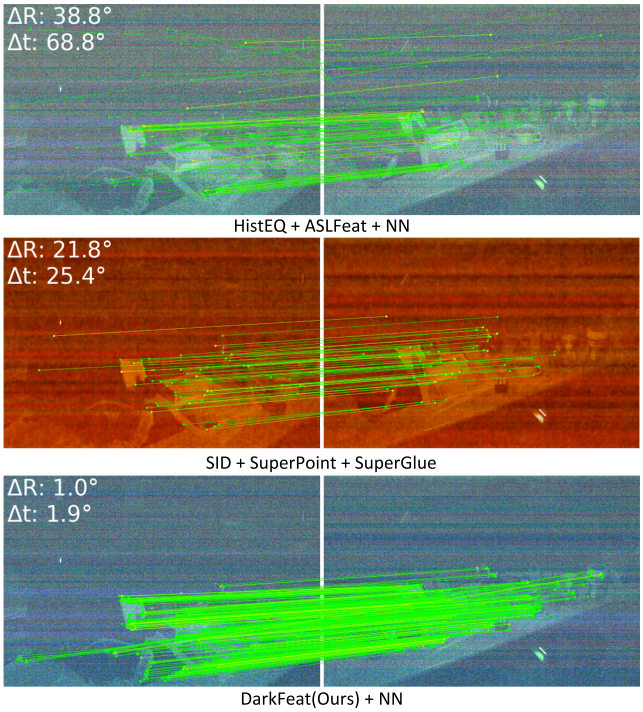

Figure 5: Visualization of low-light image matching results of different methods on MID benchmark. The angular error is shown at the upper left corner. DarkFeat achieves the lowest error with sufficient and accurate pixel matches.

age; MSEDesc refers to applying MSE loss to the dense descriptors of noise image and normal image; Attention refers to the attention loss from multi-level feature map of noise image and noise-free image (conv1, conv3 and conv8) as defined in (Zagoruyko and Komodakis 2016).

Group	Method	$N\tau$@5	$N\tau$@10
Indoor	Finetune	0.461	0.571
	+ Attention	0.448	0.554
	+ MSE	0.470	0.571
	+ MSEDesc	0.535	0.626
	+ selective suppression constraint	**0.606**	**0.689**
Outdoor	Finetune	0.395	0.507
	+ Attention	0.374	0.503
	+ MSE	0.404	0.525
	+ MSEDesc	0.463	0.574
	+ selective suppression constraint	**0.497**	**0.620**

Table 2: Ablation experiments on selective suppression constraint. The results demonstrate the superiority of selective suppression constraint over other domain-invariant losses.

As shown in Table 2, our selective suppression constraint is crucial to the success of DarkFeat, removing it and directly training the model with the noisy images results in much degraded performance. Also, other domain-invariant constraints such as MSE loss help for the training, but not as effective as our proposed selective suppression constraint, since different image regions should be treated differently.

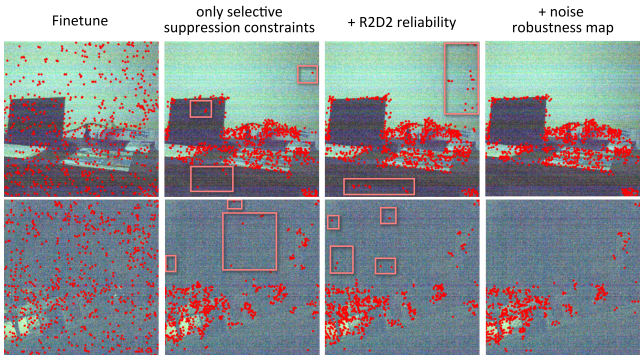

Figure 6: Visualization of ablations for keypoint extraction. Our selective suppression constraint and noise robustness map can effectively reduce the inappropriate keypoints.

Group	Method	$N\tau$@5	$N\tau$@10
Indoor	Only selective suppression	0.606	0.689
	+ Reliability (normal)	0.606	0.698
	+ Reliability (noise)	0.571	0.657
	+ Noise robustness map	**0.618**	**0.705**
Outdoor	Only selective suppression	0.497	0.620
	+ Reliability (normal)	0.493	0.622
	+ Reliability (noise)	0.486	0.595
	+ Noise robustness map	**0.519**	**0.633**

Table 3: Ablation experiments on noise robustness map. By integrating the average precision (AP) information from both noise-free and noisy images, noise robustness map helps to better exclude non-discriminative regions.

See Figure 6 for the visualization of ablations.

Noise robustness map. We further conduct ablations on our proposed noise robustness map, and summarize the results in Table 3. Reliability (normal) refers to the usage of vanilla reliability map loss proposed in R2D2 with only using AP of normal images for training; Reliability (noise) also refers to the vanilla reliability map loss but using AP of noisy images for training. Table 3 indicates that using AP loss separately for normal images or noisy images does not bring improvement, while our noise robustness map integrates both and better guide the training under strong noise.

Conclusion

Different from traditional *denoise-then-detect* strategies, in this paper, we proposed the DarkFeat framework, which outputs local feature detection and description from RAW-format images in an end-to-end way. By building a high-fidelity low-light RAW data generation pipeline for building a large-scale training dataset and proposing a novel noise robustness map and selective suppression constraints, DarkFeat can reliably extract sufficient keypoints while suppressing regions with insignificant and noise-susceptible features. Experimental results show that DarkFeat achieves state-of-the-art performance while reducing runtime by up to 70%.

Acknowledgments

This work was partially supported by the Natural Science Foundation of China (61725204).

References

Alcantarilla, P. F.; Bartoli, A.; and Davison, A. J. 2012. KAZE features. In *European conference on computer vision*, 214–227. Springer.

Balntas, V.; Lenc, K.; Vedaldi, A.; and Mikolajczyk, K. 2017. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5173–5182.

Bay, H.; Tuytelaars, T.; and Gool, L. V. 2006. Surf: Speeded up robust features. In *European conference on computer vision*, 404–417. Springer.

Bellavia, F.; and Mishkin, D. 2022. HarrisZ+: Harris corner selection for next-gen image matching pipelines. *Pattern Recognition Letters*, 158: 141–147.

Brown, M.; Hua, G.; and Winder, S. 2010. Discriminative learning of local image descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 33(1): 43–57.

Calonder, M.; Lepetit, V.; Strecha, C.; and Fua, P. 2010. Brief: Binary robust independent elementary features. In *European conference on computer vision*, 778–792. Springer.

Chen, C.; Chen, Q.; Do, M. N.; and Koltun, V. 2019. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3185–3194.

Chen, C.; Chen, Q.; Xu, J.; and Koltun, V. 2018. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3291–3300.

Cui, Z.; Qi, G.-J.; Gu, L.; You, S.; Zhang, Z.; and Harada, T. 2021. Multitask aet with orthogonal tangent regularity for dark object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2553–2562.

Dabov, K.; Foi, A.; Katkovnik, V.; and Egiazarian, K. 2007. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8): 2080–2095.

DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 224–236.

Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; and Sattler, T. 2019. D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 8092–8101.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

He, K.; Lu, Y.; and Sclaroff, S. 2018. Local descriptors optimized for average precision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 596–605.

Hnewa, M.; and Radha, H. 2021. Multiscale domain adaptive yolo for cross-domain object detection. In *2021 IEEE International Conference on Image Processing (ICIP)*, 3323–3327. IEEE.

Li, C.; Qu, X.; Gnanasambandam, A.; Elgendy, O. A.; Ma, J.; and Chan, S. H. 2021. Photon-limited object detection using non-local feature matching and knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3976–3987.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, 1150–1157. Ieee.

Luo, Z.; Shen, T.; Zhou, L.; Zhu, S.; Zhang, R.; Yao, Y.; Fang, T.; and Quan, L. 2018. Geodesc: Learning local descriptors by integrating geometry constraints. In *European Conference on Computer Vision (ECCV)*.

Luo, Z.; Zhou, L.; Bai, X.; Chen, H.; Zhang, J.; Yao, Y.; Li, S.; Fang, T.; and Quan, L. 2020. Aslfeat: Learning local features of accurate shape and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6589–6598.

Mildenhall, B.; Hedman, P.; Martin-Brualla, R.; Srinivasan, P. P.; and Barron, J. T. 2022. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16190–16199.

Pizer, S. M.; Amburn, E. P.; Austin, J. D.; Cromartie, R.; Geselowitz, A.; Greer, T.; ter Haar Romeny, B.; Zimmerman, J. B.; and Zuiderveld, K. 1987. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3): 355–368.

Revaud, J.; Weinzaepfel, P.; De Souza, C.; Pion, N.; Csurka, G.; Cabon, Y.; and Humenberger, M. 2019. R2D2: repeatable and reliable detector and descriptor. *arXiv preprint arXiv:1906.06195*.

Rublee, E.; Rabaud, V.; Konolige, K.; and Bradski, G. 2011. ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision*, 2564–2571. Ieee.

Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4938–4947.

Sasagawa, Y.; and Nagahara, H. 2020. Yolo in the dark-domain adaptation method for merging multiple models. In *European Conference on Computer Vision*, 345–359. Springer.

Sattler, T.; Maddern, W.; Toft, C.; Torii, A.; Hammarstrand, L.; Stenborg, E.; Safari, D.; Okutomi, M.; Pollefeys, M.; Sivic, J.; et al. 2018. Benchmarking 6dof outdoor visual

localization in changing conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8601–8610.

Sattler, T.; Weyand, T.; Leibe, B.; and Kobbelt, L. 2012. Image Retrieval for Image-Based Localization Revisited. In *BMVC*, volume 1, 4.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Shen, T.; Luo, Z.; Zhou, L.; Zhang, R.; Zhu, S.; Fang, T.; and Quan, L. 2018. Matchable Image Retrieval by Learning from Surface Reconstruction. In *The Asian Conference on Computer Vision (ACCV*.

Shi, J.; et al. 1994. Good features to track. In *1994 Proceedings of IEEE conference on computer vision and pattern recognition*, 593–600. IEEE.

Simo-Serra, E.; Trulls, E.; Ferraz, L.; Kokkinos, I.; Fua, P.; and Moreno-Noguer, F. 2015. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE international conference on computer vision*, 118–126.

Song, W.; Suganuma, M.; Liu, X.; Shimobayashi, N.; Maruta, D.; and Okatani, T. 2021. Matching in the Dark: A Dataset for Matching Image Pairs of Low-light Scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6029–6038.

Tardos, J. D. 2016. Feature-Based Visual SLAM. In *2016 IEEE International Conference on Robotics and Automation*. IEEE.

Tian, Y.; Fan, B.; and Wu, F. 2017. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 661–669.

Wang, W.; Wang, X.; Yang, W.; and Liu, J. 2022. Unsupervised Face Detection in the Dark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, W.; Yang, W.; and Liu, J. 2021. Hla-face: Joint high-low adaptation for low light face detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16195–16204.

Wei, K.; Fu, Y.; Yang, J.; and Huang, H. 2020. A physics-based noise formation model for extreme low-light raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2758–2767.

Xing, Y.; Qian, Z.; and Chen, Q. 2021. Invertible image signal processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6287–6296.

Yi, K. M.; Trulls, E.; Lepetit, V.; and Fua, P. 2016. Lift: Learned invariant feature transform. In *European conference on computer vision*, 467–483. Springer.

Yi, K. M.; Trulls, E.; Ono, Y.; Lepetit, V.; Salzmann, M.; and Fua, P. 2018. Learning to find good correspondences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2666–2674.

Zagoruyko, S.; and Komodakis, N. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.