A RNN for Temporal Consistency in Low-Light Videos Enhanced by Single-Frame Methods

Claudio Rota, Marco Buzzelli, Simone Bianco, and Raimondo Schettini

***Abstract*—Low-light video enhancement (LLVE) has received little attention compared to low-light image enhancement (LLIE) mainly due to the lack of paired low-/normal-light video datasets. Consequently, a common approach to LLVE is to enhance each video frame individually using LLIE methods. However, this practice introduces temporal inconsistencies in the resulting video. In this work, we propose a recurrent neural network (RNN) that, given a low-light video and its per-frame enhanced version, produces a temporally consistent video preserving the underlying frame-based enhancement. We achieve this by training our network with a combination of a new forward-backward temporal consistency loss and a content-preserving loss. At inference time, we can use our trained network to correct videos processed by any LLIE method. Experimental results show that our method achieves the best trade-off between temporal consistency improvement and fidelity with the per-frame enhanced video, exhibiting a lower memory complexity and comparable time complexity with respect to other state-of-the-art methods for temporal consistency.**

***Index Terms*—Low-light video enhancement, temporal consistency, video processing.**

I. Introduction

LOW-light image enhancement (LLIE) is an image processing task focused on improving the quality of images taken in low-light conditions [1], [2], [3], [4]. With the increasing popularity of video data, low-light video enhancement (LLVE) has become essential for a wide range of applications, including surveillance, social media, and autonomous driving [5], [6], [7]. Compared to LLIE methods, LLVE methods can consider the temporal dependency among frames to exploit additional data that may inform the enhancement of the sequence [8]. Despite the growing demand for LLVE methods, their development has been severely limited by the lack of dynamic low-light video datasets [2]. Collecting such datasets is a challenging task, as it requires complex acquisition mechanisms to acquire aligned low-/normal-light video pairs [9]. For these reasons, a common approach to achieve LLVE is to consider video frames as independent images and enhance them using existing LLIE methods. Unfortunately, this approach may introduce temporal artifacts, such as flickering and abrupt changes in brightness or color, as the temporal relationships among frames are not considered. A possible solution to extend LLIE methods to videos is the use of post-processing methods for temporal consistency [10], [11], [12], [13], [14], which aim to convert a temporally inconsistent video into a temporally consistent one preserving as much as possible the appearance of the enhanced frames. Lai et al. [11] introduce the first learning-based method for temporal consistency, where a recurrent network is trained with a temporal consistency loss and a content-preserving loss to ensure similarity with the per-frame processed video. Zhuo et al. [12] propose Temporal Denoising Mask Synthesis Network (TDMSNet), a multi-branch recurrent network that predicts optical flow [15], a motion mask and a refinement mask. The optical flow and motion mask branches align frames while masking out regions occluded by motion, and the refinement mask corrects the result of the previous operation to improve temporal consistency. Lei et al. [14] propose Deep Video Prior (DVP), which is based on the idea that temporal inconsistency can be viewed as an overfitting problem. They propose to train a neural network directly on a video using a content-preserving loss, and to stop the network training as soon as its output is similar to the per-frame processed sequence, before temporal artifacts are overfitted. Although different methods to improve temporal consistency exist, they lack appropriate mechanisms to capture the video motion dynamics or they progressively accumulate small errors over time, introducing visible artifacts in the resulting video, as we can see in Fig. 1.

Fig. 1. Artifacts introduced by existing methods for temporal consistency. MIRNet [16] is a LLIE method. Lai et al. [11] progressively darken frames due to error accumulation (left). Both Lai et al. [11] and TDMSNet [12] introduce visible ghosting effects due to wrong frame alignment (right).

In this work, we propose a recurrent neural network that improves the temporal consistency of video frames individually enhanced by LLIE methods while preserving the underlying frame-based enhancement. Our method requires two contiguous low-light frames L_t and L_{t-1}, the current frame E_t enhanced by a LLIE method, the previous frame S_{t-1} corrected by our method, i.e. stabilized, and produces a stabilized frame S_t. We

Received 19 July 2024; revised 4 October 2024; accepted 4 October 2024. Date of publication 8 October 2024; date of current version 15 October 2024. The associate editor coordinating the review of this article and approving it for publication was Prof. Yongjie Li. *(Corresponding author: Claudio Rota.)*

The authors are with the Department of Informatics, Systems and Communication, University of Milano-Bicocca, 20126 Milan, Italy (e-mail: claudio.rota@unimib.it; marco.buzzelli@unimib.it; simone.bianco@unimib.it; raimondo.schettini@unimib.it).

Code and videos are available at https://github.com/claudiom4sir/LLTC.

Digital Object Identifier 10.1109/LSP.2024.3475969

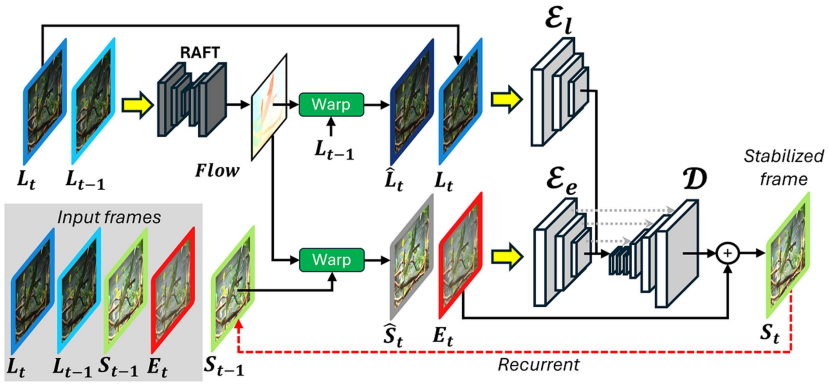

Fig. 2. Overview of the proposed method for temporal consistency. We compute the optical flow from L_{t-1} to L_t to obtain $\hat{L}_t$ and $\hat{S}_t$ via warping. We process $\hat{L}_t$ and L_t with the Low-light Encoder $\mathcal{E}_l$, and $\hat{S}_t$ and E_t with the Enhanced Encoder $\mathcal{E}_e$. The encoded features are concatenated, refined by different residual blocks, and decoded using the decoder $\mathcal{D}$. The residual is then added to E_t to obtain the stabilized frame S_t, which also serves as the new S_{t-1} for the next time step.

address the key limitations of existing methods by defining a frame alignment module that captures motion dynamics to prevent ghosting effects, and introducing a backward temporal consistency loss to avoid the progressive degradation of frame quality. Using a forward-backward temporal consistency loss and a content-preserving loss, our method learns to achieve temporal consistency without introducing new artifacts, preserving the aspect of the per-frame enhanced video. We evaluate our method on two LLVE datasets, whose videos are enhanced by different LLIE methods. The experimental results show that our method can better improve temporal consistency while preserving the underlying frame-based enhancement compared to existing methods for temporal consistency.

II. Proposed Method

We indicate with L the low-light frames, with E the frames individually enhanced by a LLIE method, and with S the frames stabilized with our method for temporal consistency. Considering a low-light video $L_{t=1}^{N}$ with N frames and its per-frame enhanced version $E_{t=1}^{N}$, our goal is to obtain a video $S_{t=1}^{N}$ that is temporally consistent and consistently similar to $E_{t=1}^{N}$. The overview of our method is shown in Fig. 2.

A. Architecture

We implement a recurrent network $\mathcal{F}$ that takes as input two contiguous low-light frames L_t and L_{t-1}, the current frame E_t enhanced by a LLIE method, and the previously stabilized frame S_{t-1} (the output of our method at the previous time step). Our method produces the stabilized frame S_t as output. Initially, we set S_1 to E_1. We first compute the forward optical flow from L_{t-1} to L_t using RAFT [17], and use it to warp L_{t-1} and S_{t-1}, obtaining $\hat{L}_t$ and $\hat{S}_t$, respectively. Then, we concatenate $\hat{L}_t$ with L_t and feed it to the Low-light Encoder $\mathcal{E}_l$. Similarly, we concatenate $\hat{S}_t$ with E_t and feed it to the Enhanced Encoder $\mathcal{E}_e$. We then concatenate the output of the two encoders and refine them using a sequence of residual blocks [18]. We then decode the output of the last residual block with the decoder $\mathcal{D}$, using skip connections from the Enhanced Encoder $\mathcal{E}_e$ via concatenation to improve reconstruction quality. Since S_t and E_t are expected to be similar in content, we compute the residual instead of the actual pixel values as:

$$S_t = E_t + \mathcal{F}(L_t, \hat{L}_t, E_t, \hat{S}_t)\ . \tag{1}$$

We use the stabilized frame S_t as input for the next time step, where it becomes the new S_{t-1}. In each time step, $\mathcal{F}$ learns to use both local and global information to ensure that S_t is temporally consistent with S_{t-1} while visually preserving the enhancement of E_t.

B. Loss Function

We optimize our method using two loss functions: the temporal consistency loss and the content-preserving loss.

The temporal consistency loss $\mathcal{L}_{TC}$ is defined as the warping error between S_t and S_{t-1}, which measures the per-pixel difference of adjacent frames after alignment via optical flow. We compute the forward optical flow from L_{t-1} to L_t using RAFT [17] and we use it to warp L_{t-1} and S_{t-1} obtaining $\hat{L}_t$ and $\hat{S}_t$. We use the occlusion mask $M_t = \exp(-\alpha||L_t - \hat{L}_t||_2^2)$ to avoid computing the loss over occluded regions as in [11]. We find that using only the loss in the forward direction, i.e. $\mathcal{L}_{TC_{fw}}$, causes small errors to accumulate at each time step, resulting in progressively darker frames. This issue is evident in existing temporal consistency methods, which are neither designed nor tested for LLVE, as illustrated in Fig. 1 (left). By incorporating the backward loss, the method enforces consistency in both directions, helping to balance these errors and preventing the gradual degradation of frame quality, ensuring that brightness and other attributes remain stable throughout the video sequence. Therefore, we also compute the loss in the backward direction, $\mathcal{L}_{TC_{bw}}$. We compute the backward flow from L_t to L_{t-1} and use it to warp L_t and S_t obtaining $\hat{L}_{t-1}$ and $\hat{S}_{t-1}$. $\hat{L}_{t-1}$ is used in occlusion mask M_{t-1}. The complete loss is defined as follows:

$$\mathcal{L}_{TC} = \mathcal{L}_{TC_{fw}} + \mathcal{L}_{TC_{bw}} = \sum_{t=2}^{T} ||M_t \odot (S_t - \hat{S}_t)||_1 + ||M_{t-1} \odot (\hat{S}_{t-1} - S_{t-1})||_1\ , \tag{2}$$

TABLE I
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS FOR TEMPORAL CONSISTENCY

LLIE methods	DID dataset [19]					SDSD dataset [2]				
	Baseline	+ Lai et al. [11]	+ TDMSNet [12]	+ DVP [14]	+ Ours	Baseline	+ Lai et al. [11]	+ TDMSNet [12]	+ DVP [14]	+ Ours
MIRNet [16] ♦	10.37 / -	3.64 / 21.25	3.29 / 26.85	4.70 / 27.16	**2.32 / 29.52**	12.40 / -	4.44 / 15.46	**4.09** / 22.88	6.44 / 29.63	4.18 / **29.93**
Kind++ [20] ♦	5.38 / -	4.85 / 26.27	4.37 / 29.25	5.12 / 31.16	**3.12 / 33.62**	16.74 / -	12.07 / 17.61	**8.44** / 23.59	14.39 / 31.72	9.19 / **34.18**
ZeroDCE++ [21] ♦	4.67 / -	3.93 / 26.54	3.45 / 30.00	4.87 / **34.51**	**2.49** / 33.66	11.88 / -	8.54 / 19.07	**5.89** / 24.31	10.08 / 34.09	6.69 / **34.57**
EnlightenGAN [22] ♦	6.25 / -	4.91 / 22.97	4.22 / 28.27	6.11 / 30.01	**3.06 / 30.98**	14.82 / -	10.18 / 17.42	**7.18** / 23.52	12.53 / 32.17	7.65 / **32.56**
BTTF [23] ♦	8.67 / -	3.22 / 21.71	3.08 / 27.83	4.08 / 26.75	**2.24 / 31.05**	9.14 / -	2.84 / 17.39	**2.67** / 24.35	4.06 / 32.42	2.76 / **33.18**
ChebyLighter [24] ♦	8.39 / -	5.97 / 22.07	4.97 / 27.35	5.17 / 29.96	**3.73 / 31.18**	12.20 / -	8.12 / 16.11	**6.62** / 22.01	9.42 / **32.88**	7.10 / 32.12
RetinexNet [25] ♦	8.34 / -	7.02 / 23.27	5.43 / 28.23	7.77 / **32.96**	**3.86** / 30.35	22.08 / -	15.05 / 16.13	**10.64** / 22.75	20.46 / **32.52**	10.78 / 30.24
URetinexNet [26] ♦	5.14 / -	3.75 / 23.58	3.30 / 27.92	5.17 / 29.96	**2.24 / 31.56**	11.10 / -	7.23 / 16.16	**6.16** / 22.32	9.78 / **33.64**	6.42 / 32.14

The baseline results refer to frames enhanced by llie methods without any stabilization, while the other results correspond to frames stabilized by a temporal consistency method. Results are reported as WE (↓) / PSNR (↑). Best results are in bold, second-best results are underlined. Our method achieves the best trade-off between temporal consistency improvement and fidelity with the per-enhanced video.

where $\odot$ is pixel-wise multiplication and M are occlusion masks. The content-preserving loss $\mathcal{L}_P$ [27] is the distance between features of the 4th layer ϕ of a pretrained VGG-16 model [28], here extracted from S_t and E_t to encourage the method to visually preserve the enhancement of E_t in S_t. The loss is defined as follows:

$$\mathcal{L}_P = \sum_{t=2}^{T} ||\phi(S_t) - \phi(E_t)||_1 \,. \tag{3}$$

The overall loss function is defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{TC} + \lambda_2 \mathcal{L}_P \,. \tag{4}$$

Note that $\mathcal{L}_{TC}$ and $\mathcal{L}_P$ optimize two contrasting aspects: the former forces a similarity between S_t and S_{t-1}, while the latter does it between S_t to E_t. Using only $\mathcal{L}_{TC}$ (i.e., $\lambda_2 = 0$) may lead to temporally consistent frames, but very different from the enhanced frames. We empirically set λ_1 to 100 and λ_2 to 0.1. We set $\alpha = 50$ in M and $T = 11$ during training.

III. EXPERIMENTS

A. Experimental Setup

We compare our method with existing video temporal consistency methods: Lai et al. [11], TDMSNet [12], and DVP [14]. We use videos from the DID dataset [19] enhanced by MIRNet [16], ZeroDCE++ [21], Kind++ [20], EnlightenGAN [22] and BTTF [23] for training and evaluation. We consider the results of these LLIE methods as our baseline. We train a single model for Lai et al. [11], TDMSNet [12] and our method considering all the above LLIE methods. Instead, we optimize multiple DVP [14] models as, by design, it needs a new model to be trained on each test video. We also evaluate the temporal consistency methods on videos enhanced by ChebyLighter [24], RetinexNet [25], and URetinexNet [26] to show generalization performance to other LLIE methods not used during training. For this purpose, we also evaluate the temporal consistency methods on videos from another dataset, i.e. SDSD [2].

Following [14], we use the Warping Error (WE) [11] between each pair (S_{t-1}, S_t) to evaluate temporal consistency. Optical flow is computed using RAFT [17]. We detect and mask out occlusions as in [29]. Lower WE values indicate better temporal consistency. We use the PSNR [30] between each pair (E_t, S_t) for the evaluation of fidelity with the per-frame enhanced sequence. Higher PSNR values indicate better fidelity.

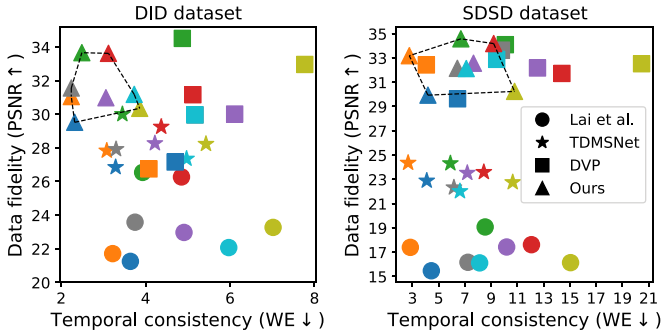

Fig. 3. Trade-off between temporal consistency improvement and fidelity with the per-frame enhanced sequences. The upper-left corner is the optimal spot (high temporal consistency, high fidelity). Each point corresponds to an entry of Table I, except for baseline (colors represent LLIE methods, see symbol colors in the table). Shapes represent temporal consistency methods. Our method (▲) is closer to the optimal spot, showing the best trade-off.

B. Quantitative and Qualitative Comparison

The quantitative comparison is reported in Table I. We can see that all the temporal consistency methods improve the temporal consistency with respect to the baseline. Lai et al. [11] obtain poor temporal consistency and fidelity performance. TDMSNet [12] is effective in improving temporal consistency but is unable to preserve the fidelity with the per-frame enhanced sequence. In contrast, DVP [14] achieves high fidelity but the resulting frames are still inconsistent. Our method represents the best trade-off: it achieves high temporal consistency with high fidelity. Indeed, it obtains the best temporal consistency performance on all the LLIE methods on the DID dataset [19] while better preserving the appearance of the enhanced frames on six out of eight LLIE methods. Observing the results on the SDSD dataset [2], we can see that the difference in temporal consistency performance between TMDSNet [12] and our method is very small, while our method is much better at preserving the fidelity with the enhanced frames. In Fig. 3, we provide a visualization of the entries from Table I to better show the trade-off between temporal consistency improvement and fidelity with the per-frame enhanced video. The points related to our method are closer to the upper-left corner, which represents the optimal spot (high temporal consistency, high fidelity).

We show a qualitative comparison in Fig. 4. For a better visualization, we blend two consecutive frames along the diagonal (i.e., upper-right vs bottom-left parts). Note that a visible

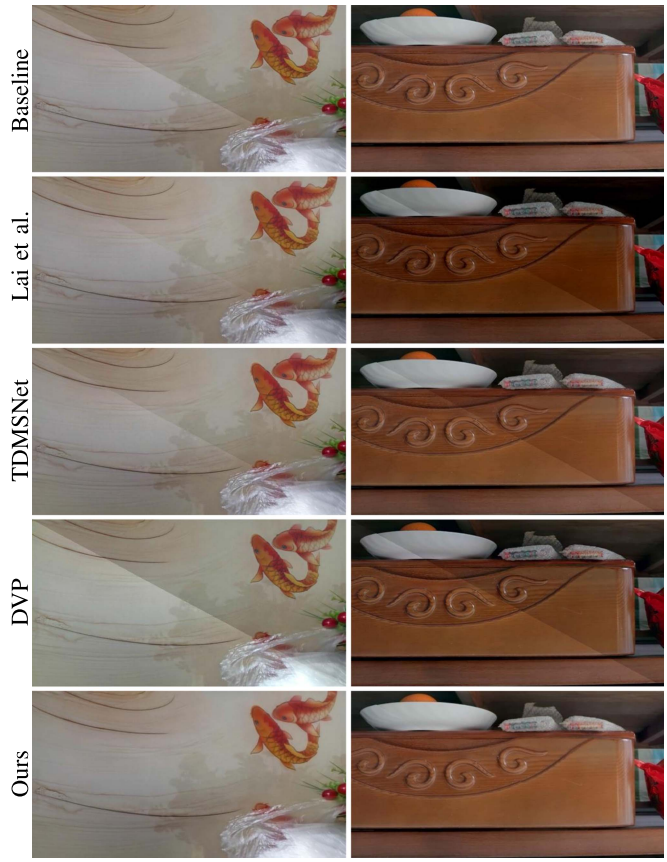

Fig. 4. Qualitative comparison with state-of-the-art methods for temporal consistency. Sequences are from the DID dataset [19]. Baseline results refer to frames enhanced by MIRNet [16] (first column) and BTTF [23] (second column). We blend two consecutive frames across the diagonal (upper-right vs bottom-left parts). A visible blending is an indicator of temporal inconsistency. The blending is not visible in our method, and the obtained frames preserve the aspect of baseline frames.

blending is an indicator of temporal inconsistency. We can see the temporal inconsistency of two consecutive frames enhanced by LLIE methods. Lai et al. [11], TDMSNet [12] and DVP [14] cannot properly improve temporal consistency, as the blending is still noticeable. In addition, Lai et al. [11] introduce artifacts that darken the corrected frames, leading to low fidelity with the baseline. In contrast, the blending is almost invisible in our method and the corrected frames appear to be very similar to the baseline, showing higher temporal consistency and fidelity, respectively.

C. Complexity Comparison

Model complexity is critical to ensure temporal consistency methods can operate on resource-limited devices. We distinguish complexity in memory and time. We use the number of model parameters and Giga FLOating-Point operations (GFLOPs) to evaluate these aspects, respectively. The results are reported in Table II. Our method is the most lightweight, with about half of the parameters compared to Lai et al. [11]. Concerning GFLOPs, our method has comparable complexity. In the last row, we can observe that 69% of the parameters are in the alignment module, which has a high impact on time complexity. Replacing the alignment module with a more efficient one may improve the overall complexity [31].

TABLE II
COMPARISON OF MODEL COMPLEXITY

Methods	Parameters	GFLOPs		
		480p	720p	1080p
Lai et al. [11]	2.89 M	109.29	247.44	560.87
TDMSNet [12]	3.47 M	150.03	339.68	769.94
DVP [14]	8.63 M	101.89	230.70	522.91
Ours	1.48 M	99.62	240.38	621.07
Ours (w/o alignment)	0.46 M	22.72	51.45	116.62

Our method is the most lightweight and has comparable efficiency.

TABLE III
ABLATION STUDY ON $\mathcal{L}_{TC_{bw}}$ AND EXPLICIT FRAME ALIGNMENT

LLIE methods	Ours	w/o $\mathcal{L}_{TC_{bw}}$	w/o align.
MIRNet [16]	2.32 / 29.52	2.25 / 24.12	2.73 / 28.26
Kind++ [20]	3.12 / 33.62	3.03 / 27.22	3.53 / 31.47
ZeroDCE++ [21]	2.49 / 33.66	2.46 / 27.32	2.79 / 32.29
EnlightenGAN [22]	3.06 / 30.98	3.00 / 24.96	3.50 / 29.53
BTTF [23]	2.24 / 31.05	2.19 / 23.66	2.54 / 29.52

Results reported as WE (↓) / PSNR (↑). With these components, we achieve better performance. The artifacts introduced without using $\mathcal{L}_{TC_{bw}}$(frames becoming darker over time) are well captured by PSNR, but not by WE.

D. Ablation Study

We conduct ablation experiments to evaluate the impact of frame alignment and the addition of backward temporal consistency loss $\mathcal{L}_{TC_{bw}}$. The results are presented in Table III. Without $\mathcal{L}_{TC_{bw}}$, the resulting frames become darker over time, as captured by the lower PSNR values. We observe the same behavior in Lai et al. [11]. Temporal consistency is not affected by this problem, as there are no abrupt changes in brightness or colors, but the transition is smooth. Indeed, completely dark frames are very consistent with each other but have very low similarity with the enhanced frames. Without explicit frame alignment, misalignment artifacts and ghosting effects negatively impact both WE and PSNR. We can conclude that both components are necessary to achieve better results.

IV. CONCLUSION

We presented a method to address the problem of temporal consistency in video frames enhanced individually by methods for low-light image enhancement (LLIE). Our method improves temporal consistency and preserves the underlying frame-based enhancement regardless of the LLIE method applied. We compared our method with existing methods for temporal consistency, using two low-light video datasets enhanced by eight LLIE methods. The results showed our method achieves the best trade-off between temporal consistency improvement and fidelity with the per-frame enhanced video. Moreover, it has a lower memory complexity and comparable time complexity. In future works, we plan to extend our work to other tasks such as video color constancy [32], and other techniques such as content deformation fields [33].

REFERENCES

[1] W. Kim, "Low-light image enhancement: A comparative review and prospects," *IEEE Access*, vol. 10, pp. 84535–84557, 2022.

[2] R. Wang, X. Xu, C.-W. Fu, J. Lu, B. Yu, and J. Jia, "Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9700–9709.

[3] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, Apr. 2018.

[4] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3291–3300.

[5] S. Ai and J. Kwon, "Extreme low-light image enhancement for surveillance cameras using attention U-net," *Sensors*, vol. 20, no. 2, 2020, Art. no. 495.

[6] L. H. Pham, D.-N. Tran, and J. W. Jeon, "Low-light image enhancement for autonomous driving systems using DriveRetinex-Net," in *Proc. IEEE Int. Conf. Consum. Electron.-Asia*, 2020, pp. 1–5.

[7] L. Fu, H. Yu, F. Juefei-Xu, J. Li, Q. Guo, and S. Wang, "Let there be light: Improved traffic surveillance via detail preserving night-to-day transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8217–8226, Dec. 2022.

[8] C. Li et al., "Low-light image and video enhancement using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9396–9416, Dec. 2022.

[9] C. Chen, Q. Chen, M. N. Do, and V. Koltun, "Seeing motion in the dark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3185–3194.

[10] N. Bonneel, J. Tompkin, K. Sunkavalli, D. Sun, S. Paris, and H. Pfister, "Blind video temporal consistency," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–9, 2015.

[11] W.-S. Lai, J.-B. Huang, O. Wang, E. Shechtman, E. Yumer, and M.-H. Yang, "Learning blind video temporal consistency," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 170–185.

[12] Y. Zhou, X. Xu, F. Shen, L. Gao, H. Lu, and H. T. Shen, "Temporal denoising mask synthesis network for learning blind video temporal consistency," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 475–483.

[13] H. Thimonier, J. Despois, R. Kips, and M. Perrot, "Learning long term style preserving blind video temporal consistency," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[14] C. Lei, Y. Xing, H. Ouyang, and Q. Chen, "Deep video prior for video consistency and propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 356–371, Jan. 2023.

[15] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, no. 1–3, pp. 185–203, 1981.

[16] S. W. Zamir et al., "Learning enriched features for fast image restoration and enhancement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1934–1948, Feb. 2023.

[17] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 402–419.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[19] H. Fu, W. Zheng, X. Wang, J. Wang, H. Zhang, and H. Ma, "Dancing in the dark: A benchmark towards general low-light video enhancement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 12877–12886.

[20] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *Int. J. Comput. Vis.*, vol. 129, pp. 1013–1037, 2021.

[21] C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4225–4238, Aug. 2022.

[22] Y. Jiang et al., "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.

[23] S. Zini, C. Rota, M. Buzzelli, S. Bianco, and R. Schettini, "Back to the future: A night photography rendering ISP without deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2023, pp. 1465–1473.

[24] J. Pan, D. Zhai, Y. Bai, J. Jiang, D. Zhao, and X. Liu, "ChebyLighter: Optimal curve estimation for low-light image enhancement," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1358–1366.

[25] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *Proc. Brit. Mach. Vis. Conf.*, Newcastle, U.K.: BMVA Press, Sep. 3–6, 2018, p. 155. [Online]. Available: http://bmvc2018.org/contents/papers/0451.pdf

[26] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang, "URetinex-Net: Retinex-based deep unfolding network for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5901–5910.

[27] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, May 7–9, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[29] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos," in *Proc. German Conf. Comput. Vis.*, 2016, pp. 26–36.

[30] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, 2010, pp. 2366–2369.

[31] L. Kong, C. Shen, and J. Yang, "FastFlowNet: A lightweight network for fast optical flow estimation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 10310–10316.

[32] M. Buzzelli and I. Erba, "On the evaluation of temporal and spatial stability of color constancy algorithms," *J. Opt. Soc. Amer. A*, vol. 38, no. 9, pp. 1349–1356, 2021.

[33] H. Ouyang et al., "CoDeF: Content deformation fields for temporally consistent video processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 8089–8099.