Zero-TIG: Temporal Consistency-Aware Zero-Shot Illumination-Guided Low-light Video Enhancement

1st Yini Li
Visual Information Laboratory, University of Bristol
Bristol, UK
ub24017@bristol.ac.uk

2nd Nantheera Anantrasirichai
Visual Information Laboratory, University of Bristol
Bristol, UK
N.Anantrasirichai@bristol.ac.uk

***Abstract*—Low-light and underwater videos suffer from poor visibility, low contrast, and high noise, necessitating enhancements in visual quality. However, existing approaches typically rely on paired ground truth, which limits their practicality and often fails to maintain temporal consistency. To overcome these obstacles, this paper introduces a novel zero-shot learning approach named Zero-TIG, leveraging the Retinex theory and optical flow techniques. The proposed network consists of an enhancement module and a temporal feedback module. The enhancement module comprises three subnetworks: low-light image denoising, illumination estimation, and reflection denoising. The temporal enhancement module ensures temporal consistency by incorporating histogram equalization, optical flow computation, and image warping to align the enhanced previous frame with the current frame, thereby maintaining continuity. Additionally, we address color distortion in underwater data by adaptively balancing RGB channels. The experimental results demonstrate that our method achieves low-light video enhancement without the need for paired training data, making it a promising and applicable method for real-world scenario enhancement. Code is available at https://github.com/liyinibristol/Zero-TIG.**

***Index Terms*—video enhancement, zero-shot learning, low-light, optical flow, underwater**

I. Introduction

Low-light conditions significantly challenge video quality. Camera sensors in dim environments capture fewer photons, crucial for image clarity and brightness, resulting in darker, less detailed footage. To compensate, cameras increase ISO sensitivity, which amplifies light signals but also heightens noise. Moreover, increasing exposure time can brighten videos but causes motion blur. These factors degrade color accuracy, contrast, and sharpness, making videos appear washed out and poorly defined. This issue is exacerbated in videos captured in deep water, where color distortion occurs due to wavelength-dependent light attenuation [1]. These distortions affect not only visibility, but also decision-making, and automation in many applications like security, surveillance, autonomous vehicles, medical imaging, and remote sensing.

The performance of low-light *image* enhancement has significantly improved with modern deep learning techniques, and research in this field continues to advance [2]. However, applying these *image*-based methods to low-light *videos* introduces temporal flickering. Some supervised learning approaches extend to video by utilizing multiple frames as input [3], yet the effectiveness of low-light video enhancement remains limited. A major challenge is the lack of high-quality video-pair datasets, particularly for underwater scenes, which makes supervised deep learning limits its practicality. This underscores the need for self-supervised solutions to address these challenges effectively.

In this paper, we propose a novel self-supervised learning method, enabling high-quality video enhancement with only a single input video for training. Specifically, we introduce a novel zero-shot method, integrated with Retinex theory [4], for real-world low-light video enhancement that simultaneously enhances contrast, reduces noise and corrects color at the same time. Additionally, our method recursively processes the reflectance and illumination components over time, ensuring temporal consistency and improved visual quality. In summary, the main contributions of this paper are as follows:

- We proposed a new zero-shot learning method, Zero-TIG, for low-light enhancement. It combines an enhancement module and a temporal feedback module, achieving visual quality without paired data.
- We designed a feedback module that combining histogram equalization, optical flow (OF) estimation and image warping to recursively incorporate the enhanced results from the previous frame into the current pipeline, leading to less image noise and flickering.
- For underwater data, we address the color distortion by computing the gain for each RGB channel. This ensures the illumination is constrained according to the average of individual channels and achieving the white balance automatically.

II. Related work

A. Low-Light Video Enhancement

Low-light video enhancement process typically begins with aligning feature maps of neighboring frames to the current frame [5], followed by iterative feature re-weighting to reduce merging errors [6]. Recent advances include a light-adjustable network based on Retinex theory [7] and a deep unfolding model using MAP optimization [8]. State-of-the-art performance has been achieved by a wavelet conditional diffusion model [9], surpassing CNN- and transformer-based methods. Due to limited paired datasets, unpaired approaches

This work was supported by the UKRI MyWorld Strength in Places Programme (SIPF00006/1) and EPSRC ECR (EP/Y002490/1).

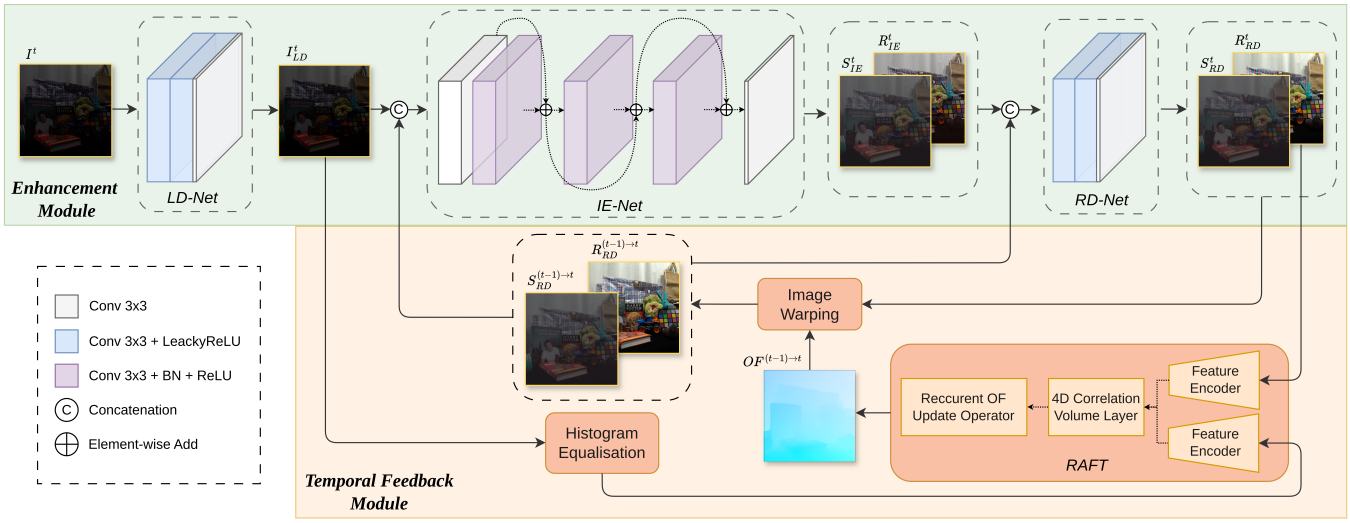

Fig. 1: The proposed network includes an enhancement module with LD-Net, IE-Net, and RD-Net, and a temporal feedback module that employs histogram equalization, RAFT network for OF computation, and image warping. R^t_{RD} is the final output.

such as CycleGAN [10] and EnlightenGAN [11] have also been explored.

B. Zero-Shot Learning for Enhancement

Zero-shot learning addresses low-light enhancement without paired data. Zero-DCE [12] introduced image-specific curve estimation using a lightweight network and non-reference losses. Subsequent work includes SGZSL [13], which integrates semantic segmentation, and RUAS [14], which combines Retinex theory with architecture search. [15] proposed a three-stage model for iterative enhancement, while Zero-IG [16] extended Noise2Noise [17] for joint denoising and enhancement. Although [18] leverages temporal information to reduce flickering, its pixel-wise frame combination introduces ghosting artifacts. Notably, [13] processes frames independently, limiting its applicability to video enhancement.

III. Methodology

The proposed network is shown in Fig. 1, consisting of two main modules: the enhancement module and the novel temporal feedback module. The enhancement module is built on the basis of Zero-IG [16], and the temporal feedback module warps the previous enhanced output to the current frame, ensuring temporal consistency across the video sequence.

A. Enhancement module

Following [16], the enhancement module comprises three subnetworks which are the low-light denoising network (LD-Net), the illumination estimation network (IE-Net), and the reflection denoising network (RD-Net).

To mitigate the impact of noise on the illumination distribution, the low-light image I^t at time t undergoes an initial denoising process. The LD-Net utilizes two downsamplers, G_1 and G_2, adopted from the framework proposed in [19] to generate a preliminary denoised image, denoted as I^t_{LP}. Inspired by Retinex theory, the IE-Net then decomposes I^t_{LP} into two components: reflectance R^t_{IE} and illumination S^t_{IE} as described by the following element-wise multiplication:

$$I^t_{LP} = R^t_{IE} \cdot S^t_{IE}, \tag{1}$$

where R^t_{IE} and S^t_{IE} represent the intrinsic properties of the scene and the lighting conditions at time t, respectively. To further enhance the performance of the denoising, the reflectance R^t_{IE} and illumination S^t_{IE} are concatenated and fed into the RD-Net. This subnetwork refines the output to produce R^t_{RD} and S^t_{RD}, effectively removing residual noise and improving the overall quality of the enhanced image. The final output from the network is R^t_{RD}, which represents the enhanced frame of the video.

B. Temporal feedback module

Although the enhancement network can achieve relatively good enhancement results, for videos, directly using single frames can lead to flickering. Consequently, we introduce the temporal feedback module to address the temporal consistency between consecutive frames in a video sequence.

The core concept of this module is to align the output result of the previous frame to the current frame, and then feed the aligned results back into IE-Net and RD-Net for continuous refinement. To achieve the alignment of the two frames, we used a highly efficient RAFT [20] pre-trained model to calculate the optical flow (OF). Instead of directly applying the original noisy input I^t, we utilize the pre-denoised image I^t_{LD}, which reduces mismatches in optical flow computation caused by noise. Moreover, we apply histogram equalization to I^t_{LD} to align its intensity distribution with that of the enhanced frame, ensuring a more accurate optical flow estimation. We denote the histogram equalization process as $HE(\cdot)$, and output as $\hat{I}^t_{LD}$:

$$\hat{I}^t_{LD} = HE(I^t). \tag{2}$$

Then, we input R_{RD}^{t-1}, which is the final output of the network, and $\hat{I}_{LD}^{t}$ into the RAFT network to compute the optical flow displacement map, denoted as $OF^{(t-1)\to t}$:

$$OF^{(t-1)\to t} = RAFT(R_{RD}^{t-1},\ \hat{I}_{LD}^{t}) \tag{3}$$

In our experiments, we observed that weak texture regions in the images often lead to mismatches in the optical flow computation. To address this issue, we downsample both images by a factor of 3 along their height and width. This downsampling strategy not only mitigates the impact of noise, but also significantly accelerates the computation process.

Based on $OF^{(t-1)\to t}$, the image warping is implemented in both R_{RD}^{t-1} and S_{RD}^{t-1}. These warped results, denoted as $R_{RD}^{(t-1)\to t}$ and $S_{RD}^{(t-1)\to t}$, are then fed into the IE-Net concatenated with I_{LD}^{t}. Meanwhile, $R_{RD}^{(t-1)\to t}$ and $S_{RD}^{(t-1)\to t}$ are also fed into RD-Net concatenated with R_{IE}^{t} and S_{IE}^{t}. For the first frame in a sequence, $R_{RD}^{(t-1)\to t}$ and $S_{RD}^{(t-1)\to t}$ are initialised as zero vectors. By using optical flow estimation and warping techniques, the context information from previous frames is fused to assist the denoising of the current frame and achieve smooth transitions.

C. Loss functions

Following [16], our method incorporates a total of 11 loss functions to optimize the model. Inspired by [19], the I_{LP}^{t} is downsampled into two subimages and is self-supervised by minimizing the difference between these subimages through a residual loss L_{res1} and a consistency loss L_{cons1}. Given the two downsamplers denoted as (G_1, G_2), the noise predicted by LD-Net as $f_{LD}()$, and I as input I_{LD}^{t} in short, L_{res1} and L_{cons1} is described as:

$$\begin{aligned} L_{res1} = &||G_1(I) - f_{LD}(G_1(I)) - G_2(I)||_2^2 \\ &+ ||G_2(I) - f_{LD}(G_2(I)) - G_1(I)||_2^2 \end{aligned} \tag{4}$$

$$\begin{aligned} L_{cons1} = &||G_1(I) - f_{LD}(G_1(I)) - G_1(I - f_{LD}(I))||_2^2 \\ &+ ||G_2(I) - f_{LD}(G_2(I)) - G_2(I - f_{LD}(I))||_2^2\ , \end{aligned} \tag{5}$$

Three constraints are applied for illumination estimation S_{IE}^{t}. Specifically, the overall mean error loss L_{over} which computes the L2 loss between S_{IE}^{t} and a predefined brightness coefficient α is leveraged:

$$L_{over} = ||S_{IE}^{t} - \alpha^{-1}||_2^2, \tag{6}$$

where $\alpha = 0.5Y_L^{-1}$ is set by [16]. Y_L indicates the mean value of I_{LD}^{t}.

To achieve amplitude adjustments for varying intensities, the pixel-wise adjustment loss L_{pix} is expressed as:

$$L_{pix} = ||S_{IE}^{t} - \beta(\alpha I_{LD}^{t})^{\alpha}||_2^2, \tag{7}$$

where the scaling factor is defined as $\beta = \alpha^{-1}0.7^{-\alpha}$ in [16].

The illumination is additionally constrained by the smoothness loss L_{smooth}, which regularizes the L1 loss of the horizontal and vertical gradient of S_{IE}^{t}. In RD-Net, R_{RD}^{t} and S_{RD}^{t} are downsampled and optimized using a residual loss L_{res2} and a consistency loss L_{cons2}. Additionally, S_{RD}^{t} is further constrained by an illumination consistency loss L_{ill} which is the mean squared error (MSE) between S_{RD}^{t} and S_{IE}^{t}. The interactive denoising loss L_{inter} is used, which identifies noise by comparing the brightness channel differences between two downsampled versions of R_{RD}^{t}. Furthermore, to refine the reflectance estimation, a local variance loss L_{var} and a color loss L_{color} are introduced to confine the variance and color change between R_{RD}^{t} and R_{IE}^{t}.

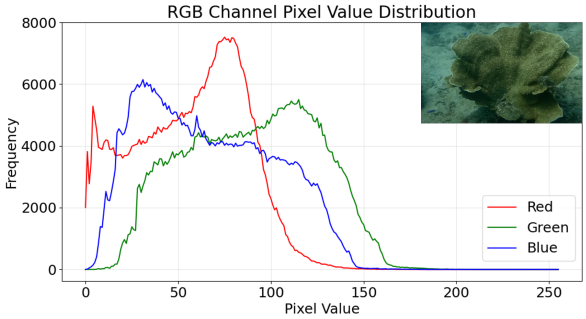

Fig. 2: Frequency distribution of a underwater image.

Adaptive white balance for underwater data. The establishment of L_{over} and L_{pix} is based on the assumption that the distributions of the RGB channels of input are approximately uniform. However, as shown in Fig. 2, this assumption does not hold for underwater data because of the absorption and scattering of light in the water. Inspired by the Chromatic Retinex in [21], we compute the mean value of luminance plane Y_L for each RGB channel independently, ensuring that all channels in S_{IE}^{t} are constrained individually. Accordingly, we extend (6) and (7) as:

$$L_{over} = \sum_{c} ||S_{IEc}^{t} - \alpha_c^{-1}||_2^2, \tag{8}$$

$$L_{pix} = \sum_{c} ||S_{IEc}^{t} - \beta_c(\alpha_c I_{LD}^{t})^{\alpha_c}||_2^2, \tag{9}$$

where $c \in \{R, G, B\}$, and we empirically set $\alpha_c = 0.3Y_{Lc}^{-1}$, as the brightness of the underwater footage is generally higher than that of low-light test videos.

In conclusion, the final loss function is defined as $L_{total} = L_{res1} + L_{cons1} + L_{over} + L_{pix} + L_{smooth} + L_{res2} + L_{cons2} + L_{ill} + L_{inter} + L_{var} + L_{color}$.

IV. Experimental Results and Discussion

A. Datasets and implementation

The BVI-RLV dataset [22] consists of paired videos featuring low-light footage and their corresponding normal-light versions. The sequences with normal light (100%) are used as references for evaluation. We utilized 16 sequences from the BVI-RLV dataset to evaluate our proposed method. Each sequence is in RGB format (HD resolution, 25 fps). The contents are varied and dynamic, with different camera and object movement speeds.

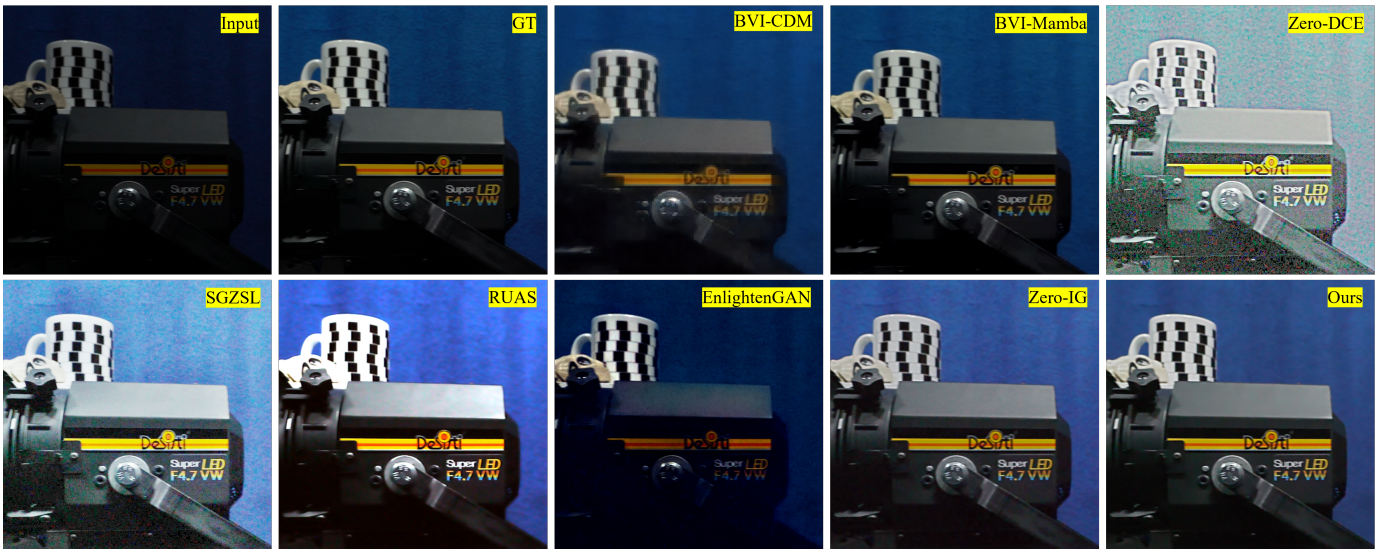

Fig. 3: Comparison of visual results of our proposed Zero-TIG with supervised methods and self-supervised methods on BVI-RLV dataset.

We tested our proposed method on an underwater video from the BVI-Coral dataset [23], which was captured at a depth of 10 meters, where the effects of low light and noise are present.

The Adam optimizer was employed with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a weight decay of 3×10^{-4} and a learning rate of 10^{-4} on an NVIDIA V100 GPU. The RAFT network used a pre-trained model from the Sintel dataset [24], with frames downsampled by a factor of 3 to improve efficiency and accuracy. For BVI-RLV, Zero-TIG was initialized with a pre-trained weights on Zero-IG for 5 epochs and fine-tuned for another 5 epochs. For BVI-Coral, we modified the loss functions (6) (7) to (8) (9) and trained for 5 epochs.

B. Performance for low-light video enhancement

We compared the performance of our proposed Zero-TIG method with five state-of-the-art self-supervised learning methods for low-light enhancement: Zero-DCE [12], SGZSL [13], RUAS [14], Zero-IG [16], and EnlightenGAN [11]. Other methods proposed for video enhancement so far, e.g. [8] and [18], do not provide available codes, and we were unable to reproduce their results. We also include the results from supervised learning (BVI-CDM [9] and BVI-Mamba [25]) as upper-bound references.

Table I presents the results, evaluated using well-known metrics: PSNR, SSIM, and LPIPS, as the BVI-RLV dataset provides normal light references. Also, since brightness can be subjective, we included results after applying histogram matching (HM) to the references to remove brightness differences, allowing us to evaluate the denoising performance of each method more accurately. These results demonstrate that our Zero-TIG outperforms most methods and is relatively comparable to Zero-IG. The EnlightenGAN is prone to generating hallucinated color artifacts in images, as subjectively evaluated and illustrated in Fig. 3. Compared to Zero-IG, our method demonstrates superior denoising performance. Although supervised methods achieve higher quantitative metrics, our visual results are comparable to, and in some cases even surpass, those of supervised approaches, outperforming BVI-CDM in terms of visual quality.

TABLE I: Comparison of PSNR, SSIM, and LPIPS across different methods with and without histogram matching (HM)

	PSNR		SSIM		LPIPS	
	w/o HM	w/ HM	w/o HM	w/ HM	w/o HM	w/ HM
BVI-CDM*	30.51	-	0.888	-	0.089	-
BVI-Mamba*	31.22	-	0.912	-	0.071	-
Zero-DCE	10.540	18.932	0.430	0.488	0.528	0.507
SGZSL	13.416	24.026	0.577	0.723	0.420	0.380
RUAS	15.305	18.520	0.631	0.712	0.481	0.515
EnlightenGAN	15.486	17.875	0.518	0.550	0.515	0.522
Zero-IG	**19.374**	27.840	0.639	0.834	0.398	0.370
Zero-TIG (ours)	19.340	**28.052**	**0.790**	**0.854**	**0.360**	**0.368**

* Supervised learning methods

We also evaluated temporal consistency using Mean Absolute Brightness Differences (MABD), as proposed in [26]. A lower MABD value indicates better sequential continuity. For comparative analysis, we selected a representative video named *S11_gift_wrap* with 10% illuminance and computed the MABD values across consecutive frames. To enhance clarity, we applied a moving average technique with a window size of 15. Fig. 4 illustrates the flickering effects of different methods after enhancement. Our approach exhibits markedly smaller amplitude, indicating more stable brightness variations over time. which confirm that our method demonstrates the superior temporal coherence of our method compared to Others.

C. Performance of low-light underwater video enhancement

In this experiment, we compared our method with Zero-IG. Since ground truth for the underwater scene is unavailable, we used non-reference metrics and subjective assessment. The evaluation metrics include two underwater evaluation

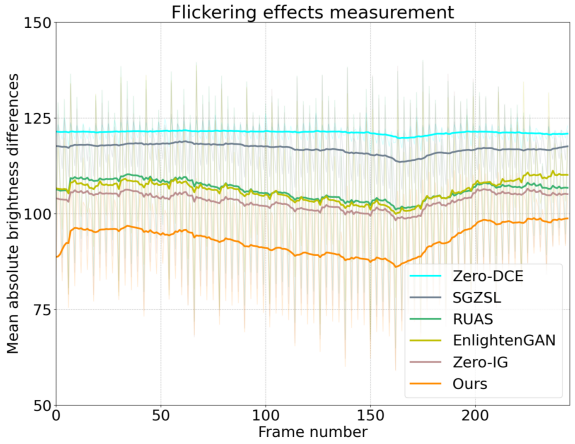

Fig. 4: MABD vectors of Zero-IG and our method. Moving average method is applied to data for clarity.

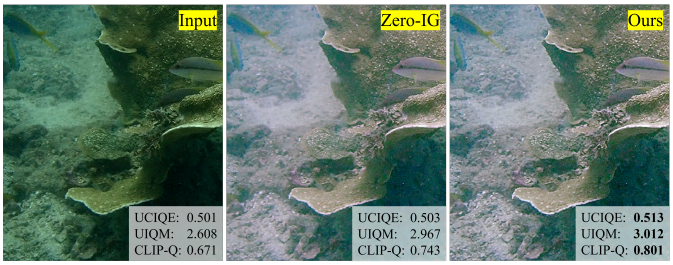

Fig. 5: Results of underwater video enhancement

metrics UIQM [27] and UCIQE [28], and the quality score in CLIP-IQA [29]. Higher values of these metrics indicate better image quality. We selected a representative video, *Original210120.MOV*, from [23] for evaluation and visualized the results of one frame, as shown in Fig. 5.

Our results demonstrate superior contrast compared to Zero-IG, with clearer edges of coral and sand in the background, as illustrated in Fig. 5. Quantitatively, our method achieves higher scores across all non-reference metrics. These improvements confirm that our approach effectively addresses distortions in underwater videos and achieves adaptive white balance, enhancing both visual quality and metric performance.

V. Conclusion

This paper proposes Zero-TIG, a zero-shot self-supervised method for low-light video enhancement that integrates Retinex theory with a novel temporal feedback module to reduce flickering and noise. Additionally, an adaptive white balance is introduced for underwater data by constraining RGB channels of illumination, achieving superior temporal coherence and color accuracy without paired training data.

References

[1] S. Lu, F. Guan, H. Zhang, and H. Lai, "Underwater image enhancement method based on denoising diffusion probabilistic model," *Journal of Visual Communication and Image Representation*, vol. 96, 2023.

[2] N. Anantrasirichai and D. Bull, "Artificial intelligence in the creative industries: a review," *Artificial Intelligence Review*, p. 589–656, 2022.

[3] R. Lin, N. Anantrasirichai, A. Malyugina, and D. Bull, "A spatio-temporal aligned SUNet model for low-light video enhancement," in *IEEE International Conference on Image Processing*, 2024.

[4] E. H. Land and J. J. McCann, "Lightness and retinex theory," *Journal of the Optical Society of America*, vol. 61, no. 1, pp. 1–11, 1971.

[5] R. Wang, X. Xu, C.-W. Fu, J. Lu, B. Yu, and J. Jia, "Seeing dynamic scene in the dark: High-quality video dataset with mechatronic alignment," in *ICCV*, 2021.

[6] K. Zhou, W. Li, L. Lu, X. Han, and J. Lu, "Revisiting temporal alignment for video restoration," in *CVPR*, 2022.

[7] 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 *IEEE/CVF International Conference on Computer Vision*, 2023.

[8] L. Zhu, W. Yang, B. Chen, H. Zhu, Z. Ni, Q. Mao, and S. Wang, "Unrolled decomposed unpaired learning for controllable low-light video enhancement," in *ECCV*, 2024.

[9] R. Lin, Q. Sun, and N. Anantrasirichai, "Low-light video enhancement with conditional diffusion models and wavelet interscale attentions," in *21st ACM SIGGRAPH Conference on Visual Media Production*, 2024.

[10] N. Anantrasirichai and D. Bull, "Contextual colorization and denoising for low-light ultra high resolution sequences," in *IEEE ICIP*, 2021.

[11] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *TIP*, vol. 30, pp. 2340–2349, 2021.

[12] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *CVPR*, June 2020.

[13] S. Zheng and G. Gupta, "Semantic-guided zero-shot learning for low-light image/video enhancement," in *WACV*, 2022, pp. 581–590.

[14] L. Risheng, M. Long, Z. Jiaao, F. Xin, and L. Zhongxuan, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *CVPR*, 2021.

[15] M. Saeed and M. Torki, "Lit the Darkness: Three-stage zero-shot learning for low-light enhancement with multi-neighbor enhancement factors," in *IEEE ICASSP*, 2023, pp. 1–2.

[16] Y. Shi, D. Liu, L. Zhang, Y. Tian, X. Xia, and X. Fu, "ZERO-IG: Zero-shot illumination-guided joint denoising and adaptive enhancement for low-light images," in *IEEE/CVF CVPR*, 2024, pp. 3015–3024.

[17] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2Noise: Learning image restoration without clean data," in *ICML*, vol. 80, 2018, pp. 2965–2974.

[18] X. Han, T. Bao, and H. Yang, "Exploring fast and flexible zero-shot low-light image/video enhancement," *Computer Graphics Forum*, vol. 43, no. 7, pp. 327–338, 2024.

[19] Y. Mansour and R. Heckel, "Zero-Shot Noise2Noise: Efficient image denoising without any data," in *IEEE/CVF CVPR*, 2023.

[20] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *ECCV*, 2020, pp. 402–419.

[21] H. Huang, W. Yang, L. Duan, and J. Liu, "Seeing dark videos via self-learned bottleneck neural representation," *AAAI*, vol. 38, no. 3, 2024.

[22] R. Lin, N. Anantrasirichai, G. Huang, J. Lin, Q. Sun, A. Malyugina, and D. R. Bull, "BVI-RLV: A fully registered dataset and benchmarks for low-light video enhancement," *arXiv preprint arXiv:2401.10166*, 2024.

[23] L. Gough, A. Azzarelli, F. Zhang, and N. Anantrasirichai, "AquaNeRF: Neural radiance fields in underwater media with distractor removal," in *ISCAS*, 2025.

[24] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *ECCV*, 2012, pp. 611–625.

[25] G. Huang, R. Lin, D. R. Bull, and N. Anantrasirichai, "BVI-Mamba: Video enhancement using a visual state-space model for low-light and underwater environments," in *SPIE Defense + Commercial Sensing (DCS25)*, 2025.

[26] H. Jiang and Y. Zheng, "Learning to see moving objects in the dark," in *ICCV*, 2019, pp. 7323–7332.

[27] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE Journal of Oceanic Engineering*, vol. 41, no. 3, pp. 541–551, 2016.

[28] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *TIP*, vol. 24, no. 12, pp. 6062–6071, 2015.

[29] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *AAAI*, 2023.