RetinexMCNet: A Memory Controller Dominated Network for Low-Light Video Enhancement Based on Retinex – Supplemental Document –

Meiao Wang[1], Xuejing Kang[1]*, Yaxi Lu[2], Jie Xu[1]
[1]Beijing University of Posts and Telecommunications, [2]Tsinghua University
{meiaowang, kangxuejing}@bupt.edu.cn

In the supplementary document, we describe the following parts:

- Complexity analysis;
- Explanation and Robustness of the $\mathcal{L}_{LTS}$;
- Generalization ability on real-world captured videos;
- Efficiency analysis of our MC on in-distribution (SDSD [11] and DID [5]) and out-of-distribution (Loli-Phone [7]) datasets;
- Hyper-parameter analysis;
- Failure case analysis.

All experiments are conducted on an NVIDIA RTX 4090 GPU using PyTorch. Notably, all critical settings are keep consistent among different methods for fair comparison.

1. Complexity Analysis

To evaluate the training overhead and complexity of our RetinexMCNet in contrast to state-of-the-art (SOTA) LLIE and LLVE methods, we present their entire training process in Fig. 1 and the complexity in Tab. 1.

As shown in Fig. 1, our two-stage training strategy (50 epochs for per-frame enhancement with low complexity, 20 epochs for temporal memory integration by activating MC) achieves SOTA performance with moderate training cost. As shown in Tab. 1, our RetinexMCNet attains the highest performance on both datasets while maintaining competitive complexity and inference time.

In summary, our method ensures efficiency and practicality, achieving SOTA with balanced training overhead, complexity, and inference time.

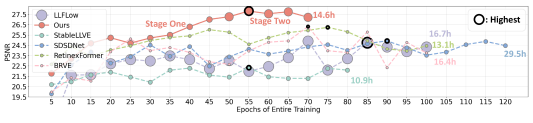


Figure 1. Entire training process. Circle size denotes complexity. (Same as Fig. 10 in the main paper)

*Corresponding author

2. Explanation and Robustness of the LTS Loss

2.1. Explanation

Limited by the length of the main paper, we further explain the symbols defined in $\mathcal{L}_{LTS}$:

- $\boldsymbol{L_t}$: The input illumination map of low-light frame X_t;
- $\hat{\boldsymbol{L}}_{\boldsymbol{t}}$: The output illumination map of our network;
- **Alg. 1**: It calculates the relative value of each map in the vertical direction; To simplify understanding, you can reduce 3D to 2D by ignoring unchanged dimensions, *e.g.*, $(H \times W \times 1, 1 \times W \times H)$ to $(H \times 1, 1 \times H)$. Moreover, in practice, we randomly compute horizontal and vertical directions for robustness;
- $\hat{\mathbf{I}}$: It is the relative lightness order matrices of L_t computed by Alg. 1;
- $\hat{\mathbf{O}}$: It is the relative lightness order matrices of $\hat{L}_t$ computed by Alg. 1;
- $\hat{\mathbf{d}}$: For each pixel, we take the values of the corresponding position from $\hat{\mathbf{I}}$ and $\hat{\mathbf{O}}$ to form a coordinate pair $(\hat{\imath}, \hat{o})$, which can be mapped to a 2D plane to get $\hat{\mathbf{d}}$;
- $\boldsymbol{\theta}$: The angle of $\hat{\mathbf{d}}$.

Finally, our *lightness* term of $\mathcal{L}_{LTS}$ prevent relative level reversal to mitigate overexposure by constraining $\hat{\mathbf{d}}$ and θ.

2.2. Robustness

As discussed in the main paper, our $\mathcal{L}_{LTS}$ not only effectively mitigates overexposure and preserves texture details in our scheme (Fig. 9 in the main paper), but also can be applied to any other intra-frame enhancement techniques to improve their performance robustly.

Here, to verify the robustness of our dual-perspective $\mathcal{L}_{LTS}$ to different models, we apply it to three effective models, RetinexFormer [1], BRVE [15], and StableLLVE [14], selected from LLIE, adjacent, and random frames methods (defined according to $\mathcal{S}$ in the *Consistency Regularization Term* in Eq. (1)). As shown in Fig. 2, $\mathcal{L}_{LTS}$ effectively mitigates overexposure while simultaneously preserving texture details across all models.

Table 1. Quantitative results and complexity of LLVE and LLIE methods. The highest values are in **red**, the second highest values are in **blue**, and the underlined one is unsupervised. The code of LAN[5] is not open yet. We report the total number of parameters (Params), the foating point operations (FLOPs), and the inference time. Patch size: 128×128.

Type	Methods	Params (M)	GFLOPs	Inference Time (ms)	SDSD		DID	
					PSNR↑	SSIM↑	PSNR↑	SSIM↑
LLIE	LLFlow[12]	39.91	409.50	53.74	24.90	0.78	25.71	**0.92**
	SNRNet[13]	4.01	26.35	3.67	25.27	0.82	24.05	0.90
	Retinexformer[1]	1.61	15.57	2.50	26.24	0.83	27.39	0.89
	EvLight[8]	22.73	180.90	12.33	18.16	0.68	22.23	0.78
	Zero-IG[10]	0.12	8.10	0.61	12.31	0.50	17.31	0.81
LLVE	MBLLEN[9]	2.78	114.38	118.16	21.79	0.65	24.82	0.91
	SMID[3]	85.65	0.17	0.04	24.09	0.69	22.97	0.87
	SDSDNet[11]	4.30	9.80	0.10	24.92	0.73	21.88	0.83
	StableLLVE[14]	4.32	2.52	8.74	22.28	0.84	23.35	0.89
	Chhirolya et al.[4]	8.01	23.09	57.98	23.46	0.79	22.77	0.88
	LAN[5]		Code is not open.		**27.25**	**0.85**	**29.01**	**0.94**
	BRVE[15]	0.37	0.03	5.76	26.31	0.82	24.43	0.87
	Ours	27.97	25.16	20.85	**27.81**	**0.88**	**30.09**	0.91

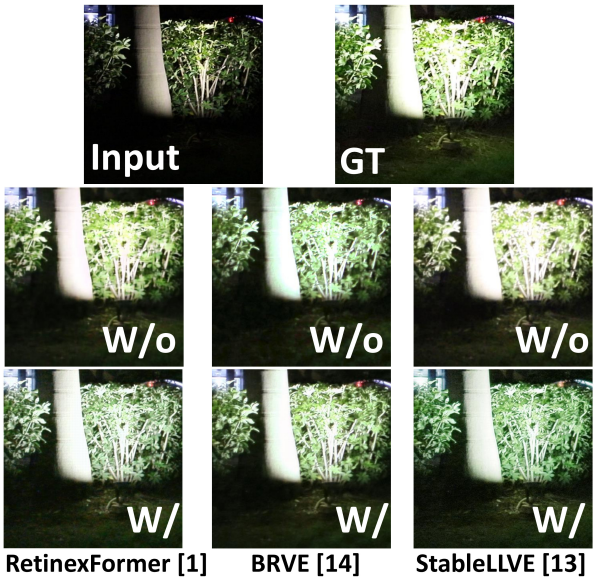


Figure 2. Visual comparison without and with $\mathcal{L}_{LTS}$. From left to right, the columns correspond to RetinexFormer [1], BRVE [15], and StableLLVE [14], respectively.

3. Generalization Ability

To further test the generalization ability of models, we captured three representative low-light videos by Xiaomi 14 Pro at 23:00p.m. in Beijing, China on November 20, 2024, *i.e.*, an extremely dark video, an outdoor video and an indoor video. The quantitative results of consistency and visual comparisons of state-of-the-art low-light video enhancement (LLVE) and low-light image enhancement (LLIE) methods are shown in Tab. 2 and Fig. 3, including SNRNet [13], Retinexformer [1], MBLLEN [9], SDSDNet [11], StableLLVE [14], and BRVE [15].

In the three real-world scenarios, particularly under extremely low-light conditions, our method demonstrates significant advantages quantitatively and qualitatively. As shown in Fig. 3, for individual frames, unlike other methods that produce black blurs (BRVE, MBLLEN, SDSDNet in the extremely dark) and local under-exposure (RetinexFormer and SNRNet in the extremely dark), over-exposure (StableLLVE in the indoor), our method delivers well-balanced lighting and shadow effects. Across frames, our method effectively minimizes inter-frame flickering and artifacts, ensuring superior visual consistency (Tab. 2 and Fig. 3). Thus, this demonstrates the strong generalization ability and robustness of our model.

Table 2. Quantitative results of consistency on real-shot videos. To measure the consistency level of a video, Mean absolute brightness differences (MABD) [6] is proposed as a general level of time derivatives of brightness value on each pixel location. Here, we regard its average value as the brightness flicker degree of a video (smaller value denotes better consistency). The lowest value is in **red**, the second one is in **blue**.

Methods	Extremely Dark	Outdoor	Indoor
SNRNet [13]	3.354	16.744	5.969
RetinexFormer [1]	2.901	**9.270**	3.104
MBLLEN[9]	3.767	9.311	3.079
SDSDNet[11]	4.926	10.464	6.810
StableLLVE[14]	**2.209**	9.846	**2.812**
BRVE[15]	2.219	9.403	3.640
Ours	**1.826**	**9.113**	**2.716**

Please view the **.mp4** files in our additional supplementary materials, which clearly illustrates the strengths and robustness of our method on both in-distribution and out-of-distribution data.

4. Efficiency Analysis of Our MC

We explore the relationship between memory pool size, the average utilization of stored key-value pairs, and perfor-

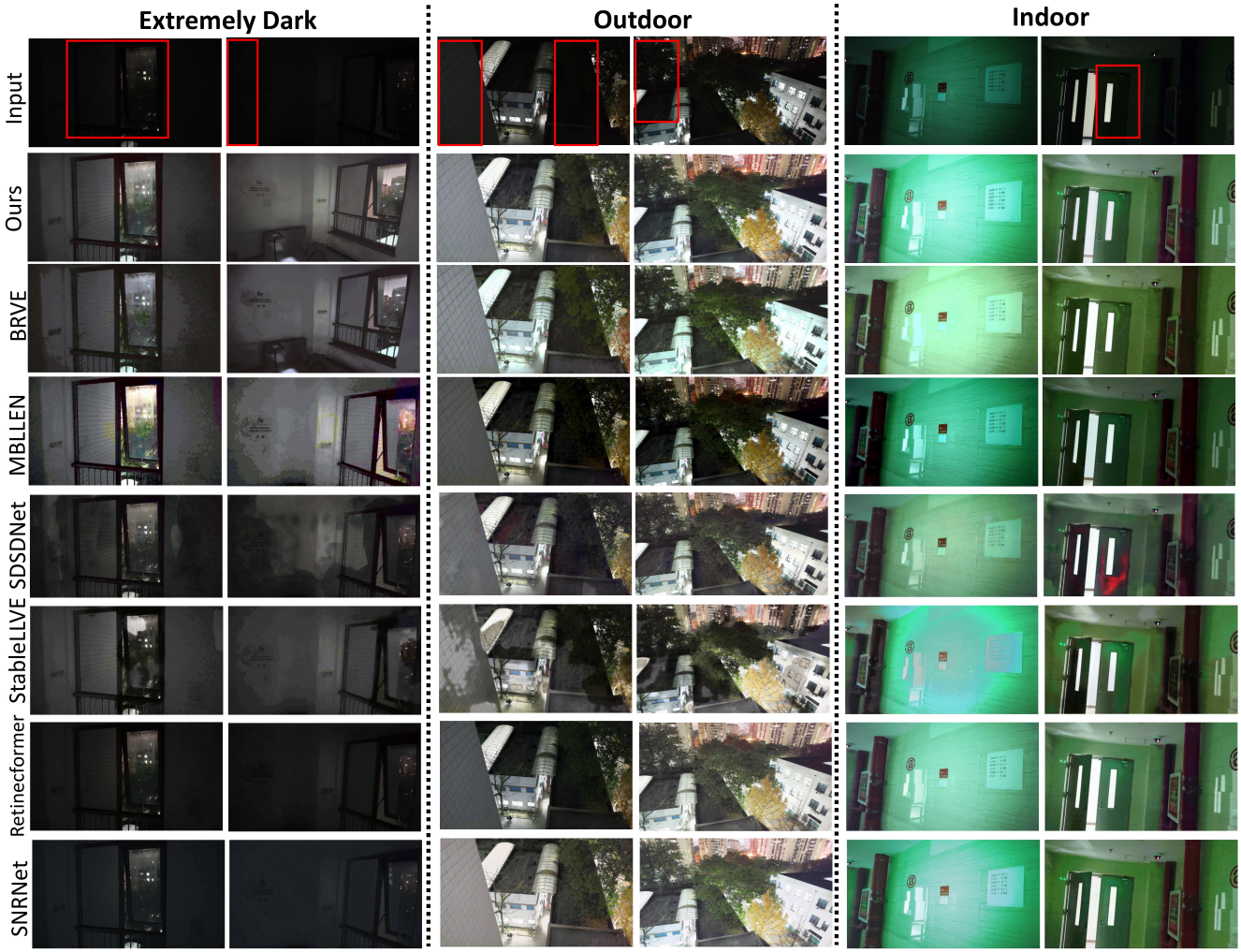


Figure 3. Visual comparisons of state-of-the-art LLVE and LLIE methods on real-shot videos. We all use models trained on the DID dataset for fair comparison.

mance during model inference.

To further filter redundant information, our MC operates as a channel-level module. For a newly generated key-value pair $K_t, V_t \in \mathbb{R}^{C \times HW}$, the adaptive storage mechanism fuses high-affinity channels while selectively storing low-affinity channels. This design enables our MC effectively retain global key temporal features with minimal memory consumption.

We first randomly select a video from each of the three datasets. Next, we set the total memory size within the range of $2C \sim 7C$ where C is the number of channels in the key and value. By default, the memory sizes for *Working Memory* and *Long-term Memory* are set at a 3:1 ratio. The average utilization is formulated as:

$$Average\ Utilization = \frac{\text{The number of times all stored channels are loaded}}{\text{The number of channels in stored key-value pairs}} \tag{1}$$

A higher average utilization indicates better inter-frame consistency. We use PSNR to measure the performance. A higher value indicates better intra-frame enhancement.

Conclusion. Efficiency analysis of our MC on in-distribution datasets is presented in Tab. 3. When the total memory size is $5C$, the average utilization on both datasets reaches the highest, indicating optimal inter-frame consistency performance. However, the total size has no direct correlation with the quality of intra-frame enhancement performance (PSNR). Therefore, in practical applications, the memory pool size should be freely selected according to the specific intra-frame and inter-frame visual experiences.

Efficiency analysis of our MC on out-of-distribution datasets is shown in Tab. 4. Even though this random video sequence with 351 frames exceeds the maximum sequence length of the training set, our MC still achieves the largest average utilization of 76.63% when the total size is $3C$. This highlights the robustness and adaptability of our

approach in handling challenging scenarios.

Table 3. Efficiency analysis of our MC on in-distribution datasets. We randomly select SDSD-out-pair9 with 287 frames and DID-video77 with 100 frames.

Total Size	SDSD		DID	
	Average Utilization	Performance	Average Utilization	Performance
$2C$	50.36%	**27.1155**	131.3%	29.1613
$3C$	49.42%	26.9687	215.4%	29.1620
$4C$	47.74%	26.9221	126.6%	29.1622
$5C$	**52.26%**	26.9943	**293.7%**	29.1622
$6C$	43.68%	27.0191	164.9%	29.1623
$7C$	39.29%	26.9680	119.4%	**29.1626**

Table 4. Efficiency analysis of our MC on out-of-distribution dataset. We randomly select LoliPhone-VID20210208191851 with 351 frames.

Total Size	Average Utilization
$2C$	64.96%
$3C$	**76.63%**
$4C$	49.97%
$5C$	55.96%
$6C$	36.46%
$7C$	31.03%

5. Hyper-parameter Analysis

In our work, we introduce three hyper-parameters, *i.e.*, α in Eq. (3), μ in Eq. (7) and τ in Alg. 2. Next, we detail their effects.

μ in Eq. (7). To alleviate overexposure, our $\mathcal{L}_{LTS}$ contains two parts: the *lightness* term H and the *texture* term T, formulated in Eq. (6). Since the TV Loss [2] makes the value of T too small, we introduce a scaling hyper-parameter μ to increase T, aligning its magnitude with that of H and ensuring a balanced contribution from both terms. Thus, by calculation, we set $\mu = 2000$ for all experiments.

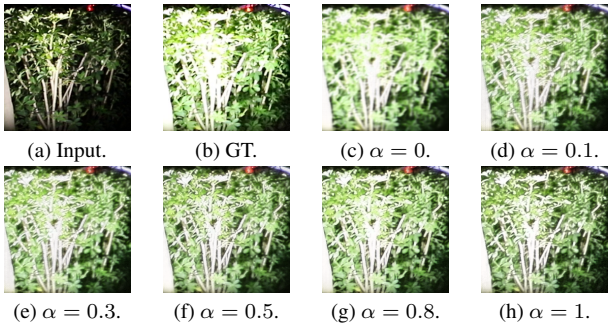
α in Eq. (3). As defined in $\mathcal{L}_{total}$, α is the weight of $\mathcal{L}_{LTS}$, which balances the *Intra-frame Regularizer* $\mathcal{L}_{LTS}$ and the *Fidelity Term*. Here, we investigate the influence of α on visual quality, as shown in Fig. 4.

We can observe that the texture level became higher along with the increase of α. However, when $\alpha > 0.5$, the influence of *Fidelity Term* begins to weaken, reducing the PSNR value. Therefore, based on the results, we set $\alpha = 0.5$ for all experiments.

Table 5. Effect of α. We test PSNR values on the SDSD dataset.

α	0	0.1	0.3	0.5	0.8	1
PSNR	27.71	27.73	27.77	**27.81**	27.80	27.78

τ in Alg. 2. It thresholds homogeneous content averaging in the adaptive storage strategy. Our MC adaptively updates the memory pool based on this threshold. A high value reduces fusion and weakens consistency, while a low value increases fusion and lowers efficiency. The threshold of 0.9 is chosen as a trade-off between efficiency and performance.



(a) Input. (b) GT. (c) $\alpha = 0$. (d) $\alpha = 0.1$. (e) $\alpha = 0.3$. (f) $\alpha = 0.5$. (g) $\alpha = 0.8$. (h) $\alpha = 1$.

Figure 4. Visual results of different settings for α in Eq. (3).

6. Failure Case Analysis

Due to the inherent limitations of camera hardware, including aperture constraints and limited dynamic range, certain visual artifacts are inevitable in the output frames, such as halos, stiff light and shadows, and vignetting.

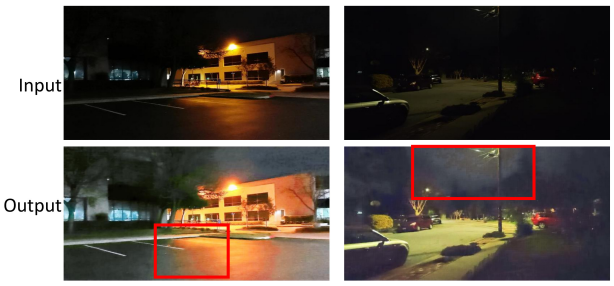


Figure 5. Failure cases on the Loli-Phone [7] dataset.

Moreover, as pointed out by the reviewer, checkerboard artifacts are observed in Fig. 2. We speculate that employing TV Loss (Eq. (7) in the main paper) to measure texture complexity may inadvertently harm image structure. Future work will explore more effective alternatives to assess texture complexity without introducing artifacts.

References

[1] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12504–12513, 2023. 1, 2

[2] Stanley H. Chan, Ramsin Khoshabeh, Kristofor B. Gibson, Philip E. Gill, and Truong Q. Nguyen. An aug-

mented lagrangian method for total variation video restoration. *IEEE Transactions on Image Processing*, 20(11):3097–3111, 2011. 4

[3] Chen Chen, Qifeng Chen, Minh N. Do, and Vladlen Koltun. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[4] Shivam Chhirolya, Sameer Malik, and Rajiv Soundararajan. Low light video enhancement by learning on static videos with cross-frame attention. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 2

[5] Huiyuan Fu, Wenkai Zheng, Xicong Wang, Jiaxuan Wang, Heng Zhang, and Huadong Ma. Dancing in the dark: A benchmark towards general low-light video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12877–12886, 2023. 1, 2

[6] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[7] Chongyi Li, Chunle Guo, Ling-Hao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. Low-light image and video enhancement using deep learning: a survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, 2021. 1, 4

[8] Guoqiang Liang, Kanghao Chen, Hangyu Li, Yunfan Lu, and Lin Wang. Towards robust event-guided low-light image enhancement: A large-scale real-world event-image dataset and novel approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23–33, 2024. 2

[9] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *BMVC*, page 4. Northumbria University, 2018. 2

[10] Yiqi Shi, Duo Liu, Liguo Zhang, Ye Tian, Xuezhi Xia, and Xiaojing Fu. Zero-ig: Zero-shot illumination-guided joint denoising and adaptive enhancement for low-light images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3015–3024, 2024. 2

[11] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. 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 (ICCV)*, pages 9700–9709, 2021. 1, 2

[12] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex C. Kot. Low-light image enhancement with normalizing flow. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 2604–2612. AAAI Press, 2022. 2

[13] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17714–17724, 2022. 2

[14] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4967–4976, 2021. 1, 2

[15] Gengchen Zhang, Yulun Zhang, Xin Yuan, and Ying Fu. Binarized low-light raw video enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25753–25762, 2024. 1, 2