DarkVisionNet: Low-Light Imaging via RGB-NIR Fusion with Deep Inconsistency Prior

Shuangping Jin,[1, 2*] **Bingbing Yu,** [1, 3*] **Minhao Jing** [1] **Yi Zhou** [1] **Jiajun Liang** [1] **Renhe Ji** [1†]

[1] Megvii Technology
[2] Southeast University
[3] Dalian University Of Technology
220191583@seu.edu.cn, 21901037@mail.dlut.edu.cn, jingminhao@megvii.com, zhouyi@megvii.com, liangjiajun@megvii.com, jirenhe@megvii.com

Abstract

RGB-NIR fusion is a promising method for low-light imaging. However, high-intensity noise in low-light images amplifies the effect of structure inconsistency between RGB-NIR images, which fails existing algorithms. To handle this, we propose a new RGB-NIR fusion algorithm called Dark Vision Net (DVN) with two technical novelties: Deep Structure and Deep Inconsistency Prior (DIP). The Deep Structure extracts clear structure details in deep multiscale feature space rather than raw input space, which is more robust to noisy inputs. Based on the deep structures from both RGB and NIR domains, we introduce the DIP to leverage the structure inconsistency to guide the fusion of RGB-NIR. Benefiting from this, the proposed DVN obtains high-quality low-light images without the visual artifacts. We also propose a new dataset called Dark Vision Dataset (DVD), consisting of aligned RGB-NIR image pairs, as the first public RGB-NIR fusion benchmark. Quantitative and qualitative results on the proposed benchmark show that DVN significantly outperforms other comparison algorithms in PSNR and SSIM, especially in extremely low light conditions.

Introduction

High-quality low-light imaging is a challenging but significant task. On the one hand, it is the cornerstone of many important applications such as 24-hour surveillance, smartphone photography, etc. On the other hand, though, massive noise of images in extremely dark environments hinders algorithms from the satisfactory restoration of low-light images. RGB-NIR fusion techniques provide a new perspective for the challenge: it enhances the low-light noisy color (RGB) image through rich, detailed information in the corresponding near-infrared (NIR) image (The high quality of NIR images in dark environments comes from invisible near-infrared flash), which greatly improves the signal-to-noise ratio (SNR) of the restored RGB image. Under the constraint of cost, size and other factors, RGB-NIR fusion becomes the most promising technique to restore the vanished textual and structural details from noisy RGB images taken in an extremely low-light environments, as shown in Figure 1(a).

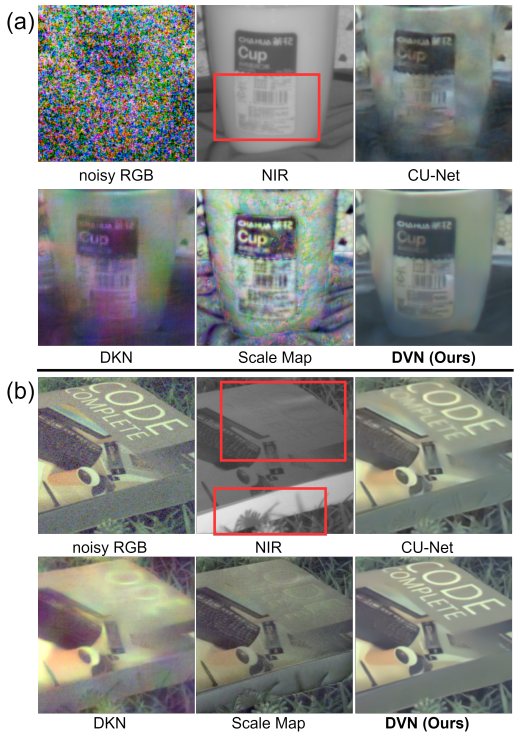

Figure 1: (a) and (b) are fusion examples from DVD, compared to CU-Net (Deng and Dragotti 2020), DKN (Kim, Ponce, and Ham 2021) and Scale Map (Yan et al. 2013), our method, Dark Vision Net (DVN), effectively handle the structure inconsistency between RGB-NIR images. Regions with inconsistent structures are framed in red.

However, the existing RGB-NIR fusion algorithms suffers from the problem of structure inconsistency between RGB and NIR images, resulting in unnatural appearance and loss of key information, which limits the application of RGB-NIR fusion algorithm in low-light imaging. Figure 1 illustrates two typical examples of structure inconsistency between RGB and NIR images: Figure 1(b) shows the absence of NIR shadows in the RGB image (grass shadows only appear on the book edge in the NIR image), and

*Authors contributed equally.
†Corresponding author.

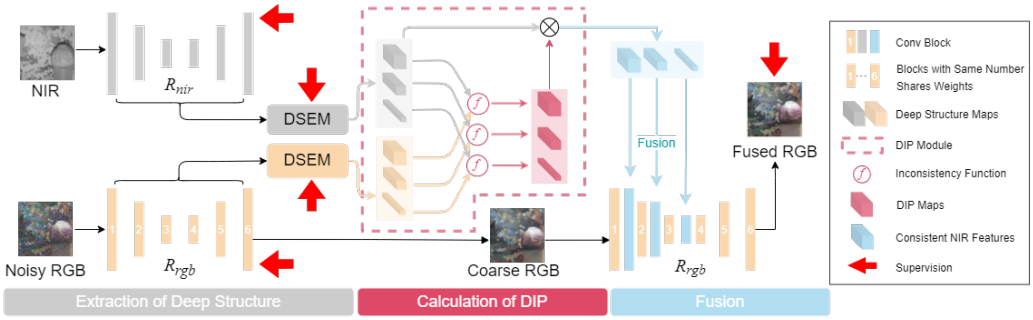

Figure 2: Overview of the proposed DVN. In the first stage, the network predicts deep structure maps utilising the multi-scale features maps from restoration network R by the proposed Deep Structure Extraction Module (DSEM) for noisy RGB and NIR respectively. In the second stage, taking advantage of the predicted deep structures, the DIP can be calculated by inconsistency function $\mathcal{F}$. In the third stage, the DIP-weighted NIR structures are fused with the RGB features to obtain the final fusion result without obvious structure inconsistency.

the nonexistence of RGB color structure in the NIR image (text 'complete' almost disappears on the book cover in the NIR image). Fusion algorithms need to tackle these structure inconsistency to avoid visual artifacts in output images. There are two categories of RGB-NIR fusion methods currently, *i.e.* traditional methods and neural-network-based methods, and modeling the structure of the paired RGB-NIR images plays an important role in both of them. Traditional methods, such as Scale Map (Yan et al. 2013), tackle the structure inconsistency problem by manually designed functions. Some neural-network-based methods (Kim, Ponce, and Ham 2021; Li et al. 2019), on the other hand, utilize deep learning techniques to automatically learn the structure inconsistency by a large amount of data. Both of them perform well under certain circumstances.

However, when confronted with extreme low-light environments, existing methods fail to maintain satisfactory performance, since the structure inconsistency is dramatically exacerbated by massive noise in the RGB image. As shown in Figure 1(b), the dense noise in the RGB image makes it difficult for Scale Map to extract structural information, causing the failure of distinguishing which structures in the NIR image should be eliminated, and result in the unnatural ghost images on the book edge. Deformable Kernel Networks (DKN) (Kim, Ponce, and Ham 2021) falsely weakens gradients of input RGB image that do not exist in the corresponding NIR image, which leads to the blurriness of letters on the book cover. Even though these structural inconsistency of corresponding RGB and NIR images can be captured by human eyes effortlessly, they still confuse most of the existing fusion algorithms.

In this paper, we focus on improving the RGB-NIR fusion algorithm for extremely low SNR images by tackling the structure inconsistency problem. Based on the above analysis, we argue that the structure inconsistency under extremely low light can be handled well by introducing prior knowledge into deep features. To achieve this, we propose a deep RGB-NIR fusion algorithm called Dark Vision Net (DVN), which explicitly leverages the prior knowledge of structure inconsistency to guide the fusion of RGB-NIR deep features, as shown in Figure 2. With DVN, two technical novelties are introduced: (1) We find a new way, referred to as deep structures, to represent clear structure information encoded in deep features extracted by the proposed Deep Structure Extraction Module (DSEM). Even facing images with low SNR, the deep structures can still be effectively extracted and represent reliable structural information, which is critical to the introduction of prior knowledge. (2) We propose Deep Inconsistency Prior (DIP), which indicates the differences between RGB-NIR deep structures. Integrated into the fusion of RGB-NIR deep features, the DIP empowers the network to handle the structure inconsistency. Benefiting from this, the proposed DVN can obtain high-quality low-light images.

In addition, to the best of our knowledge, there is no available benchmark dedicated for the RGB-NIR fusion task so far. The lack of benchmarks to evaluate and train fusion algorithms greatly limits the development of this field. To fill this gap, we propose a dataset named Dark Vision Dataset (DVD) as the first RGB-NIR fusion benchmark. Based on this dataset, we give qualitative and quantitative evaluations to prove the effectiveness of our method. In summary, the main contributions of this paper are as follows:

- We propose a novel RGB-NIR fusion algorithm called Dark Vision Net (DVN) With Deep Inconsistency Prior (DIP). The DIP explicitly integrates the prior of structure inconsistency into the deep features, avoiding overrelying on NIR features in the feature fusion. Benefits from this, DVN can obtain high-quality low-light images without visual artifacts.
- We propose a new dataset Dark Vision Dataset (DVD) as the first public dataset for training and evaluating RGB-NIR fusion algorithms.
- The quantitative and qualitative results indicate that DVN is significantly better than other compared methods.

Related Work

Image Denoising. In recent years, denoising algorithms based on deep neural networks have continually emerged and overcome the drawbacks of analytical methods (Lucas et al. 2018). The image noise model is gradually improved simultaneously (Wei et al. 2020; Wang et al. 2020).

(Mao, Shen, and Yang 2016) applied an encoder-decoder network to suppress the noise and recover high-quality images. (Zhang, Zuo, and Zhang 2018) presented a denoising network to process blind noise denoising. (Guo et al. 2019; Cao et al. 2021) attempted to remove the noise from real noisy images. There are also deep denoising algorithms trained without clean data supervision (Lehtinen et al. 2018; Krull, Buchholz, and Jug 2019; Huang et al. 2021). However, in extremely dark environments, fine texture details damaged by the high-intensity noise are very difficult to restore. In that case, denoising algorithms tend to generate over-smoothed outputs. By the way, low-light image enhancement algorithms (Chen et al. 2018; Lamba and Mitra 2021; Gu et al. 2019) try to directly restore high-quality images in terms of brightness, color, etc. However, these algorithms cannot deal with such high-intensity noise as well.

RGB-NIR Fusion. To obtain high-quality low-light images, researchers (Krishnan and Fergus 2009) try to fuse NIR images with RGB images. Recently, (Yan et al. 2013) pointed out the gradient inconsistency between RGB-NIR image pairs, and proposed Scale Map to try to solve it. Among the methods based on deep neural network, Joint Image Filtering with Deep Convolutional Networks (DJFR) (Li et al. 2019) constructs a unified two-stream network model for image fusion, CU-Net (Deng and Dragotti 2020) combines sparse encoding with Convolutional Neural Networks (CNNs), DKN (Kim, Ponce, and Ham 2021) explicitly learns sparse and spatially-variant kernels for image filtering. (Lv et al. 2020) innovatively constructs a network that directly decouples RGB and NIR signals for 24-hour imaging. In general, current RGB-NIR fusion algorithm has two main problems: insufficient ability to deal with RGB-NIR texture inconsistency, leading to heavy artifacts on the final fusion images, inadequate noise suppression capability especially when dealing with high-intensity noise in extremely low-light environments.

Datasets. There is only a small amount of data that can be used for RGB-NIR fusion studies because of the difficulty to obtain aligned RGB-NIR image pairs. Some studies (Foster et al. 2006) focus on obtaining hyperspectral datasets, and strictly aligned RGB-NIR image pairs can be obtained by integrating hyperspectral images on the corresponding band. (Krishnan and Fergus 2009) present a prototype camera to collect RGB-NIR image pairs. However, these datasets are too small to be used to comprehensively measure the performance of fusion algorithms. More importantly, due to the lack of data on actual scenarios, they cannot encourage follow-up researchers to focus on the valuable problems that RGB-NIR will encounter in applications.

Approach

Prior Knowledge of Structure Inconsistency

As previously described, the network needs to be aware of the inconsistent regions on the two inputs. We design an intuitive function to measure the inconsistency from image features. Firstly binary edge maps are extracted from each feature channel. Then the inconsistency is defined as

$$\mathcal{F}(edge^{C_:}, edge^{N}) = \lambda(1 - edge^{C_:})(1 - edge^{N}) + edge^{C_:} \cdot edge^{N} \quad (1)$$

where $C_: \in \mathbb{R}^{H\times W}$ and $N \in \mathbb{R}^{H\times W}$ denote R/G/B channel of the clean RGB image and NIR image, $edge^{C_:}$ and $edge^{N}$ respectively represent the binarized edge maps of $C_:$ and N, which is obtained by binarizing its mean value as a threshold after Sobel filtering.

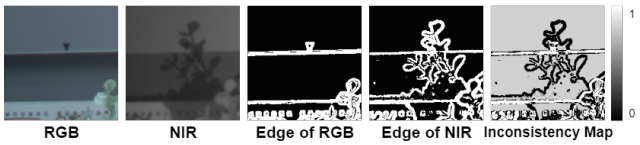

Figure 3: Through applying $\mathcal{F}$ on edge maps of clean RGB and NIR images, the calculated inconsistency map clearly shows the structure inconsistency between RGB and NIR.

As shown in Figure 3, $\mathcal{F}(\cdot,\cdot)$ equals to 0 in the regions where $edge^{C_:}$ and $edge^{N}$ shows severe inconsistency. On the contrary, $\mathcal{F}(\cdot,\cdot)$ equals to 1 in the regions where the structures of RGB and NIR are consistent. And in other regions, $\mathcal{F}(\cdot,\cdot)$ is set to a hyperparameter $\lambda(0 < \lambda < 1)$, indicating that there is no significant inconsistency. Utilising the output inconsistency map of $\mathcal{F}$, the inconsistent NIR structures can be easily suppressed by a direct multiplication.

Extraction of Deep Structures

Even though function $\mathcal{F}$ subtly describes the inconsistency between RGB and NIR images, it cannot be applied directly in extremely low light cases. As shown in Figure 4, the calculated inconsistency map contains nothing but non-informative noise when facing extremely noisy RGB image. To avoid the influence of noise in the structure inconsistency extraction, we propose the Deep Structure Extraction Module (DSEM) and Deep Inconsistency Prior (DIP), where we compute the structure inconsistency in feature space. Considering the processing flow of RGB and NIR are basically the same, we give a unified description here to keep the symbols concise.

The detailed architecture of DSEM is shown in Figure 5(a). DSEM takes multi-scale features $feat_i$ (i represents scale) from restoration network R and outputs multi-scale deep structure maps $struct_i$. In order for DSEM to predict high-quality deep structure maps, we introduce a clear supervision signal $struct_i^{gt}$ (addressed later) for DSEM and the training loss is calculated as:

$$\mathcal{L}_{stru} = \sum_{i=1}^{3}\sum_{c=1}^{Ch_i} Dist(struct_{i,c}, struct_{i,c}^{gt}), \quad (2)$$

where Ch_i is the channel number of the deep structures in the ith scale, $Dist$ is Dice Loss (Deng et al. 2018), $struct_{i,c}$ is the cth channel of the predicted deep structures in the ith scale and $struct_{i,c}^{gt}$ is the corresponding ground-truth. The Dice loss is given by $Dist(P,G) = (\sum_j^N p_j^2 + \sum_j^N g_j^2)/(2\sum_j^N p_j g_j)$, where p_j, g_j is the value of the jth pixel on the predicted structure map P and ground-truth G.

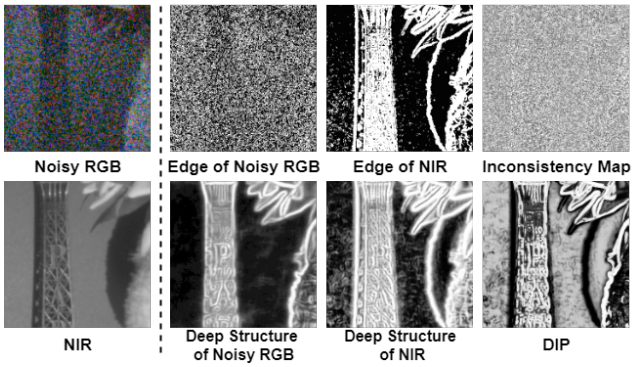

Figure 4: Applying $\mathcal{F}$ on the edge maps of noisy RGB and NIR images can only get a meaningless inconsistency map due to the heavy noise in the input RGB (as the first row shows). However, the deep structure map predicted by DSEM is very clear and the calculated DIP effectively describes the inconsistent structures (as the second-row shows). See more examples in supplementary material.

Supervision of DSEM Considering that it is almost impossible for DSEM to naturally output feature maps that only contain structural information, we have to introduce a clear supervision for the output of DSEM to predict high-quality deep structure maps. The supervision signal is set up following the idea of Deep Image Prior (Ulyanov, Vedaldi, and Lempitsky 2018) and $struct_{i,c}^{gt}$ is acquired from a pretrained AutoEncoder (Hinton and Salakhutdinov 2006) network AE [1]. The base architecture of AE is exactly the same as R with skip connections removed, as Figure 5(b) shows. Multi-scale decoder features $dec_{i,c}$ are extracted from the pretrained AutoEncoder network AE and the supervision signal is calculated by:

$$struct_{i,c,j}^{gt} = \begin{cases} 0 & \text{if } (\nabla dec_{i,c,j} - m_{\nabla dec_{i,c}}) <= 0, \\ 1 & \text{if } (\nabla dec_{i,c,j} - m_{\nabla dec_{i,c}}) > 0. \end{cases} \quad (3)$$

where $struct_{i,c,j}^{gt}$ is the jth pixel of $struct_{i,c}^{gt}$, ∇ represents the Sobel operator, $\nabla dec_{i,c,j}$ is the jth pixel in $\nabla dec_{i,c}$ and $m_{\nabla dec_{i,c}}$ is the global average pooling result of $\nabla dec_{i,c}$. The supervision signal obtained by this design effectively trains the DSEM and clear deep structure maps are predicted as shown in Figure 4.

Calculation of DIP and Image Fusion

The extracted deep structures contain rich structure information and are robust to noise. With $struct_i$ of the noisy RGB and NIR images, we can introduce inconsistency function $\mathcal{F}$ to obtain high-quality knowledge of structure inconsistency:

$$M_{i,c}^{DIP} = \mathcal{F}(struct_{i,c}^{C}, struct_{i,c}^{N}) \quad (4)$$

where $C \in \mathbb{R}^{H \times W \times 3}$ and $N \in \mathbb{R}^{H \times W}$ denote the noisy RGB image and NIR image, $struct_{i,c}$ is the cth channel of the features from the ith scale and $M_{i,c}^{DIP}$ is the corresponding inconsistency measurement. Since $M_{i,c}^{DIP}$ represents the structure inconsistency instead of intensity inconsistency, we apply $M_{i,c}^{DIP}$ directly to $struct_{i,c}^{N}$ instead of $feat_{i,c}^{N}$ in the form of:

$$\hat{struct}_{i,c}^{N} = M_{i,c}^{DIP} \cdot struct_{i,c}^{N}. \quad (5)$$

Under the guidance of the DIP, $\hat{struct}_{i,c}^{N}$ discards the structures that are inconsistent with RGB, thus empowering the deep features with prior knowledge to tackle structure inconsistency. As we will show in the experiments later, inconsistent structures in the NIR structure maps can be significantly suppressed after multiplying with $M_{i,c}^{DIP}$.

To further fuse the rich details contained in $\hat{struct}_{i,c}^{N}$ into RGB features, we designed a multi-scale fusion module as shown in Figure 5(c). As pointed out in (Jung, Zhou, and Feng 2020), denoising first and fusion later can improve the fusion quality. So we follow (Jung, Zhou, and Feng 2020) to reuse the denoised output of the restoration network R set up for noisy RGB as the input of the multi-scale fusion module.

Loss Function

The total loss function we used is formulated as:

$$\mathcal{L} = \mathcal{L}_{rec}^{C} + \mathcal{L}_{rec}^{\hat{C}} + \mathcal{L}_{rec}^{N} + \lambda_1 \cdot \mathcal{L}_{stru}^{C} + \lambda_2 \cdot \mathcal{L}_{stru}^{N} \quad (6)$$

where $\mathcal{L}_{stru}^{C}$ and $\mathcal{L}_{stru}^{N}$ are the loss function for RGB/NIR deep structures prediction, which is described above. λ_1 and λ_2 are the corresponding coefficients and set to $1/1000$ and $1/3000$. $\mathcal{L}_{rec}^{C}$, $\mathcal{L}_{rec}^{\hat{C}}$ and $\mathcal{L}_{rec}^{N}$ represent the reconstruction loss of fused-RGB/coarse-RGB/NIR image respectively. All of them are Charbonnier loss (Charbonnier et al. 1994) in the form of:

$$\mathcal{L}_{rec} = \sqrt{\|\mathbf{X} - \mathbf{X}_{gt}\|^2 + \varepsilon^2} \quad (7)$$

where $\mathbf{X}$ and $\mathbf{X}_{gt}$ represent the network output and the corresponding supervision. The constant ε is set to 10^{-3} empirically.

Experiment

Datasets

Data Collection. In order to obtain the aligned RGB-NIR image pairs in the easiest and direct way, we collect all RGB-NIR image pairs by switching an optical filter placed directly in front of the camera without an IR-Cut, and we divide them into two types of image pairs for different usages. We collect **reference image pairs** in normal-light environments. In order to obtain high-quality references, multiple still captures are averaged to remove noise and a matching algorithm (DeTone, Malisiewicz, and Rabinovich 2017) with manual double-check is applied to ensure the alignment of image pairs. In the following experiments, we add synthetic noise to these references to quantitatively evaluate the performance of fusion algorithms. To facilitate training, the collected images are cropped into 256*256 image patches. We collect **real noisy image pairs** of 1920*1080 pixels in low-light environments. The post-processing steps are the same as those used in collecting references image pairs, except that multi-frame average is not performed to noisy RGB

[1] See the training details in supplementary material.

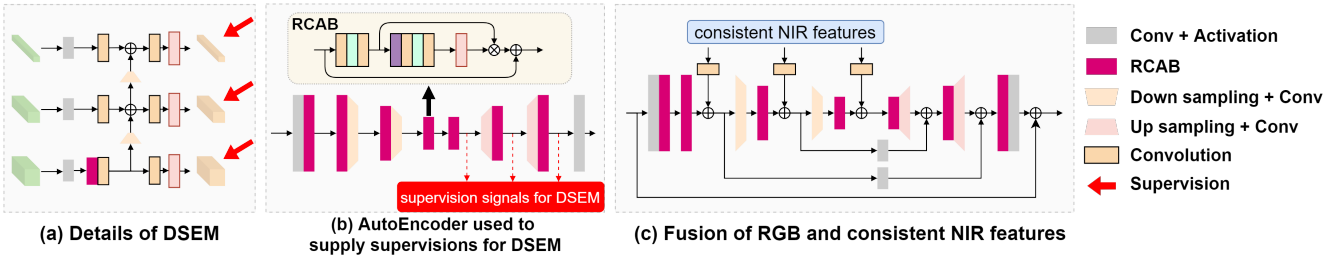

Figure 5: (a) Illustration of the Deep Structure Extraction Module (DSEM) details.(b) The architecture of the AutoEncoder network which is employed to provide supervision signals for DSEM. (c) The detailed fusion process of RGB and DIP-weighted NIR features, *i.e.* the third stage of DVN. Residual channel attention blocks (RCABs) (Zhang et al. 2018) are used to extract features.

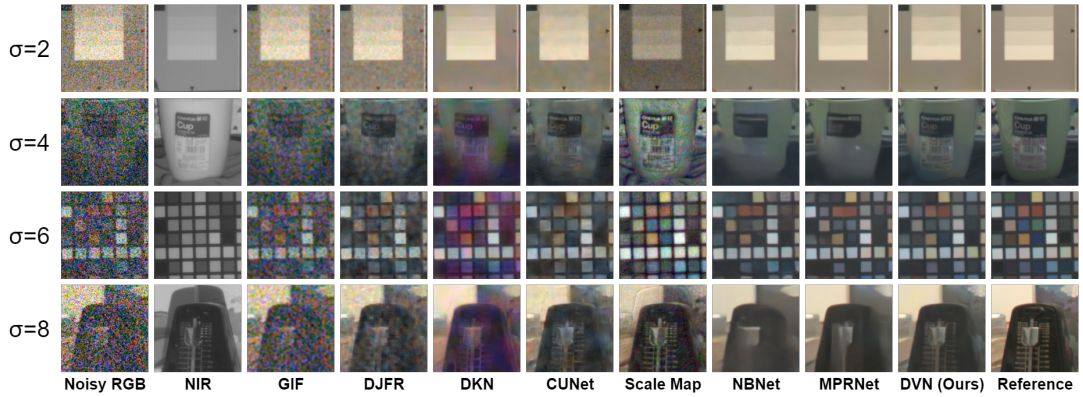

Figure 6: Fusion examples from DVD. The proposed DVN shows great superiority than other algorithms. Images are brightened for visual convenience. See supplementary material for more examples.

images. In the following experiments, we use these noisy image pairs to qualitatively evaluate the performances of fusion algorithms in handling real low-light images.

Dataset for Experiment. To make the synthetic data closer to the real images, we follow (Wang et al. 2020) to add synthetic noise to reference image pairs for training and testing. Considering that all the comparison methods use sRGB (standard Red Green Blue) images as input, we convert the collected raw data into sRGB through a simple isp-pipeline (Gray World for white balance, Gamma correction, Demosaicing) (Karaimer and Brown 2016), to make a fair comparison. In the following experiments, we use 5k reference image pairs (256*256) as the training set. Another 1k reference image pairs (256*256) along with 10 additional real noisy image pairs (1920*1080) are used for testing.

Implementation Details

All experiments are conducted on a device equipped with two 2080-Ti GPUs. We train the proposed DVN from scratch in an end-to-end fashion. Batchsize is set to 16. Training images are randomly cropped in the size of 128*128, and the value range is [0, 1]. We augment the training data following MPRNet (Zamir et al. 2021), including random flipping and rotating. Adam optimizer with momentum terms (0.9, 0.999) is applied for optimization. The whole network is trained for 80 epochs, and the learning rate is gradually reduced from 2e-4 to 1e-6. λ in function $\mathcal{F}$ is set to 0.5 for all configurations. The AutoEncoder network used to provide supervision signals for DSEM is pretrained in the same way, except that it only trained for 5 epoches and the input is clean RGB and NIR images separately. See supplementary material for more training details.

The synthesis of low-light data for training includes two steps. The first step is to reduce the average value of raw images taken under normal light to 5 (10-bit raw data). The second step is to add noise to the pseudo-dark raw images, including Gaussian noise with variance equals to σ, and Poisson noise with a level proportional to σ.

Performance Comparison

Results on DVD Benchmark. We evaluate and compare DVN with representative methods in related fields, including single-frame noise reduction algorithms NBNet (Cheng et al. 2021) and MPRNet (Zamir et al. 2021), joint image filtering algorithms GIF (He, Sun, and Tang 2012), DJFR (Li et al. 2019), DKN (Kim, Ponce, and Ham 2021) and CUNet (Deng and Dragotti 2020), as well as Scale Map (Yan et al. 2013) which specially designed for RGB-NIR fusion. All methods are trained or finetuned on DVD from scratch. We use PSNR and SSIM (Wang et al. 2004) for quantitative measurement. Qualitative comparison is shown in Figure 6, and quantitative comparison under different noise intensity settings ($\sigma = 2, 4, 6, 8$, the larger the σ, the heavier the noise) on DVD benchmark is shown in Table 1.

		GIF	DJFR	DKN	Scale Map	CUNet	NBNet	MPRNet	DVN (Ours)
$\sigma = 2$	PSNR	22.32	26.28	27.22	21.98	28.62	31.38	**31.79**	*31.50*
	SSIM	0.6410	0.8263	0.8902	0.6616	0.9138	0.9477	*0.9504*	**0.9551**
$\sigma = 4$	PSNR	19.15	23.91	24.34	21.02	26.81	29.14	*29.37*	**29.62**
	SSIM	0.5033	0.7464	0.8427	0.6225	0.8832	0.9259	*0.9276*	**0.9400**
$\sigma = 6$	PSNR	17.30	22.40	22.78	20.02	25.43	27.27	*27.68*	**28.26**
	SSIM	0.4240	0.6802	0.8067	0.5959	0.8510	0.9060	*0.9083*	**0.9273**
$\sigma = 8$	PSNR	15.98	20.72	22.50	19.07	23.75	24.81	*26.20*	**26.98**
	SSIM	0.3701	0.6177	0.7799	0.5742	0.8154	0.8822	*0.8908*	**0.9155**

Table 1: The PSNR (dB) and SSIM results of different algorithms on DVD dataset. The best and second best results are highlighted in bold and Italic respectively.

PSNR comparison on public IVRG dataset (σ=50, input PSNR = 13.44)		PSNR comparison with other methods on DVD ($\sigma = 4$)	
DJFR	23.35	SID (Chen et al. 2018)	25.26
CUNet	24.96	SGN (Gu et al. 2019)	28.40
Scale Map	25.59	SSN (Dong et al. 2018)	13.72
DVN (Ours)	30.43	DVN (Ours)	29.62

Table 2: Performance comparison (PSNR). The conclusions are the same if SSIM is applied as the metric.

The qualitative comparison in Figure 6 clearly illustrates the superiority of the proposed DVN on noise removal, detail restoration and visual artifacts suppression. In contrast, image denoising algorithms (*i.e.* NBNet and MPRNet) cannot restore texture details when the noise intensity becomes high, and the output turns into pieces of smear even though the noise is effectively suppressed. GIF and DJFR output images with heavy noise as the 3rd and 4th column in Figure 6 shows, which greatly affects the fusion quality. The fusion effect of DKN and CUNet (5th and 6th column in Figure 6) under mild noise (*e.g.* $\sigma = 2$) is acceptable. But under heavy noise, obvious color deviation appears in the DKN output, and neither of them can deal with structure inconsistency (see the 4th row in Figure 6), resulting in severe artifacts in the fusion images. Scale Map outputs images with rich details. However, it cannot reduce the noise in the areas where texture is lacking in the NIR image. In addition, it is hard to achieve a balance between noise suppression and texture migration when applying Scale Map.

Generalization on Real Noisy RGB-NIR. To evaluate the performance of algorithms when facing *real* low-light images, we conduct a qualitative experiment on several pairs of RGB-NIR images captured in real low-light environments. As shown in Figure 7, outputs of DVN have obviously low noise, rich details, and are visually more natural when handling RGB-NIR pairs with *real* noise, even if the network is trained on a synthetic noisy dataset.

Comparison on Public Dataset. So far, there is no **high-quality** public RGB-NIR dataset like DVD yet. For example, RGB-NIR pairs in IVRG (Brown and Süsstrunk 2011) are not well aligned. Even so, we retrained DVN and other methods on IVRG and give quantitative comparison in Table 2. It is clear that DVN still performs well.

Comparison with Low-Light Enhancement Methods. We also compare our method with the low-light enhancement methods. We retrained SID (Chen et al. 2018) and SGN (Gu et al. 2019), the comparison can be seen in Table 2. It is clear that our proposed DVN still shows great superiority.

Effectiveness of DIP

In this section, we verify that the proposed DIP is effective in handling the mentioned structure inconsistency. For comparison, we retrain a baseline, which is the same as the proposed DVN only without the DIP module. As Figure 8(a) shows, the NIR shadow of the grass still remains in the fusion result without DIP, but not in the fusion result with DIP. This directly proves that DIP can handle the structure inconsistency. Figure 8(b) shows that DIP can also deal with serious structure inconsistency caused by the misalignment between RGB-NIR images to a certain extent (this example pair cannot be aligned even after image registration). This has practical value because the problem of misalignment frequently occurs in applications. Taking into account the nature of DIP, the remaining artifacts are in line with expectations, since they are concentrated near the pixels with gradients in the RGB image.

In addition, Figure 8 also visualizes the deep structure of RGB, NIR, consistent NIR (DIP-weighted) as well as DIP Maps. It is obvious that even facing noisy input, the RGB deep structure still contains clear structures. The visual comparison between the NIR deep structure and the **consistent** NIR deep structure proves that the introduction of DIP can handle structure inconsistency in deep feature space.

Ablation Study

We evaluate the effectiveness of each component in the proposed algorithm on the DVD benchmark quantitatively in this section. PSNR and SSIM are reported in Table 3. The baseline network directly fuse NIR features with RGB features (row 1 in Table 3).

Intermediate supervision $\mathcal{L}_{rec}^{\hat{C}}$ and $\mathcal{L}_{rec}^{N}$ effectively improve the performance as Table 3 (row 1 and 2) shows. This

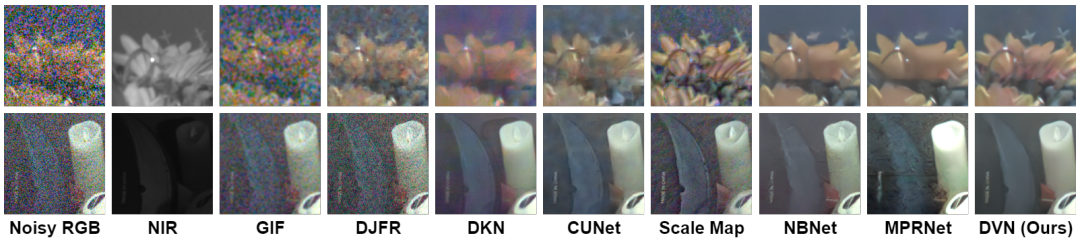

Figure 7: Fusion results on RGB-NIR image pairs with *real* noise. DVN obviously obtains better results than other algorithms.

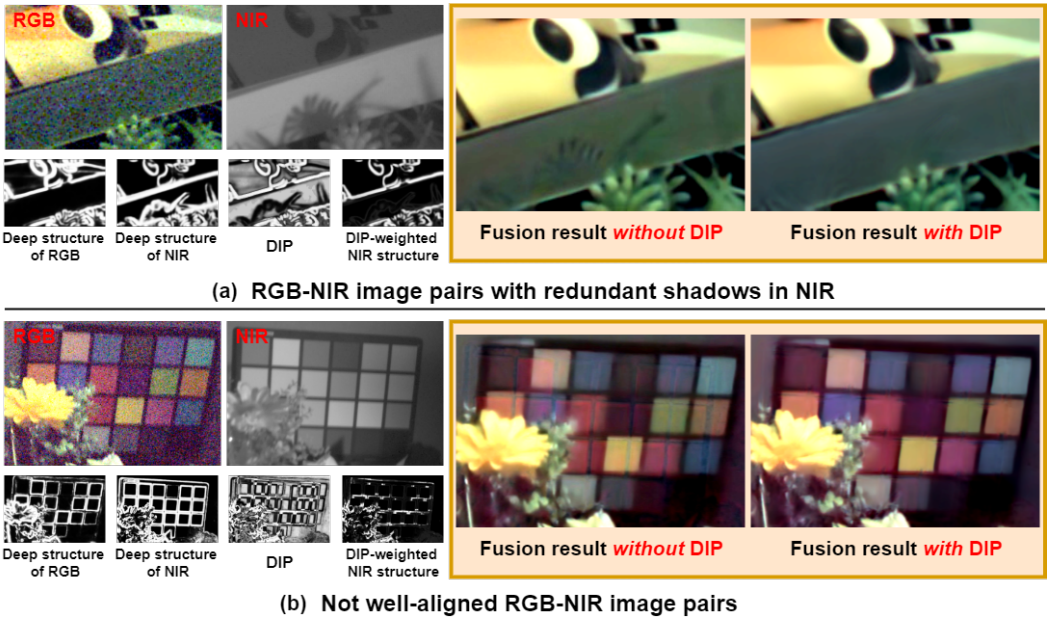

Figure 8: Illustration of the effectiveness of DIP. (a) shows a typical case of structure inconsistency caused by NIR shadows and (b) shows a case of the misalignment RGB-NIR. Fusion results and visualizations of deep structures verify the effectiveness of the DIP. Both examples are gathered from real noisy image pairs.

row.	$\mathcal{L}_{rec}^{\hat{C}} + \mathcal{L}_{rec}^{N}$	DSEM	DIP	PSNR	SSIM
1.	–	–	–	28.87	0.9356
2.	✓	–	–	29.30	0.9375
3.	–	✓	–	29.06	0.9376
4.	✓	✓	–	29.36	0.9358
5.	✓	✓	✓	**29.62**	**0.9400**

Table 3: Ablation experiment results are conducted on DVD to study the effectiveness of each component. σ is set to 4.

indicates the necessity of enhancing the noise suppression capability of the network for clean structure extraction.

Applying DSEM to learn deep structures without DIP can improve performance as well as Table 3 (row 1 and 3) shows. However, since the inconsistent structures are not removed, the benefits are not obvious, even if we use intermediate supervision and DSEM simultaneously as row 4 shows.

As Table 3 (row 5) shows, after introducing DIP to deal with the structure inconsistency, the network performance can be further improved by a large margin. This demonstrates the effectiveness of our proposed algorithm and the necessity to focus on the structure inconsistency problem on RGB-NIR fusion problem.

Conclusion

In this paper, we propose a novel RGB-NIR fusion algorithm called Dark Vision Net (DVN). DVN introduces Deep inconsistency prior (DIP) to integrate the structure inconsistency into the deep convolution features, so that DVN can obtain a high-quality fusion result without visual artifacts. In addition, we also proposed the first available benchmark, which is called Dark Vision Dataset (DVD), for RGB-NIR fusion algorithms training and evaluation. Quantitative and qualitative results prove that the DVN is significantly better than other algorithms.

Acknowledgements

This paper is supported by the National Key R&D Plan of the Ministry of Science and Technology (Project No. 2020AAA0104400) and Beijing Academy of Artificial Intelligence (BAAI).

References

Brown, M.; and Süsstrunk, S. 2011. Multi-spectral SIFT for scene category recognition. In *CVPR 2011*, 177–184. IEEE.

Cao, Y.; Wu, X.; Qi, S.; Liu, X.; Wu, Z.; and Zuo, W. 2021. Pseudo-ISP: Learning Pseudo In-camera Signal Processing Pipeline from A Color Image Denoiser. *arXiv preprint arXiv:2103.10234*.

Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; and Barlaud, M. 1994. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing*, volume 2, 168–172. IEEE.

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.

Cheng, S.; Wang, Y.; Huang, H.; Liu, D.; Fan, H.; and Liu, S. 2021. NBNet: Noise Basis Learning for Image Denoising with Subspace Projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4896–4906.

Deng, R.; Shen, C.; Liu, S.; Wang, H.; and Liu, X. 2018. Learning to predict crisp boundaries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 562–578.

Deng, X.; and Dragotti, P. L. 2020. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE transactions on pattern analysis and machine intelligence*.

DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2017. SuperPoint: Self-Supervised Interest Point Detection and Description. *CoRR*, abs/1712.07629.

Foster, D. H.; Amano, K.; Nascimento, S. M.; and Foster, M. J. 2006. Frequency of metamerism in natural scenes. *Josa a*, 23(10): 2359–2372.

Gu, S.; Li, Y.; Gool, L. V.; and Timofte, R. 2019. Self-guided network for fast image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2511–2520.

Guo, S.; Yan, Z.; Zhang, K.; Zuo, W.; and Zhang, L. 2019. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1712–1722.

He, K.; Sun, J.; and Tang, X. 2012. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6): 1397–1409.

Hinton, G. E.; and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *science*, 313(5786): 504–507.

Huang, T.; Li, S.; Jia, X.; Lu, H.; and Liu, J. 2021. Neighbor2Neighbor: Self-Supervised Denoising from Single Noisy Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14781–14790.

Jung, C.; Zhou, K.; and Feng, J. 2020. FusionNet: Multispectral fusion of RGB and NIR images using two stage convolutional neural networks. *IEEE Access*, 8: 23912–23919.

Karaimer, H. C.; and Brown, M. S. 2016. A software platform for manipulating the camera imaging pipeline. In *European Conference on Computer Vision*, 429–444. Springer.

Kim, B.; Ponce, J.; and Ham, B. 2021. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, 129(2): 579–600.

Krishnan, D.; and Fergus, R. 2009. Dark flash photography. *ACM Trans. Graph.*, 28(3): 96.

Krull, A.; Buchholz, T.-O.; and Jug, F. 2019. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2129–2137.

Lamba, M.; and Mitra, K. 2021. Restoring Extremely Dark Images in Real Time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3487–3497.

Lehtinen, J.; Munkberg, J.; Hasselgren, J.; Laine, S.; Karras, T.; Aittala, M.; and Aila, T. 2018. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*.

Li, Y.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2019. Joint image filtering with deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1909–1923.

Lucas, A.; Iliadis, M.; Molina, R.; and Katsaggelos, A. K. 2018. Using deep neural networks for inverse problems in imaging: beyond analytical methods. *IEEE Signal Processing Magazine*, 35(1): 20–36.

Lv, F.; Zheng, Y.; Li, Y.; and Lu, F. 2020. An integrated enhancement solution for 24-hour colorful imaging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 11725–11732.

Mao, X.; Shen, C.; and Yang, Y.-B. 2016. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in neural information processing systems*, 29: 2802–2810.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9446–9454.

Wang, Y.; Huang, H.; Xu, Q.; Liu, J.; Liu, Y.; and Wang, J. 2020. Practical deep raw image denoising on mobile devices. In *European Conference on Computer Vision*, 1–16. Springer.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

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.

Yan, Q.; Shen, X.; Xu, L.; Zhuo, S.; Zhang, X.; Shen, L.; and Jia, J. 2013. Cross-field joint image restoration via scale map. In *Proceedings of the IEEE International Conference on Computer Vision*, 1537–1544.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14821–14831.

Zhang, K.; Zuo, W.; and Zhang, L. 2018. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, 27(9): 4608–4622.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, 286–301.