Real-Time Restoration of Dark Stereo Images

Mohit Lamba M V A Suhas Kumar Kaushik Mitra
Indian Institute of Technology Madras

Abstract

Low-light image enhancement has been an actively researched area for decades and has produced excellent night-time single-image, video, and Light Field restoration methods. Despite these advances, the problem of extreme low-light stereo image restoration has been mostly ignored and addressing it can enable night-time capabilities to several applications such as smartphones and self-driving cars. We propose an especially light-weight and fast hybrid U-net architecture for extreme low-light stereo image restoration. In the initial few scale spaces, we process the left and right features individually, because the two features do not align well due to large disparity. At coarser scale-spaces, the disparity between left and right features decreases and the network's receptive field increases. We use this fact to reduce computations by simultaneously processing the left and right features, which also benefits epipole preservation. As our architecture does not use any 3D convolution for fast inference, we use a Depth-Aware loss module to train our network. This module computes quick and coarse depth estimates to better enforce the stereo epipolar constraints. Extensive benchmarking in terms of visual enhancement and downstream depth estimation shows that our architecture not only restores dark stereo images faithfully but also offers $4-60\times$ *speed-up with* $15-100\times$ *lower floating point operations, necessary for real-world applications.*

1. Introduction

The low-light enhancement community has witnessed the development of state-of-the-art algorithms on restoring high-quality images captured in extremely dark conditions. Years of research in this area has exploited several techniques ranging from histogram equalization [29, 50, 58, 8, 22, 35] to retinex theory [34, 33, 61, 13, 14, 37, 17, 26], and more recently convolutional neural networks [6, 15, 16, 72, 60, 38, 63, 25, 36, 52, 75, 45]. These advances have enabled several night-time applications that were previously limited to only daylight conditions, such as object detection [54, 10], semantic segmentation [64, 53], saliency detection [68, 69], or even casual photography [6, 15]. Seeing the overwhelming success of these methods for single image enhancement, many researchers have extended them for night-time video [24, 5] and Light Field restoration [32]. Although the area of low-light restoration has been studied quite extensively, an important gap still exists, which is the restoration of *extremely dark stereo images*. Filling in this gap would benefit several night-time applications that need to incorporate 3D information from the surrounding world. For example, today most self-driving cars use LiDAR to get reliable depth estimates in low-light conditions. At the same time cameras are inevitable for other ADAS-oriented tasks such as lane detection and pedestrian identification. However, if high-quality restoration can be done for low-light stereo images, the costly and bulky LiDAR may be removed for cost-efficient products. Other applications such as bokeh effect in smartphones and AR/VR headsets can similarly benefit from low-light stereo enhancement.

Leveraging existing monocular low-light methods is likely to bear sub-optimal results because they cannot harness the information present in the corresponding stereo pair and the epipolar geometry might be destroyed. Another option is to use existing stereo models, but they have been mostly optimised for depth predictions [56, 4, 27, 7, 74, 18, 71, 76] and do not visually enhance the RGB images. Also, they heavily rely on 3D convolutions which entail a huge computational burden. Consequently, recent super-resolution methods for well-lit stereo images have replaced expensive 3D convolutions with relatively cheaper attentions modules [59, 70, 73, 62, 67]. However, when we tried using them for dark stereo images they struggled to give any enhancement benefit due to the presence of acute noise and poor contrast in extreme low-light images.

We propose a hybrid U-net architecture (see Fig. 1) to restore dark stereo images in a way that not only benefits the perceptual quality of individual images but also preserves the epipolar geometry for downstream applications. Further our network is especially fast and light-weight which is necessary for real-world deployment. To benefit visual enhancement, our hybrid architecture independently processes the left and right views in the initial scale spaces because in these scale spaces the left/right features do not

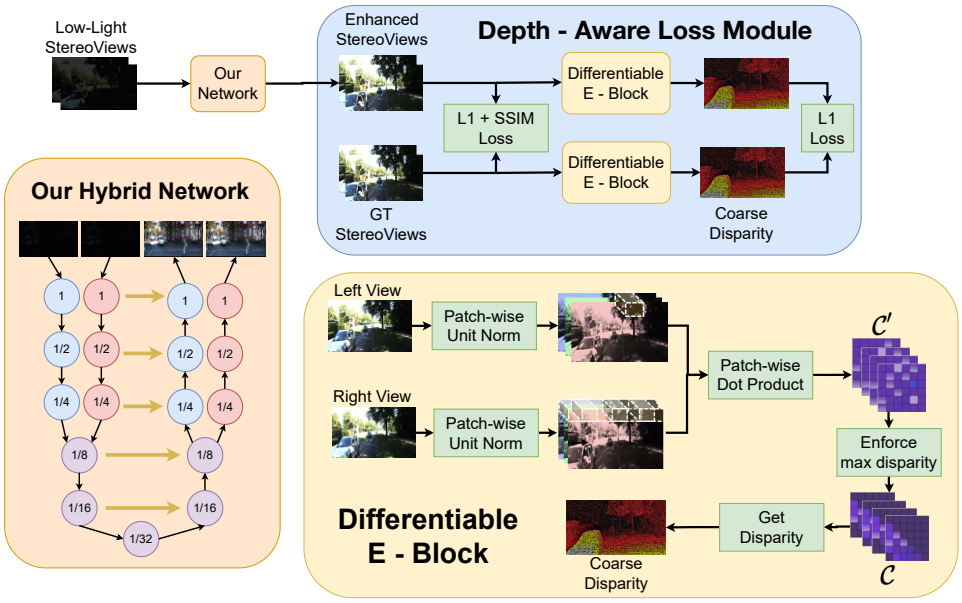


Figure 1. We propose an exceptionally light-weight & fast hybrid U-net architecture to restore dark stereo images in a way that not only benefits the visual perception of individual images but also respects the epipolar geometry for downstream applications. It is fast because it does not use any 3D convolutions and it is accurate because of the Depth-Aware Loss Module.

align well due to large disparity. But as the scale-space becomes coarser due to repeated downsampling, the view disparity decreases, and the receptive field of the network increases significantly. This allows us to process both views simultaneously for computational efficiency. We avoid using 3D convolutions or attention modules to keep the network fast. Though our proposed solution is quite simple, it is very effective and has been overlooked in the existing literature perhaps because the focus was on well-lit images.

The proposed hybrid architecture uses only 2D convolutions and so to better enforce the epipolar constraints, we train it using our Depth-Aware differentiable loss module. Our Depth-Aware loss module takes the restored stereo views and uses classical computer vision to compute the disparity in a differentiable manner. Likewise, after obtaining GT disparity from GT stereo views, the module computes an L1 loss between the two disparities. A naive approach would have used state-of-the-art depth from stereo model instead of our Depth-Aware module to enforce the geometric constraints. But this approach has two challenges: (a) back-propagation through the depth estimation models is computationally very expensive requiring multiple GPUs, and thus the primary task of training the enhancement network will suffer due to memory scarcity. Using existing depth estimation models will force the main enhancement network to use small batch/patch size which negatively affects the restoration quality. (b) almost every depth from stereo model available today has been optimized for either KITTI [43] or synthetic SceneFlow dataset [42]. Thus, they cannot be used *out-of-the-box* to train networks for images captured using any arbitrary stereo setup. To overcome these limitations, our Depth-Aware differentiable loss module uses classical computer vision for depth computation and has only one hyperparameter. It can thus be directly plugged in to train stereo enhancement models for any general stereo rectified setting. The Depth-Aware loss module is not designed to replace state-of-the-art stereo depth models, which if supplied a huge amount of training data, time and memory can deliver excellent depth estimates, but to offer quick and light-weight coarse level depth estimates sufficient for enforcing epipolar constraints during training. Our code[1] is available at `https://mohitlamba94.github.io/darkstereo/`

In summary, we make the following contributions:

- We aim for computationally light and high-speed restoration of extremely dark stereo images, which although an important problem, has been largely unexplored in the existing low-light enhancement literature.
- We propose a simple yet effective Hybrid U-net architecture for stereo image enhancement which compared to existing methods offers a good trade-off for visual restoration of stereo images, epipole preservation and real-time inference.
- Inspired by classical computer vision, we use an Depth-Aware differentiable loss module which can be

[1]This work was supported in part by *IITM Pravartak Technologies Foundation*.

used *out-of-the-box* for training stereo enhancement models for any arbitrary stereo rectified setting.

- Compared to exisitng methods our method offers $4-60\times$ speed-up with $15-100\times$ lower floating-point operations with restoration at par with computationally expensive methods.

2. Related Works

Low-light image enhancement. Early methods on low-light enhancement used histogram equalization to enhance the dynamic range [29, 50, 58, 8, 22, 35]. Later, people found that exploiting Retinex theory [34, 33] to decompose low-light images into illumination and reflectance components aided better enhancement [61, 13, 14, 37, 17, 26]. Now-a-days people use learning based networks for better low-light enhancement [16, 72, 60, 38, 63, 25, 36, 52, 75]. Chen *et al.* [6] proposed the famous SID dataset for extreme low-light image enhancement and the dataset has since then motivated several works on extreme low-light image enhancement [41, 15, 66, 1]. Most of these methods had a considerable computational overhead. Thus, few light-weight single image enhancement methods [31, 45] have also been proposed by slightly compromising visual enhancement.

Deep stereo models for different stereo applications. Stereo models have been used for wide variety of tasks such as depth estimation [4, 56, 27, 7, 18, 55, 74, 57, 65, 46, 3, 76, 71] and super resolution [59, 70, 73, 62, 67, 23]. Majority of depth models warp stereo features to generate a 4D cost volumes and then regress using 3D convolutions to compute disparity. Although these methods produce state-of-the-art results, using 3D convolution is computationally expensive. To alleviate this problem, recent stereo super resolution methods [59, 62] propose relatively cheaper attention modules. While attention modules have been beneficial for well-lit images, applying them to extremely noisy low-light images may confer similar improvements. Deep stereo models have also been used for stereo deblurring [70], correcting double refraction [28], and image compression [11] but the task of light-weight enhancement of extreme low-light stereo images has been barely studied. Only very recently DVEnet [21] was proposed that enhanced underexposed low-light stereo images. DVEnet, however, when used for extremly dark stereo images exhibits lot of halo artifacts (see Experimental section) and entails considerable computational overhead.

3. Real-Time Stereo Enhancement Network

A practical low-light stereo restoration method must simultaneously address three challenges: (a) noise suppression and color enhancement, (b) preserving the epipolar geometry of a wide baseline camera setup, and (c) low computational overhead required for real world applications. Keeping these constraints in mind, we propose a hybrid U-net architecture for real-time restoration of extremely dark stereo images. Most existing stereo methods first compute disparity between left/right views and then use this information for a specific task such as super-resolution and depth estimation. However, for extreme low-light stereo images, we first enhance the images using our Hybrid U-net and then enforce geometric constraints using the Depth-Aware Loss Module. We do so because very low-light images are too noisy with poor contrast and thus, directly retrieving the depth is prone to errors (see supplementary).

3.1. Network Architecture

Our hybrid U-net architecture accepts a pair of stereo rectified low-light images and outputs the restored stereo views. It is designed to enable each view harness the information present in the corresponding stereo view without using any computationally expensive 3D convolutions or attention mechanisms. Fig. 1 shows our proposed network, which operates at 6 scale spaces: $1, 1/2, 1/4, 1/8, 1/16$ and $1/32^{th}$ resolution of the input image. In the initial few scale spaces the stereo features do not align due to large disparity. We thus process them independently by running the convolutional kernels twice, once for each stereo feature. But as the feature dimensions reduces, due to repeated downsampling operations, the misalignment between the stereo features also reduces and the network's receptive field increases. For example, the maximum pixel disparity in the KITTI[43] dataset is between $200-256$ and for CityScape[9] it is even lower. So at the $1/8^{th}$ resolution scale space, the maximum pixel disparity will be $25-32$. But our network's receptive field just after the first convolutional layer in this scale space is 36×36. Thus, at $1/8^{th}$ resolution scale space we channel-wise concatenate the stereo features and process them jointly for the remaining scale spaces. This not only facilitates the exchange of information between the stereo features but also avoids doing repeated convolutions. To save computations, we allot more convolutional kernels to later scale-spaces and do not use too many convolutional kernels in the initial scale spaces. The network mostly depends on pixel-shuffle operation for down and upsampling features maps and uses LeakyReLU nonlinearity. More details about our architecture are present in the supplementary.

3.2. Depth-Aware Loss Module

Our hybrid U-net uses only 2D convolutions to achieve high-speed inference. We thus train it using the Depth-Aware Loss Module to better enforce the epipolar constraints. This module though required only for training, is kept light-weight to accommodate a bigger enhancement network. The module has two components, namely the photometric loss denoted by $\mathcal{L}_{ph}$ and the disparity consistency

loss denoted by $\mathcal{L}_{disp}$. $\mathcal{L}_{ph}$ computes the L1+SSIM loss between enhanced stereo views and the ground-truth (GT) stereo views. $\mathcal{L}_{disp}$ on the other hand computes the L1 loss between disparity obtained from the enhanced views and disparity obtained from the GT stereo views. Note that unlike many stereo methods our method does not require GT depth but only the GT stereo rectified RGB enhanced views for training. This is advantageous because cameras are cheap compared to LiDARs and aligning LiDAR and RGB data not straightforward. The overall loss function $\mathcal{L}$ can be thus summarised as:

$$\mathcal{L} \quad = \quad \mathcal{L}_{ph} + \lambda \cdot \mathcal{L}_{disp} \tag{1}$$

Photometric loss: Let $\mathbf{L}_{en}$ and $\mathbf{R}_{en}$ denote the enhanced left and right stereo views, and $\mathbf{L}_{GT}$ and $\mathbf{R}_{GT}$ denote the GT stereo views. Further, let $L_1(\cdot,\cdot)$ compute the difference between l_1 norm of the input tensors and $d_{ssim}(\cdot,\cdot) = 1 - SSIM(\cdot,\cdot)$. The photometric loss is thus computed as,

$$\begin{aligned} \mathcal{L}_{ph} &= 0.5 \cdot [L_1(\mathbf{L}_{GT}, \mathbf{L}_{en}) + L_1(\mathbf{R}_{GT}, \mathbf{R}_{en})] \\ + \quad & 0.5 \cdot [d_{ssim}(\mathbf{L}_{GT}, \mathbf{L}_{en}) + d_{ssim}(\mathbf{R}_{GT}, \mathbf{R}_{en})] \end{aligned} \tag{2}$$

Differentiable E - Block: Given a pair of stereo rectified views $\mathbf{L} \in \mathbb{R}^{H \times W \times 3}$ and $\mathbf{R} \in \mathbb{R}^{H \times W \times 3}$, the E - Block computes the disparity between the two views in a way that allows back-propagation through it.

Given any pixel in the left view $\mathbf{L}$ we construct a $M \times M$ (in this work $M = 31$) patch around it and compute a unit normalised dot product with every other patch of same dimension in the right stereo view along the horizontal epipolar line. In this way we construct $\mathcal{C}' \in \mathbb{R}^{H \times W \times W}$, such that each entry $\mathcal{C}'_{i,j,k}$ in $\mathcal{C}'$ is computed as:

$$\begin{aligned} \mathcal{C}'_{i,j,k} \quad = \quad & \frac{\mathbf{LP}}{||\mathbf{LP}||_1} \bullet \frac{\mathbf{RP}}{||\mathbf{RP}||_1} \\ & \forall\, i \in [1, H] \text{ and } \forall\, j, k \in [1, W] \end{aligned} \tag{3}$$

where, $\mathbf{LP} \in \mathbb{R}^{M \times M \times 3}$ is a patch around the pixel (i, j) in $\mathbf{L}$ and $\mathbf{RP} \in \mathbb{R}^{M \times M \times 3}$ is a patch around the pixel (i, k) in $\mathbf{R}$. However, computing $\mathcal{C}'_{i,j,k}$ can be computationally expensive and we prefer a lighter operation to not compromise the training of our hybrid U-net architecture. For example, if we ignore the unit normalisation step for simplicity, each $\mathcal{C}'_{i,j,k}$ requires at least $3M^2$ multiplications. We however found that more than the size of the chosen path, it is the context which matters more. We thus introduce a dilation term d, wherein every d^{th} rows and columns of the patch are considered to compute $\mathcal{C}'_{i,j,k}$. This way only $3\left(\frac{M}{d}\right)^2$ multiplications are required. In this work we set $d = 3$ and experiment with this idea during the ablation studies.

$\mathcal{C}'$ now has all the information required for computing a quick and coarse level disparity. The disparity for any pixel (i, j) in $\mathbf{L}$ is computed as,

$$j - k', \text{where } k' = \underset{k}{argmax}(\mathcal{C}'_{i,j,k}) \tag{4}$$

Though $argmax$ conventionally does not allow back-propagation, we make it differentiable by forcing the gradients through (i, j, k') in $\mathcal{C}'$ as 1 and all other gradient to 0. This simple workaround is also sometimes used to make the *maxpooling* layer differentiable. More sophisticated methods like SGM [20] additionally enforce smoothness constraints which definitely help in obtaining finer disparities. But for the extreme low-light enhanced views, very fine textures are hard to recover and so we found coarse level disparity good enough to train our network. This not only avoids the challenges involved in making additional constraints differentiable but also keeps the operation computationally light. Once we have computed the disparity, we also compute a confidence map $\mathbf{C} \in \mathbb{R}^{H \times W}$ as follows,

$$\mathbf{C}_{i,j} \quad = \quad \begin{cases} 1 & \text{if } \mathcal{C}'_{i,j,k'} \geq \Psi \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where Ψ is the mean of all the entries in $\mathcal{C}'$ where *argmax* in Eq. 4 was obtained.

For most stereo setup we have a fair amount of idea on the maximum disparity, $disp_{max}$, required because information such as baseline and camera focal length is generally known. We incorporate this information to regularise $\mathcal{C}'$. Specifically, we define a new tensor $\mathcal{C} \in \mathbb{R}^{H \times W \times W}$,

$$\begin{aligned} \mathcal{C}_{i,j,k} \quad = \quad & \begin{cases} \mathcal{C}'_{i,j,k} & \text{if } 0 \leq j - k \leq disp_{max} \\ \text{invalid} & \text{otherwise} \end{cases} \\ & \forall i \in [1, H] \text{ and } j, k \in [1, W] \end{aligned} \tag{6}$$

and then use only the valid entries in $\mathcal{C}$ to compute the disparity and the confidence. For all our experiments we only vary this single hyperparameter $disp_{max}$. Further, in the ablation studies we show that in rare cases when $disp_{max}$ is not known, it can be set to $disp_{max} = \infty$ and the predicted disparity is still quite good.

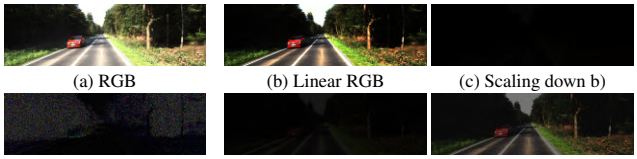
Disparity Consistency Loss: Let, $\mathbf{D}_{en}$ and $\mathbf{C}_{en}$ be the disparity and confidence map produced by the E - Block for the enhanced stereo views. Likewise, let $\mathbf{D}_{GT}$ and $\mathbf{C}_{GT}$ be the disparity and confidence map for the GT stereo views. The disparity consistency loss, $\mathcal{L}_{disp}$ is thus computed as,

$$\mathcal{L}_{disp} = L_1(\mathbf{D}_{GT} \cdot \mathbf{C}_{en} \cdot \mathbf{C}_{GT}, \mathbf{D}_{en} \cdot \mathbf{C}_{en} \cdot \mathbf{C}_{GT}) \tag{7}$$

4. Experiments

4.1. Experimental settings

Datasets: We assess the performance of our method on three publicly available datasets, namely KITTI [43], CityScape [9] and L3F [32].



(a) RGB (b) Linear RGB (c) Scaling down b)

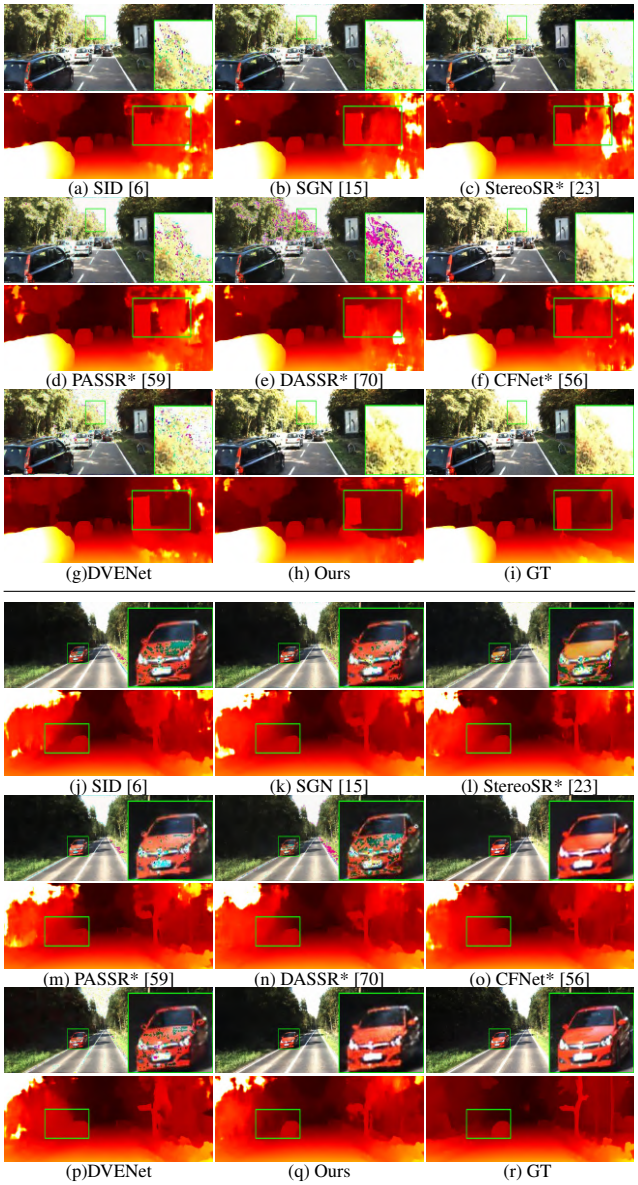
(d) Add Heteroscedastic noise (e) Low light sRGB (f)10x Amplified e)

Figure 2. Various steps involved in converting well-lit images into low-light.

There is no publicly available dataset on extreme low-light stereo enhancement for quantitative benchmarking. Thus following previous works, which faced a similar challenge [51, 48, 38, 19, 15, 70, 39], we transformed KITTI2015 well-lit stereo images into low-light images. But instead of naively adding Gaussian noise and darkening the image using gamma functions we follow a more principled approach for a realistic modelling [2]. A detailed description of our low-light modelling method is given in the supplementary and a brief overview can be found in Fig. 2. Low-light photon noise is added mainly during the image acquisition which is a linear space. But the processed images produced by cameras are sRGB images which is a non-linear space. We thus first go back to the linear RGB space, scale down the image, add the heteroscedastic noise and then come back to the sRGB space. The KITTI dataset contains 400 pairs of 1240×376 stereo images with corresponding ground-truth (GT) LiDAR depth map. We used 200 pairs for training and reserved 200 for testing. During low-light conversion, for each stereo pair we chose, $erms \sim \mathcal{N}(2, 0.01)$, $gain \sim \mathcal{N}(2, 0.01)$, $QE \sim \mathcal{U}(0.55, 0.66)$ and $scale = 1/40$. Refer the supplementary for details. To compute depth from the enhanced stereo views we used LEAStereo [7] which at the time of writing the paper holds the top position in the KITTI2015 stereo leaderboard amongst the published works. To further assess the performance of our method, we repeated this procedure for the CityScape dataset. The CityScape dataset has 5000 pairs of 1024×2048 stereo images, of which 500 were reserved for testing and remaining for training. CityScape does not have LiDAR depth maps, and uses SGM [20] to provide GT depth. Thus, we also use SGM for computing depth from enhanced low-light CityScape images.

We tried using other stereo datasets such as the Oxford Robot Car dataset [40] but it lacks the GT enhanced stereo images, and so cannot be used for benchmarking. In Sec. 4.4 we, however, show qualitative results on *real* extreme low-light stereo images obtained from the L3F Light Field dataset and real night time stereo images captured by us.

Comparison with other methods: We compare our method with SID [6], SGN [15], StereoSR [23], PASSR[59], DASSR[70], CFnet[56], and DVEnet [21]. SID and SGN were proposed for monocular extreme low-light enhancement and noise suppression. We used them to enhance left/right views individually. As there are no methods on extreme low-light stereo enhancement, we compare with DASSR, which has shown good performance for denoising additive white Gaussian noise added to well-lit stereo images. DASSR in stage-I obtains features independently for left/right view, in stage-II computes disparity to warp the features and in stage-III restores only the left view. We thus replicated stage-II and stage-III to get both left and right low-light enhanced views. We also com-



(a) SID [6] (b) SGN [15] (c) StereoSR* [23]

(d) PASSR* [59] (e) DASSR* [70] (f) CFNet* [56]

(g)DVENet (h) Ours (i) GT

(j) SID [6] (k) SGN [15] (l) StereoSR* [23]

(m) PASSR* [59] (n) DASSR* [70] (o) CFNet* [56]

(p)DVENet (q) Ours (r) GT

Figure 3. The figure shows the left view enhanced by different methods, and the depth computed by LEAStereo [7] using the low-light enhanced stereo views. Our method performs significantly better than most methods. With respect to CFNet*, our visual results are comparable but with $40\times$ higher inference speed.

Method	Perceptual		Depth		Inference Speed		
	PSNR (dB)↑	SSIM ↑	RMSE ↓	D1% ↓	CPU (s) ↓	GPU (ms) ↓	GFLOPs ↓
SID [6]	16.32	0.696	7.78	22.2	2.23	30.49	200.50
SGN [15]	16.30	0.692	7.93	22.2	2.10	30.86	203.92
StereoSR* [23]	20.97	0.664	7.91	20.2	12.17	147.50	1101.86
PASSR* [59]	14.86	0.699	6.69	18.9	38.31	473.82	2301.25
DASSR* [70]	20.09	0.673	8.12	23.9	6.40	132.27	407.62
CFNet* [56]	24.56	0.718	7.37	21.9	15.75	312.33	1278.44
DVENet [21]	16.08	0.645	**4.81**	**14.85**	5.23	142.14	173.02
Ours	**25.16**	**0.726**	5.70	17.7	**0.35**	**7.47**	**13.12**

Table 1. Quantitative comparisons on the KITTI dataset. The best scores are in **bold** and second best underlined. Our method achieves the best performance on all metrics while delivering real-time inference speed.

pare with StereoSR and PASSR designed for well-lit RGB stereo super-resolution. As our goal is enhancement and not super-resolution we reduced the scaling factor of last layer of these methods from $\times 2, \times 4, ...$ to $\times 1$. These methods also output only the left view and so we replicate the warping and final merging stages for enhancing both left and right views. We also compare with CFNet a light-weight model for computing depth from stereo. CFNet has three stages: stage-I individually computes the left/right features; in stage-II features are warped to obtain a 4D volume; in stage-III 3D convolutions operates on the 4D volume to output a single channel tensor denoting the depth map. We slightly modified the last stage by making 3D convolutions output a 3 channel tensor for RGB color images. We then used L1 and SSIM to train it for RGB image enhancement. In our benchmarking we denote these slightly modified stereo models by adding a '*' in the superscript. We re-trained all models for fair comparison. Finally, few works on raw low-light enhancement, have reported better enhancement using ratio pre-amplification. But as we directly train on sRGB images and not raw images we did not find any performance difference and so do not use it.

We used PyTorch [49], running on Intel Xeon E5-1620V4 CPU with 64GB RAM and RTX 3090 GPU to design our model. We trained our model using ADAM optimizer [30] with default parameters. We trained the model for $100,000$ iterations with learning rate set to 10^{-4}. The training was carried out on randomly chosen patches with no data augmentation as sufficient number of random patches can be obtained from the full images. For KITTI dataset patch size was set to 352×704 and for CityScape it was set to 512×512. For KITTI dataset, $disp_{max} = 200$. And since CityScape's baseline is roughly half of KITTI's baseline, we set $disp_{max} = 100$ for CityScape.

Method	Perceptual		Depth	Inference Speed	
	PSNR↑	SSIM↑	RMSE↓	CPU(s)↓	GPU(ms)↓
SID [6]	27.58	0.840	0.224	10.82	248.96
StereoSR* [23]	28.08	0.810	0.221	52.95	635.03
DASSR* [70]	27.28	0.831	0.223	32.84	616.64
CFNet* [56]	29.17	0.852	0.214	72.71	1361.61
DVENet [21]	29.02	0.847	0.189	30.05	685.74
Ours	**30.49**	**0.853**	**0.177**	**1.63**	**23.02**

Table 2. Quantitative comparisons on the CityScape dataset. Our method achieves the best performance on all other metrics while maintaining a fast inference speed.

4.2. Quantitative and Qualitative comparisons

In Tab. 1 we benchmark our method on 7 metrics: PSNR and SSIM for comparing visual enhancement; RMSE and D1 bad pixel percentage [43] for comparing depth computed from enhanced views; and CPU time, GPU time and Floating-Point Operations (FLOPS) for measuring inference computational complexity. For measuring computational overhead we considered the time/operations required to enhance both left and right views with full spatial resolution.

We find that our method performs significantly better than most methods while exhibiting real-time inference speed, necessary for real-world applications. These quantitative results are also supported by the qualitative results shown in Fig. 3. For computing RMSE and D1 metric we have used the LiDAR GT available in the KITTI dataset. But as LiDAR outputs semi-dense depth, for visual comparison in Fig. 3 we have shown the dense depth map obtained from GT stereo views.

In general, we observe that stereo methods perform better than SID and SGN monocular methods. This is expected as monocular methods do not benefit from the corresponding views. Further, SID uses max-pooling for downsampling which suffers from gradient sparsity and transposed convolution for upsampling, which has been reported to lower the performance [47]. Stereo models like PASSR*, StereoSR* and DVEnet do feature matching for final restoration. Contrary to their approach CFNet* relies only on 3D convolutions for enhancement and thus achieves the best results compared to existing stereo models. This is because, using attention modules or feature correlation is beneficial for well-lit images but not for extremely low-light images having poor contrast and large amount of noise. This fact is also evident from Fig. 3, where the enhancement done by all previous methods except for CFNet* and ours, suffers from the 'Halo Artifacts' resulting from in-

Figure 4. Qualitative results on *real* low-light stereo images using our network. Not only is the enhancement almost identical to the GT views but results of downstream tasks such as depth estimation and semantic segmentation on enhanced views at par with estimates obtained from GT views.

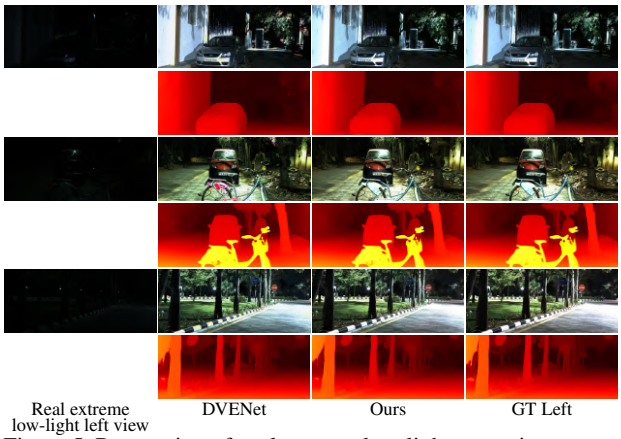


Figure 5. Restoration of *real* extreme low-light stereo images captured in midnight. while DVEnet's & our method's restoration are comparable our method offers atleast $10\times$ speedup.

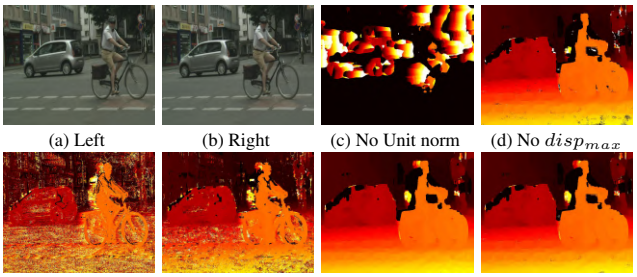


Figure 6. Coarse-level disparity (masked by confidence map) computed by the E-Block under different settings. Small patch sizes produce noisy disparities. We thus use larger patches but increase dilation d to 3 for saving computations. Even if $disp_{max}$ is unknown (easily calculated using baseline and focal length), the disparity map in (d) is quite close to (h), except for tiny white dots, denoting very large disparity.

correct color restoration in the small vicinity of saturated pixels [12, 61, 44]. This is the main reason for superior PSNR of CFNet* and our method. Finally, methods like CFNet* and DASSR* estimate intermediate disparity to warp the views. We, however, do not leverage such ideas in our model because view warping using disparity computed from intermediate low-light features is prone to errors. We rather simply channel-wise concatenate the features and let the network implicitly learn using 2D convolutions by enforcing epipolar constraints using the Depth-Aware loss module over enhanced views. Doing so not only helps our method achieve better visual enhancement and depth estimates than all previous methods in both Tab. 1 and Fig. 3 but also keeps the memory footprint low.

In Tab. 2 we also assess the performance of our model on the CityScape dataset and find that our method does much better than most methods and continues to maintain significantly higher speedup. Overall, KITTI [43] contains images with rich, vibrant colors, while CityScape's [9] well-lit images have relatively lower saturation. Enhancing CityScape images is, therefore, easier than KITTI images. At the same time, CityScape has much higher resolution stereo images, which makes it difficult for models to maintain good inference speed.

4.3. Time Complexity

Generally, stereo models are computationally heavy but as reported in Tab. 1 and Tab. 2, our network is extremely fast. The main reason is that, unlike CFNet*, PASSR* and DEVnet, we do not use 3D convolutions or attention modules. Rather we only use 2D convolutions. StereoSR* also primarily relies on 2D convolution, but it does so mainly at full resolution. However, in our model, most convolutions happen at 1/16 and 1/32 resolution. StereoSR* also needs additional time to convert RGB images into YCrCb color space.

The inference speed of our network is even faster than mono methods such as SID. SID needs to do computations twice to enhance both left and right views. Our model, however channel-wise, concatenates both views at lower resolution and so need not do repeated computations. This not only keeps memory footprint low but also better enforces epipolar constraints.

4.4. Real low-light stereo images

Here we show qualitative results on real low-light stereo images obtained from the Low-Light Light Field (L3F) dataset [32]. The dataset has three subsets, of which we used the L3F-20 subset. The L3F dataset was captured in the evening when the light falling on the camera lens was a maximum of 20 lux. The captured LFs are thus extremely dark. The dataset also captured the corresponding well-lit Light Fields (LFs) by having a large exposure time of $5-10$ seconds. The LFs are arranged in a 15×15 grid, with each SAI having 434×625 spatial resolution. To construct the stereo dataset we ignored the peripheral views as they were severely affected by the vignetting effect and considered only the central 9×9 SAIs. We then selected the extreme

	Independent Feature Extraction	Our Feature Extraction	λ $\mathcal{L}_{disp}$ weight	Perceptual (PSNR)	Depth (RMSE)
Net-I	✓	✗	0	**25.26**	7.82
Net-II	✗	✓	0	25.17	6.00
Net-III	✗	✓	1	24.90	**5.38**
Proposed	✗	✓	0.1	25.16	5.70

Table 3. Ablation Study on the proposed method using the KITTI dataset. Our style of feature extraction benefits epipolar constraints while only slightly lowering the PSNR for visual enhancement. The table also shows the trade-off between perceptual enhancement and depth estimation.

left and right SAIs from the middle row to obtain the stereo pair. We thus retrained our network on 384×384 patches by setting $disp_{max} = 10$ as the SAIs have sub-pixel disparity. Thus, we show that our method can easily switch between large/small baseline systems even though it is optimised for large baseline systems. The qualitative results are shown in Fig. 4. We see that the input stereo views are extremely dark, and yet the restored views look almost identical to GT views. Performance of other tasks such as depth estimation and semantic segmentation on enhanced views is also at par with results computed from the GT stereo views. We couldn't do quantitative benchmarking since the dataset lacks GT depth.

We even took two FLIR machine vision cameras, placed them rigidly at 24cm baseline and captured rectified stereo pairs at midnight with light lux values < 10 on the camera lens. Like SID, we also captured high-exposure images for qualitative benchmarking. Fig. 5 shows the restoration done by DVEnet and our method. We find that both results are comparable but as noted in Tab. 1 and Tab. 2 our method is significantly faster and light-weight.

4.5. Ablation study

Tab. 3 reports the quantitative comparison for stereo enhancement by re-training different versions of our method on the KITTI dataset. Results on Net-I and Net-II demonstrates the benefit of our hybrid architecture. For Net-I throughout our U-net the features for left and right views were processed independently. For Net-II our hybrid style of feature extraction, as shown in Fig. 1, was used. Compared to Net-1, the depth prediction metric is much better by 1.82 units while experiencing only a tiny drop of 0.09 dB PSNR. Moreover, Net-II also has computational advantages. For example, for a 2MP image Net-I requires 77.08 GFLOPS while Net-II only requires 57.42 GFLOPS. We next train Net-II by including our disparity consistency loss, $\mathcal{L}_{disp}$ with weightage of $\lambda = 1$. This improves the depth metric RMSE but lowers the PSNR. This perception-depth trade-off was also noticed in [70]. To favour both restoration and depth we choose $\lambda = 0.1$.

In Sec. 4.2 and Sec. 4.4 we showed results by training our network on KITTI (baseline=54cm), CityScape (baseline=22cm) and L3F dataset (very small baseline) having very different types of acquisition sensors and stereo setup. To train our network on these different datasets, only one hyperparameter, $disp_{max}$, was changed in the Depth-Aware loss module. To further understand the role of different components in the Depth-Aware loss module, we show some coarse-level disparity map computed by the E-Block in Fig. 6. Here the disparity maps are masked by the confidence map estimated by the E-Block. In Fig. 6 c) we do not perform patch-wise unit normalisation leading to incoherent disparity. Next in Fig. 6 d) we assume that stereo metadata namely baseline and focal length is not known and thus $disp_{max}$ is unknown. Except for very tiny sporadic white dots, the estimation is still quite good. Thus, not knowing $disp_{max}$ is not a sever drawback. Next, we reduce the patch size from $M = 31$ to $M = 5, 11$ and find that the estimated disparity are very noisy. Finally, we bring M back to 31 but reduce the dilation d from 3 to 1. The estimated disparity in this case is almost identical to the proposed disparity shown in (h) except for marginal improvement at few isolated points. But at the same time the computational complexity is exceedingly high. For example as noted in Sec. 3.2, computing just one entry of $\mathcal{C}$ in this case requires at least $3 \cdot 31^2$ multiplications. But for the disparity shown in Fig. 6 h) only $3 \cdot \left(\frac{31}{3}\right)^2$ multiplications are required.

5. Conclusion & Future Work

Low-light enhancement has been extensively studied, addressing the nighttime restoration of single-image, videos, and Light Fields. Yet a very important area of fast & light-weight restoration of extreme low-light stereo enhancement has been almost unexplored, which we address in this work. We proposed a hybrid U-net architecture which faithfully restores the stereo images belonging to various datasets, while preserving the epipolar geometry. The inference speed of our network is significantly better than existing stereo methods because we use only 2D convolutions and enforce the epipolar constraints during training by using the Depth-Aware loss module. We showed that this module can be used out-of-the-box for training on different types of datasets such as KITTI which has a very large baseline and CityScape which has medium-sized baseline. Finally, our network is even faster than mono methods such as SID. This is because our hybrid architecture at lower resolutions jointly operates on both stereo features and unlike mono methods we need not do repeated convolutions. Overall, our network offers $4 - 60\times$ speedup with $15 - 100\times$ lower floating-point operations compared to existing strategies. As a future work, we wish to parameterise the scale at which feature merging happens in our model as a function of stereo baseline though it will make training harder.

References

[1] Yousef Atoum, Mao Ye, Liu Ren, Ying Tai, and Xiaoming Liu. Color-wise attention network for low-light image enhancement. In *CVPR Workshop*, 2020.

[2] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *CVPR*, 2019.

[3] Rohan Chabra, Julian Straub, Christopher Sweeney, Richard Newcombe, and Henry Fuchs. Stereodrnet: Dilated residual stereonet. In *CVPR*, pages 11786–11795, 2019.

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, 2018.

[5] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *ICCV*, 2019.

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

[7] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *NIPS*, 33:22158–22169, 2020.

[8] Dinu Coltuc, Philippe Bolon, and J-M Chassery. Exact histogram specification. *IEEE Transactions on Image Processing*, 15(5):1143–1152, 2006.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[10] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *ICCV*, 2021.

[11] Xin Deng, Wenzhe Yang, Ren Yang, Mai Xu, Enpeng Liu, Qianhan Feng, and Radu Timofte. Deep homography for efficient stereo image compression. In *CVPR*, 2021.

[12] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)*, 36(6):1–15, 2017.

[13] Xueyang Fu, Delu Zeng, Yue Huang, Yinghao Liao, Xinghao Ding, and John Paisley. A fusion-based enhancing method for weakly illuminated images. *Signal Processing*, 129:82–96, 2016.

[14] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *CVPR*, 2016.

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

[16] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*, 2020.

[17] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016.

[18] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *CVPR*, pages 3273–3282, 2019.

[19] Yanhui Guo, Xue Ke, Jie Ma, and Jun Zhang. A Pipeline Neural Network for Low-Light Image Enhancement. *IEEE Access*, 7:13737–13744, 2019.

[20] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE TPAMI*, 30(2):328–341, 2008.

[21] Jie Huang, Xueyang Fu, Zeyu Xiao, Feng Zhao, and Zhiwei Xiong. Low-light stereo image enhancement. *IEEE Transactions on Multimedia*, 2022.

[22] Haidi Ibrahim and Nicholas Sia Pik Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(4):1752–1758, 2007.

[23] Daniel S. Jeon, Seung-Hwan Baek, Inchang Choi, and Min H. Kim. Enhancing the spatial resolution of stereo images using a parallax prior. In *CVPR*, pages 1721–1730, 2018.

[24] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *ICCV*, 2019.

[25] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *arXiv:1906.06972*, 2019.

[26] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image processing*, 6(7):965–976, 1997.

[27] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, 2017.

[28] Hakyeong Kim, Andreas Meuleman, Daniel S Jeon, and Min H Kim. High-quality stereo image restoration from double refraction. In *CVPR*, 2021.

[29] Yeong-Taeg Kim. Contrast enhancement using brightness preserving bi-histogram equalization. *IEEE transactions on Consumer Electronics*, 43(1):1–8, 1997.

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[31] Mohit Lamba, Atul Balaji, and Kaushik Mitra. Towards fast and light-weight restoration of dark images. In *BMVC*, 2020.

[32] Mohit Lamba, Kranthi Kumar Rachavarapu, and Kaushik Mitra. Harnessing multi-view perspective of light fields for low-light imaging. *IEEE Transactions on Image Processing*, 30:1501–1513, 2021.

[33] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977.

[34] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971.

[35] Chulwoo Lee, Chul Lee, and Chang-Su Kim. Contrast enhancement based on layered difference representation of 2d histograms. *IEEE transactions on image processing*, 22(12):5372–5384, 2013.

[36] Chongyi Li, Jichang Guo, Fatih Porikli, and Yanwei Pang. Lightennet: a convolutional neural network for weakly illuminated image enhancement. *Pattern Recognition Letters*, 104:15–22, 2018.

[37] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust retinex model. *IEEE Transactions on Image Processing*, 27(6):2828–2841, 2018.

[38] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017.

[39] Feifan Lv, Yu Li, and Feng Lu. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *arXiv preprint arXiv:1908.00682*, 2019.

[40] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017.

[41] Paras Maharjan, Li Li, Zhu Li, Ning Xu, Chongyang Ma, and Yue Li. Improving extreme low-light image denoising via residual learning. In *Int. Conf. Multimedia and Expo (ICME)*, 2019.

[42] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. arXiv:1512.02134.

[43] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[44] Laurence Meylan and Sabine Susstrunk. High dynamic range image rendering with a retinex-based adaptive filter. *IEEE Transactions on image processing*, 15(9):2820–2830, 2006.

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

[46] Guang-Yu Nie, Ming-Ming Cheng, Yun Liu, Zhengfa Liang, Deng-Ping Fan, Yue Liu, and Yongtian Wang. Multi-level context ultra-aggregation for stereo matching. In *CVPR 2019*, pages 3278–3286, 2019.

[47] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.

[48] Seonhee Park, Soohwan Yu, Byeongho Moon, Seungyong Ko, and Joonki Paik. Low-light image enhancement using variational optimization-based retinex model. *IEEE Trans. Consumer Electronics*, 63(2):178–184, 2017.

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*. 2019.

[50] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.

[51] Tal Remez, Or Litany, Raja Giryes, and Alex M Bronstein. Deep Convolutional Denoising of Low-Light Images. *arXiv:1701.01687*, 2017.

[52] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. Low-light image enhancement via a deep hybrid network. *IEEE Transactions on Image Processing*, 28(9):4364–4375, 2019.

[53] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, 2019.

[54] Yukihiro Sasagawa and Hajime Nagahara. Yolo in the dark-domain adaptation method for merging multiple models. In *ECCV*, 2020.

[55] Aashish Sharma and Loong-Fah Cheong. Into the twilight zone: Depth estimation using joint structure-stereo optimization. In *ECCV*, pages 103–118, 2018.

[56] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *CVPR*, 2021.

[57] Xiao Song, Guorun Yang, Xinge Zhu, Hui Zhou, Zhe Wang, and Jianping Shi. Adastereo: A simple and efficient approach for adaptive stereo matching. In *CVPR*, pages 10328–10337, 2021.

[58] J Alex Stark. Adaptive image contrast enhancement using generalizations of histogram equalization. *IEEE Transactions on image processing*, 9(5):889–896, 2000.

[59] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *CVPR*, pages 12250–12259, 2019.

[60] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, 2019.

[61] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing*, 22(9):3538–3548, 2013.

[62] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Symmetric parallax attention for stereo image super-resolution. In *CVPR workshop*, pages 766–775, 2021.

[63] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018.

[64] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *CVPR*, 2021.

[65] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *CVPR*, pages 1959–1968, 2020.

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

[67] Qingyu Xu, Longguang Wang, Yingqian Wang, Weidong Sheng, and Xinpu Deng. Deep bilateral learning for stereo image super-resolution. *IEEE Signal Processing Letters*, 28:613–617, 2021.

[68] Xin Xu and Jie Wang. Extended non-local feature for visual saliency detection in low contrast images. In *ECCV Workshops*, 2018.

[69] Xin Xu, Shiqin Wang, Zheng Wang, Xiaolong Zhang, and Ruimin Hu. Exploring image enhancement for salient object detection in low light images. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1s):1–19, 2021.

[70] Bo Yan, Chenxi Ma, Bahetiyaer Bare, Weimin Tan, and Steven Hoi. Disparity-aware domain adaptation in stereo image restoration. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13176–13184, 2020.

[71] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *CVPR*, pages 5515–5524, 2019.

[72] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *CVPR*, 2020.

[73] Xinyi Ying, Yingqian Wang, Longguang Wang, Weidong Sheng, Wei An, and Yulan Guo. A stereo attention module for stereo image super-resolution. *IEEE Signal Processing Letters*, 27:496–500, 2020.

[74] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *CVPR*, pages 185–194, 2019.

[75] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM International Conference on Multimedia*, 2019.

[76] Yiran Zhong, Yuchao Dai, and Hongdong Li. Self-supervised learning for stereo matching with self-improving ability. *arXiv preprint arXiv:1709.00930*, 2017.