Supplementary Materials: Learning Exposure Correction in Dynamic Scenes

Anonymous Authors

1 OVERVIEW

In this supplementary material, we provide more explanation, discussion, visualization, and experimental results. They are organized as follows:

- We present additional information and visualization of the DIME dataset in Sec. 2.
- Implementation details of the capture system, metrics, and evaluated models are covered in Sec. 3.
- More ablation studies and qualitative results are covered in Sec. 4.

2 DATASET VISUALIZATION

We present the luminance curves of the video pairs in our DIME dataset to visualize the exposure difference in the inter-frame and inter-video dimensions, as shown in Fig. 1.

3 IMPLEMENTATION DETAILS

3.1 Quantitative Metrics

The reference-based PSNR and SSIM measure the difference between the enhanced results and the ground truth. NIQE estimates image quality by measuring the deviations from the statistical regularities of natural images. A smaller NIQE indicates a more naturalistic and perceptually favored quality. Note that Average Luminance Variance (ALV) measures the difference in the average luminance of the same object between adjacent frames. A smaller ALV suggests better temporal consistency in the enhanced video. ALV scores are calculated by: $ALV = \frac{1}{N}\sum_{i=1}^{N}(L_i - L_{avg})^2$, where N is the number of frames of a video, L_i represents the average luminance value of the i-th frame, and L_{avg} denotes the average luminance value of all frames in the video.

3.2 Alignment Metrics

We evaluate alignment performance and motion activity on two evaluation metrics, LOE and optical flow, employed for the assessment of our DIME datasets. We provide comprehensive details regarding the implementation and enhancements of these metrics, as shown in Algorithm 1 and 2. To mitigate potential interference from video resolution and noise in the LOE metric, we standardize the sizes of videos across all datasets and perform LOE calculations on their grayscale images. Algorithm 2 presents the pseudocode of optical flow calculation. Similarly, we standardize video sizes across all datasets and compute optical flow on their grayscale images. The evaluation of scene motion is conducted through an analysis of the average magnitude of the optical flow vectors.

3.3 Capture System

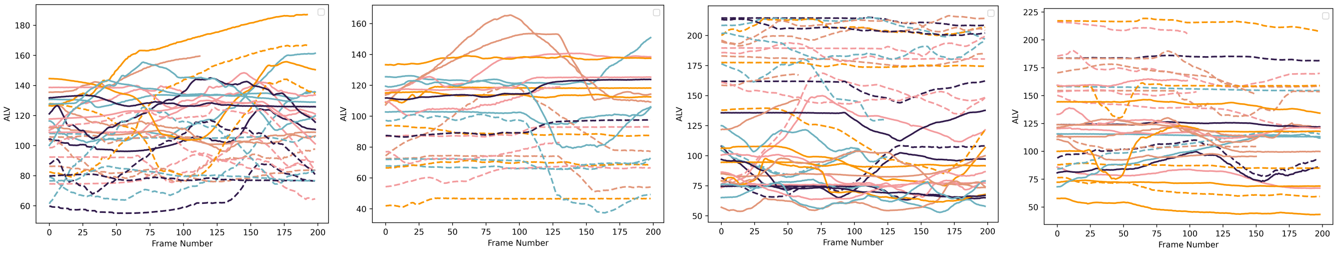
For dynamic Low-Quality-Ground-Truth (LQ-GT) video capture, we need to utilize two DSLR cameras with different settings, as shown in Fig. 2. Specifically, we configure both cameras in automatic mode to ensure adaptive exposure. The ISO range is set between 100 and 12800, while the lens focal length is fixed at 45mm. Shutter

Algorithm 1 Calculate LOE (H, W, win)

```
Input: Under-/Over-exposed video $\mathcal{I}^{o/u}$ and normal exposed video $\mathcal{I}^{gt}$
Output: LOE scores
Params: $H, W$ (resized height/width), $win$ (window size)
$LOE \leftarrow []$
for each frame $(\mathcal{I}_t^{o/u}, \mathcal{I}_t^{gt})$ of $(\mathcal{I}^{o/u}, \mathcal{I}^{gt})$ do
    $\mathcal{I}_t^{o/u} \leftarrow$RGB2GRAY(Resize($\mathcal{I}_t^{o/u}, (H, W)$))
    $\mathcal{I}_t^{gt} \leftarrow$ RGB2GRAY(Resize($\mathcal{I}_t^{gt}, (H, W)$))
    $LOE_t \leftarrow []$
    for $x \leftarrow 0$ to $w - 1$ step $win$ do
        for $y \leftarrow 0$ to $h - 1$ step $win$ do
            $RD \leftarrow 0$
            for $win_x \leftarrow 0$ to $win - 1$ do
                for $win_y \leftarrow 0$ to $win - 1$ do
                    $E \leftarrow (F_l[x + win_x, y + win_y] > F_l[x : x + win, y : y + win] \oplus (F_h[x + win_x, y + win_y] > F_h[x : x + win, y : y + win])$
                    $RD \leftarrow RD + sum(E)$
                end for
            end for
            $LOE_t.append(RD/(win \times win))$
        end for
    end for
    $LOE.append(mean(LOE_t))$
end for
return $mean(LOE)$
```

Algorithm 2 Calculate Optical Flow (H, W, f)

```
Input: Random video $\mathcal{I}$
Output: Optical Flow scores
Params: $H, W$ (resized height and width), Farneback flow parameters
Mean normalized flow $f_n^m \leftarrow 0$
$count \leftarrow 0$
for Adjacent frames $(\mathcal{I}_t, \mathcal{I}_{t+1})$ of $\mathcal{I}$ do
    $\mathcal{I}_t \leftarrow$ RGB2GRAY(Resize($\mathcal{I}_t, (H, W)$))
    $\mathcal{I}_{t+1} \leftarrow$ RGB2GRAY(Resize($\mathcal{I}_{t+1}, (H, W)$))
    Flow $f \leftarrow$ CalcOpticalFlowFarneback($\mathcal{I}_t, \mathcal{I}_{t+1}$)
    $f_n \leftarrow norm(f, axis = -1)$
    $f_n^m \leftarrow f_n^m + mean(f_n)$
    $count \leftarrow count + 1$
end for
$f_n^m \leftarrow f_n^m / count$
return $f_n^m$
```



(a) Underexposure-GT Part1 (b) Underexposure-GT Part2 (c) Overexposure-GT Part1 (d) Overexposure-GT Part2

Figure 1: Lux curves of the under-/over-exposed (dotted line) and normal exposed (solid line) video pairs in the DIME dataset.

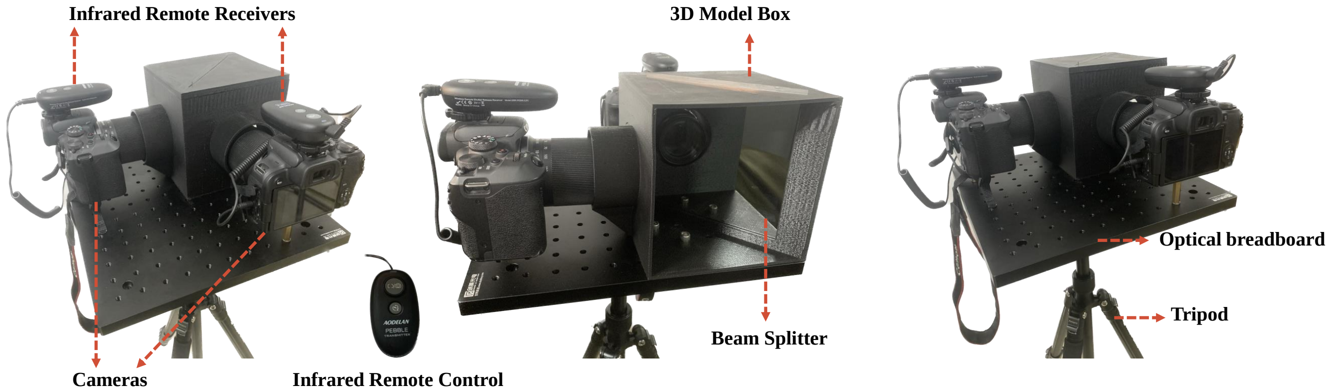


Figure 2: Visualization of the physical picture of the capture system.

Table 1: Ablation study for investigating the components of the specific modules.

Alignment	DCN Align	×	✓	×
	Fourier Align	×	×	✓
	PSNR/SSIM	19.12/0.8810	19.35/0.8948	20.11/0.9016
Illumination	Under-stream	✓	×	✓
	Over-stream	×	✓	✓
	PSNR/SSIM	19.47/0.8915	19.59/0.8942	20.11/0.9016
Synthesis	Stage-1	✓	×	✓
	Stage-2	×	✓	✓
	PSNR/SSIM	19.86/0.8972	19.79/0.8906	20.11/0.9016

mode is set to metering plus servo autofocus. The default aperture range spans from f/4 to f/22. We solve parallax problems by using a large beam splitter to cover the lens of DSLR cameras. To this end, we utilize a large and cheap beam splitter with reflectance coating and antireflection coating, instead of a small and expensive beam splitter cube. In this way, the two cameras can receive natural light from the same viewpoint. The size of the beam splitter is $150 \times 150 \times 0.7 (mm^3)$. We adjust the camera to video mode and set the resolution to 4K.

Upon concluding the video capture process, professional photographers carefully engage in color grading and rendering for the creation of the GT videos from LOG format to sRGB format. Then we apply a two-stage frame alignment strategy to obtain aligned pairs, and manually remove parts of the videos with large alignment errors.

3.4 Training Details

The training set contains 47 underexposed videos and 43 overexposed videos. In the data preprocessing stage, we first convert the video into frame-by-frame images, and then resize the resolution to 960×512. All frames of each video are organized within the same folder, and accessed in numpy format.

4 EXPERIMENTAL RESULTS

4.1 Ablation Study

We provide a series of detailed ablation studies to evaluate the effectiveness of each component in the proposed method, as shown in Table 1 and Table 2. We evaluate the key modules in our network by replacing them with other straightforward solutions.

(1) **Alignment.** We assess the efficiency of the Multi-frame Fourier Alignment module by replacing it with a standard DCN alignment. It becomes evident that Fourier alignment surpasses DCN alignment in performance. The superiority lies in the precision of offsets calculated by disentangled amplitude alignment compared to those derived from pixel-level alignment. Additionally, we present the outcome when removing alignment module, which exhibits inferior performance compared to our method.

Table 2: Ablation study on loss terms.

Method	$\mathcal{L}_{pix}$	$\mathcal{L}_{tv}$	$\mathcal{L}_{amp}$	PSNR	SSIM
(a)	✓			18.95	0.8862
(b)	✓	✓		19.47	0.8973
(c)	✓	✓	✓	20.11	0.9016

Table 3: Model efficiency and performance on LLVE tasks. The results are obtained considering a single frame resolution of 960 × 512.

Method	SDSD [9]	DID [3]	Efficiency
SDSDNet [9]	24.92/0.73	21.88/0.83	214.4GMacs/97.4ms
VECNet (Ours)	25.39/0.84	24.76/0.90	74.84GMacs/62.9ms

(2) **Dual-stream Illumination.** When one of the illumination learning branches is removed from our complete model, the results correspondingly decline. It confirms the utility of the dual-stream unit in leveraging complementary illumination information.

(3) **Two-stage Synthesis.** We further illustrate the results obtained by training each synthesis stage independently. Upon removing any single stage, the performance deteriorates to differing extents. This serves as evidence that the two-stage synthesis strategy confers beneficial improvements to the model.

(4) **Individual losses.** We conduct ablation studies to validate the effectiveness of loss functions, which are proven to be effective, demonstrating the reasonableness of the supervision manner.

4.2 More Qualitative Results

As shown in Figs. 5, 8, 11, we give more visual comparisons of methods (MSEC [1], DRBN-ENC [4], ECLNet [6], FECNet [5], and LACT [2], SMOID [7], SDSD [9], RVRT [8], and DIDNet [3].) from our DIME dataset, mobile phones, and the Internet. Our method achieves excellent visual performance in both global illumination and image details compared to other methods, indicating that our proposed approaches not only have better performance but also have excellent generalization ability. We present the enhanced videos in the zip file, including the results of our proposed method and the previous methods.

4.3 Model efficiency and performance on low-light video enhancement tasks

We first conduct experiments on two representative low-light video enhancement datasets, including SDSD and DID. Then we provide comparisons about computation costs and running time with SDSDNet [9] to illustrate the model efficiency. The performance presented on Table. 3 monitors both the effectiveness and efficiency of our method.

REFERENCES

[1] Mahmoud Afifi, Konstantinos G Derpanis, Bjorn Ommer, and Michael S Brown. 2021. Learning multi-scale photo exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE, Nashville, TN, USA, 9157–9167.

[2] Jong-Hyeon Baek, DaeHyun Kim, Su-Min Choi, Hyo-jun Lee, Hanul Kim, and Yeong Jun Koh. 2023. Luminance-aware Color Transform for Multiple Exposure Correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society, Los Alamitos, CA, USA, 6133–6142.

[3] Huiyuan Fu, Wenkai Zheng, Xicong Wang, Jiaxuan Wang, Heng Zhang, and Huadong Ma. 2023. Dancing in the Dark: A Benchmark towards General Low-light Video Enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Computer Vision Foundation/IEEE, Los Alamitos, CA, USA, 12831–12840.

[4] Jie Huang, Yajing Liu, Xueyang Fu, Man Zhou, Yang Wang, Feng Zhao, and Zhiwei Xiong. 2022. Exposure normalization and compensation for multiple-exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE, New Orleans, LA, USA, 6043–6052.

[5] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. 2022. Deep fourier-based exposure correction network with spatial-frequency interaction. In *European Conference on Computer Vision*. Springer Nature Switzerland, Cham, 163–180.

[6] Jie Huang, Man Zhou, Yajing Liu, Mingde Yao, Feng Zhao, and Zhiwei Xiong. 2022. Exposure-consistency representation learning for exposure correction. In *Proceedings of the 30th ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 6309–6317.

[7] Haiyang Jiang and Yinqiang Zheng. 2019. Learning to see moving objects in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Computer Vision Foundation/IEEE, Seoul, Korea (South), 7323–7332.

[8] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. 2022. Recurrent video restoration transformer with guided deformable attention. In *Advances in Neural Information Processing Systems*. MIT Press, New Orleans, LA, USA.

[9] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. 2021. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Montreal, QC, Canada, 9680–9689.

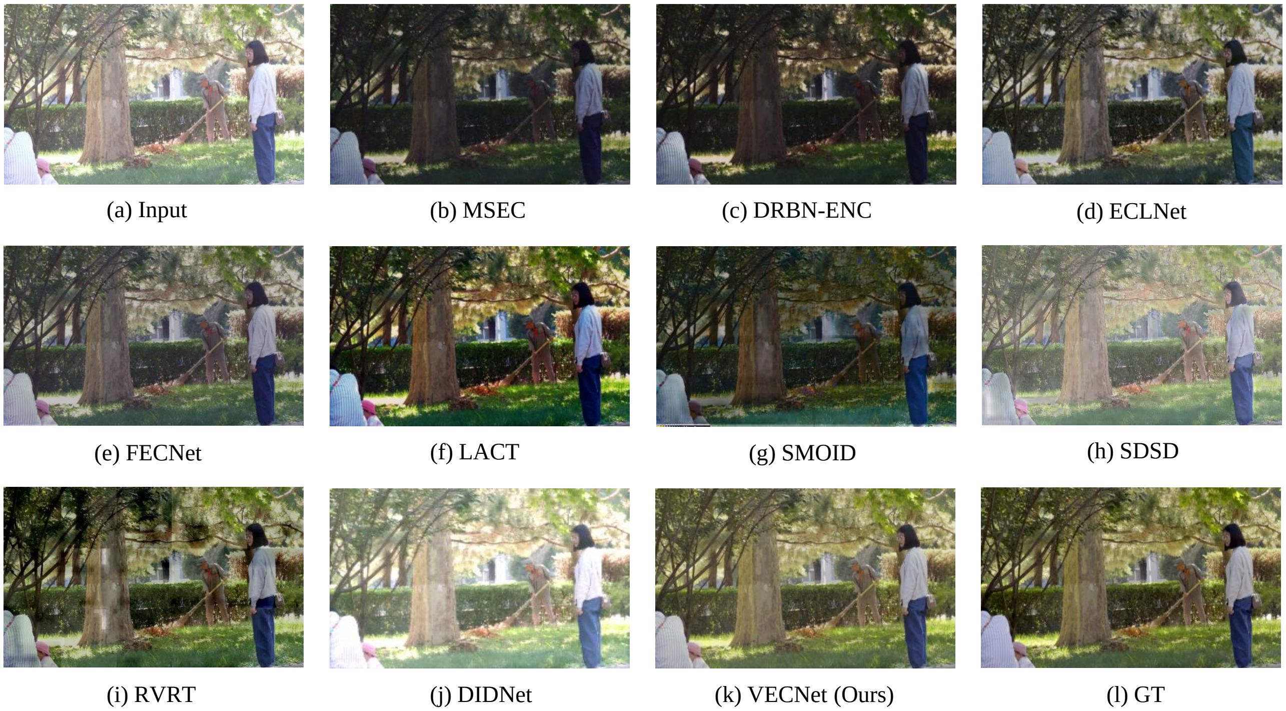


Figure 3: Visual comparisons of overexposure examples.

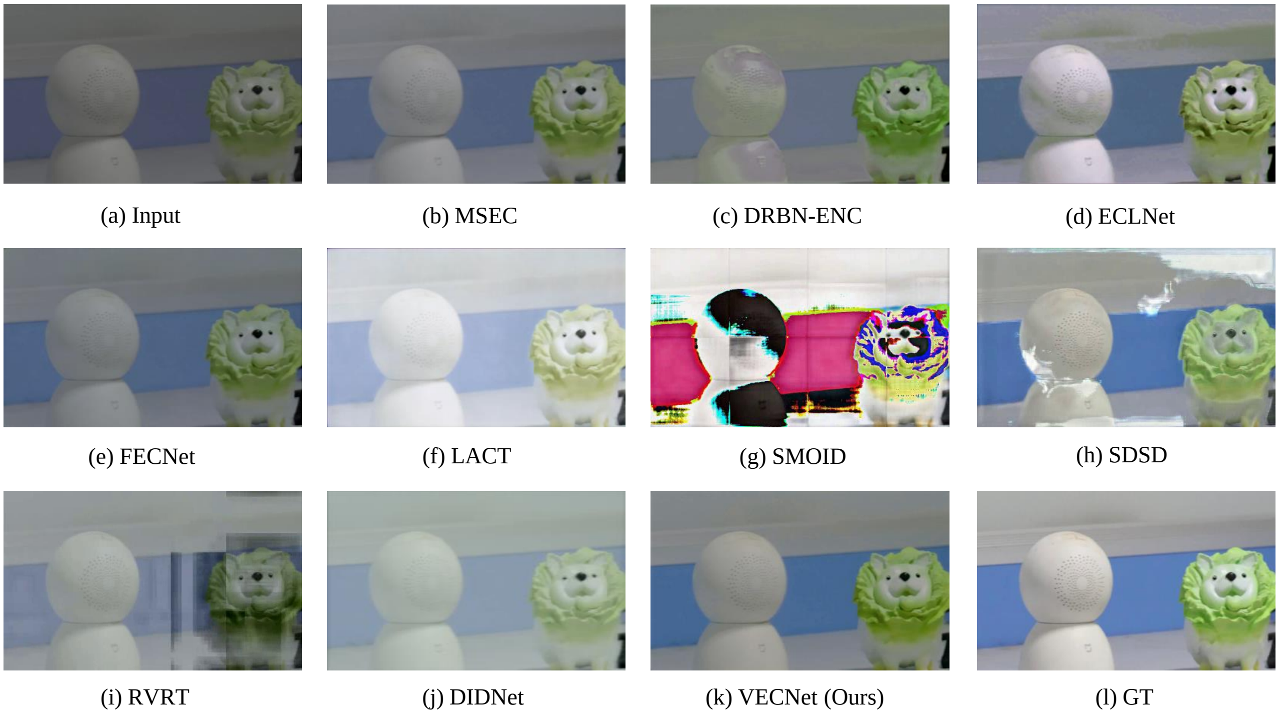
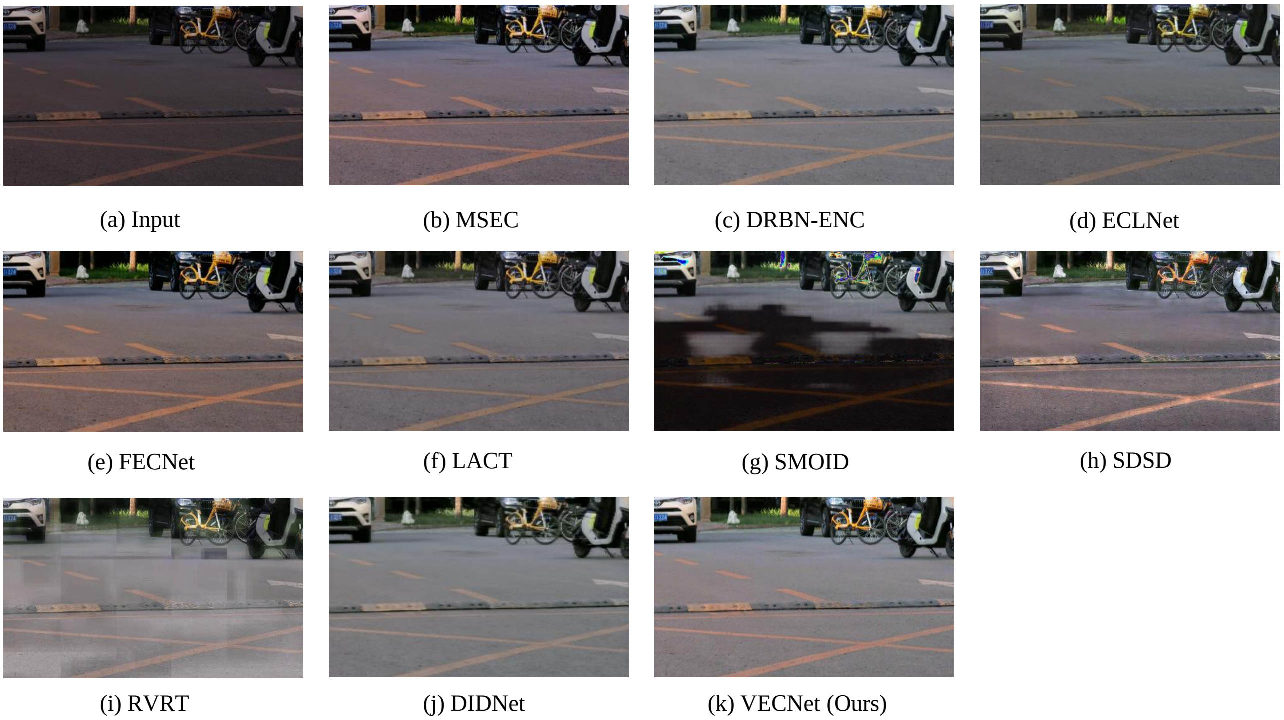


Figure 4: Visual comparisons of underexposure examples.

Figure 5: Visual comparison with state-of-the-art methods on DIME dataset.

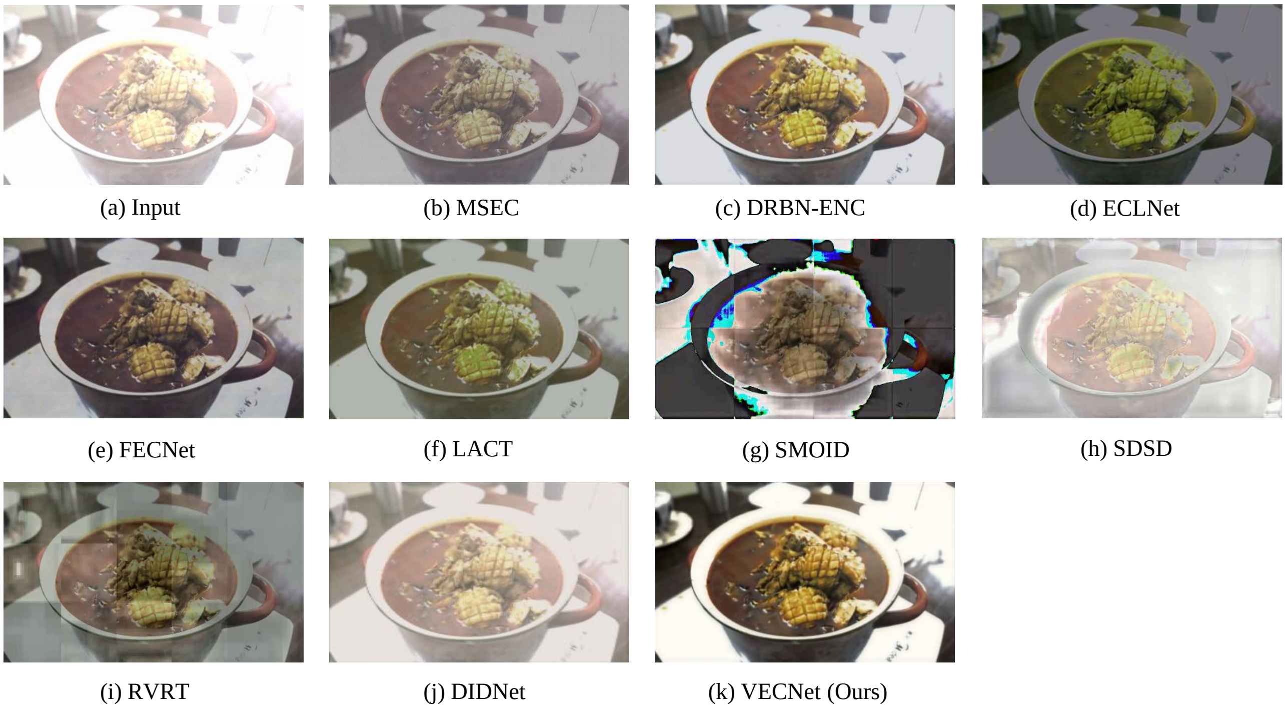


(a) Input (b) MSEC (c) DRBN-ENC (d) ECLNet

(e) FECNet (f) LACT (g) SMOID (h) SDSD

(i) RVRT (j) DIDNet (k) VECNet (Ours)

Figure 6: Visual comparisons of underexposure examples.



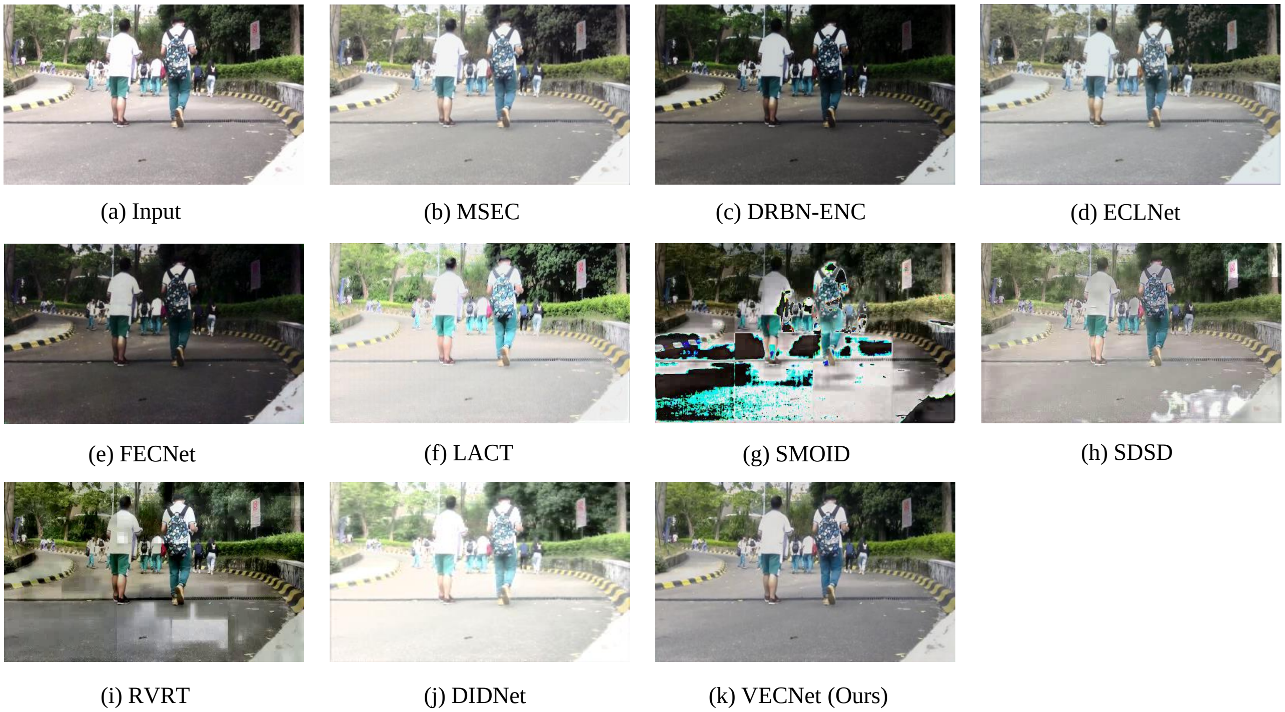
(a) Input (b) MSEC (c) DRBN-ENC (d) ECLNet

(e) FECNet (f) LACT (g) SMOID (h) SDSD

(i) RVRT (j) DIDNet (k) VECNet (Ours)

Figure 7: Visual comparisons of overexposure examples.

Figure 8: Visual comparison with state-of-the-art methods captured on mobile phones in real-world scenes.

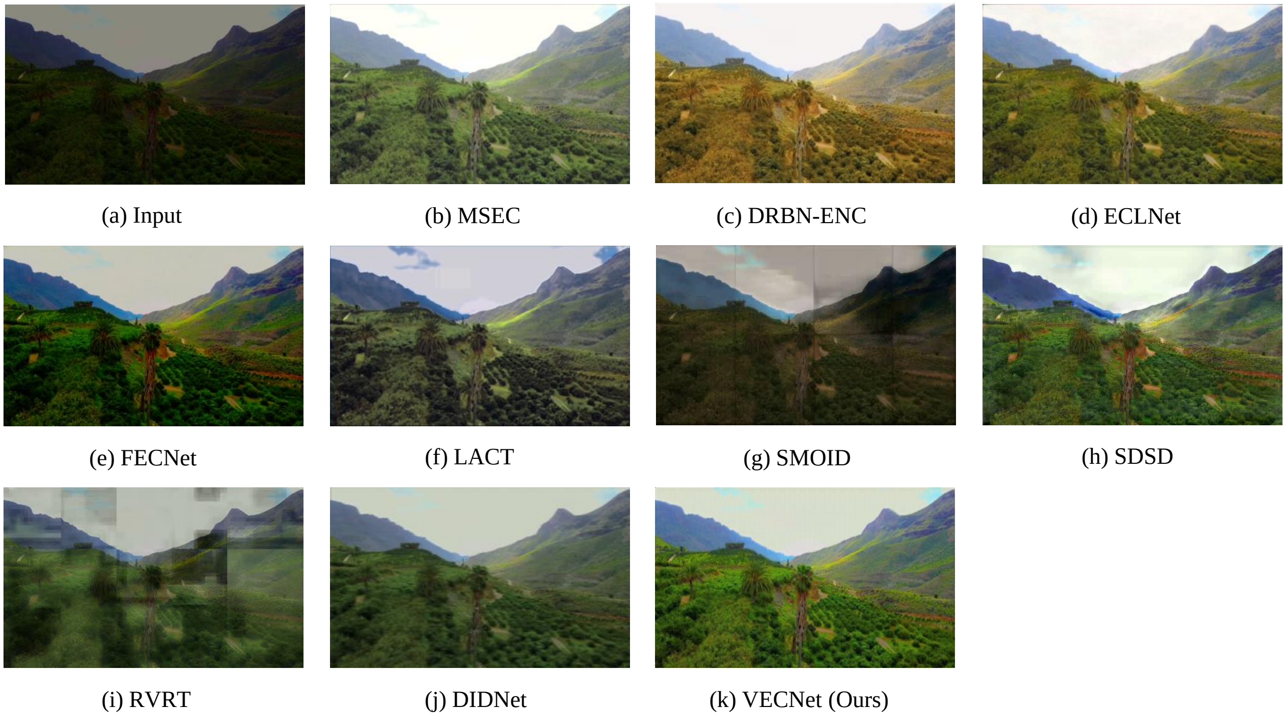


(a) Input (b) MSEC (c) DRBN-ENC (d) ECLNet

(e) FECNet (f) LACT (g) SMOID (h) SDSD

(i) RVRT (j) DIDNet (k) VECNet (Ours)

Figure 9: Visual comparisons of overexposure examples.



(a) Input (b) MSEC (c) DRBN-ENC (d) ECLNet

(e) FECNet (f) LACT (g) SMOID (h) SDSD

(i) RVRT (j) DIDNet (k) VECNet (Ours)

Figure 10: Visual comparisons of underexposure examples.

Figure 11: Visual comparison with state-of-the-art methods collected from the Internet.