Computer Vision and Image Understanding 240 (2024) 103916

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Adaptive Locally-Aligned Transformer for low-light video enhancement

Yiwen Cao [a], Yukun Su [a], Jingliang Deng [a], Yu Zhang [c], Qingyao Wu [a,b,*]

[a] School of Software Engineering, South China University of Technology, Guangzhou, China
[b] Pazhou Lab, Guangzhou, China
[c] Shenzhen Santachi Video Technology Co.,Ltd., Shenzhen, China

ARTICLE INFO

Communicated by Jinshan Pan

Keywords:
Low-light enhancement
Vision transformer
Adaptive align
Spatial-temporal sequence

ABSTRACT

Low-light enhancement is a crucial task that aims to enhance the under-exposed input in computer vision. While state-of-the-art static single-image enhancement methods have made remarkable progress, yet, few attempts are explored the spatial-temporal sequence problem in low-light video enhancement. In this paper, we propose a simple yet highly effective method, termed as **A**daptive **L**ocally-**A**ligned **T**ransformer (ALAT) for low-light video enhancement based on visual transformers. ALAT consists of three parts: feature encoder, locally-aligned transformer block (LATB) and pyramid feature decoder. Specifically, the transformer block enables the network to model the long-range spatial and appearance dependencies in videos due to its self-attention parallel computing mechanism. However, different from some previous approaches directly using the vanilla transformer, we consider that locality is significant in low-level vision tasks since the misaligned contextual local features (*i.e.*, edges, shapes) may affect the prediction quality. Therefore, the proposed LATB is designed to align the video pixel with its most relevant ones adaptively in the local region to preserve the regional content information. Furthermore, we publish a new real-world low-light video dataset, named *ExpressWay*, to fill the gaps in the lack of dynamic low-light video scenarios, which contains high-quality videos with moving objects in both dark- and bright-light conditions. We conduct experiments on five benchmarks under three comprehensive settings including synthesized, static and our proposed dynamic low-light video datasets. Extensive experimental results show that our ALAT can outperform the previous state-of-the-arts by a large margin of **0.20~1.10 dB**. Our method can be also extended to other video enhancement applications. The project is available at https://github.com/y1wencao/LLVE-ALAT.

1. Introduction

Low-Light Enhancement (LLE) is of great significance and practical value in computer vision tasks, which has been successfully applied to some real-world applications such as computational photography (Raskar and Tumblin, 2005), video surveillance (Vishwakarma and Agrawal, 2013; Li et al., 2021) and autonomous driving (Caesar et al., 2020), *etc*. While witnessing the great progress in image LLE domain (Jiang et al., 2021; Guo et al., 2020), few works try to explore its usefulness in video domain yet, which is much more challenging due to its spatial-temporal information. Generally, to model the sequential cues in videos, some 3DConv (Tran et al., 2015) and LSTM (Hochreiter and Schmidhuber, 1997) methods are proposed. However, they may suffer from local receptive field and vanishing gradient issues. Compared to the elaborate CNN and RNN networks, the recently emerged vision transformer (Dosovitskiy et al., 2020) shows the strong ability to capture the global content with its inherent self-attention mechanism, which has achieved remarkable performance in some high-level vision tasks (*e.g.*, segmentation Xie et al., 2021; Zheng et al., 2021 and detection Carion et al., 2020; Su et al., 2023; Tan et al., 2023).

In this paper, we take an early attempt to adopt transformer to perform low-light video enhancement (LLVE). However, directly applying the vanilla transformer architecture will not bring us performance gains since it ignores the local feature alignment. To be specific, the feed forward self-attention layer that lacks inductive bias (Goyal and Bengio, 2022) in the original transformer merely focuses on modeling the global semantic information among the divided tokens, it fails to pay more attention to the local regions like edges and shapes within the objects, which are the crucial cues in low-level vision tasks. The misalignment of contextual local pixels features may lead to misjudgments in the upsampling predictions, especially on the object boundaries. As depicted in Fig. 1, while this can be alleviated by using optical flows (Ilg et al., 2017), it will increase computational burden and hinder the application in real-world scenarios. Another alternative is to adopt local window like in Swin-T (Liu et al., 2021a). The fixed local windows, however, fail to align pixels from different regions across long distances.

To this end, we propose an **A**daptive **L**ocally-**A**ligned **T**ransformer (ALAT) network to tackle the aforementioned drawbacks. Specifically,

* Corresponding author at: School of Software Engineering, South China University of Technology, Guangzhou, China.
E-mail address: qyw@scut.edu.cn (Q. Wu).

https://doi.org/10.1016/j.cviu.2023.103916
Received 28 August 2023; Received in revised form 2 November 2023; Accepted 26 December 2023
Available online 29 December 2023

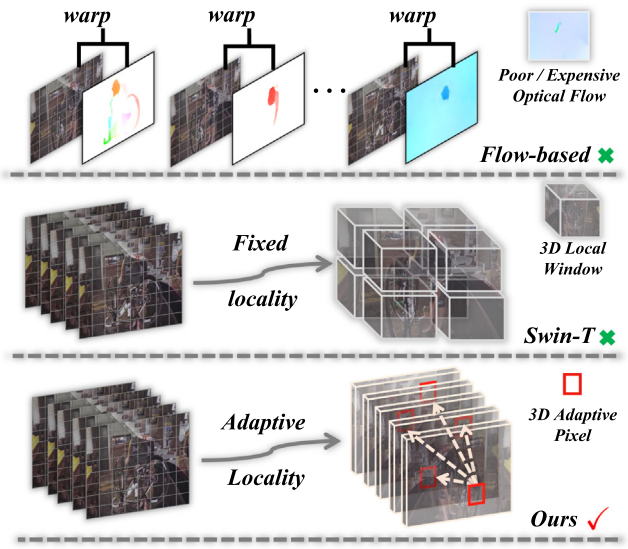

Fig. 1. Illustration of the difference among flow-based, Swin-transformer approaches and our method. The proposed operation can adaptively align the potential semantic pixels so as to preserve the video spatial-temporal consistency.

it consists of three parts including: feature encoder, locally-aligned transformer block (LATB) and pyramid feature decoder. The encoder–decoder structure in our network is similar to the feature pyramid network (Lin et al., 2017), and we utilize the convolution layers to extract features to preserve low-frequency information. The core designed LATB aims to adaptively align the current-frame video pixel with its top-k most semantically relevant features from other frames so as to learn the spatial-temporal transformation offsets. Besides, we further extend our LATB to capture multi-scale local patterns by setting scalable relevant pixel density. Features at different scales are aggregated to enhance the feature representations. This can help us improve the network robustness in a simple but effective way. The proposed strong network can produce more satisfactory results over other state-of-the-arts.

Furthermore, as is not well-explored in low-light video enhancement, some standard evaluation benchmarks are missing. For example, most of the existing works (Zhang et al., 2021b; Lv et al., 2018) use gamma correction (Bradski, 2000) to synthesize the pseudo low-light video and then train/test on these datasets, this may limit their applications in real scenes since the synthetic videos have a large gap from the real-world low-light videos. Although recent researches (Chen et al., 2019; Jiang and Zheng, 2019a; Wang et al., 2021) propose some LLVE datasets, either they have not been released publicly yet for commercial reasons, or some are static videos (*e.g.*, there are very few moving objects in the videos). Therefore, we release a new dataset, called *ExpressWay*, containing high-quality videos captured in different express highways with moving objects (*i.e.*, different vehicles) in both dark- and bright-light conditions from different surveillance scenarios. *ExpressWay* makes a leap in terms of diversity and difficulty, and it is more close to the natural real-world conditions rather than adding a filter on the camera lens to capture the low-light videos. Our contribution can be summarized as follows:

- We experimentally reveal the shortcomings in vanilla transformer and explore the transformer-variant architecture in LLVE domain. The designed LATB not only models the long-range spatial-temporal dependencies by exploiting the global self-attention mechanism, but also preserves the local patterns by aligning the semantically nearby pixels.
- We propose a challenging real-world *ExpressWay* benchmark, which is captured from both natural dark- and bright-scenes, to facilitate future research in LLVE fields.
- Extensive experiments conducted on five benchmarks under different settings including synthesized, static, and our proposed dynamic *ExpressWay* datasets validate the effectiveness and superiority of the proposed Adaptive Locally-Aligned Transformer network.

2. Related work

Low-Light Image/Video Enhancement. Low-Light Image Enhancement (LLIE) has made remarkable progress over the past decades. Most of the early conventional works (Fu et al., 2016; Guo et al., 2016; Li et al., 2018b; Ren et al., 2020) are based on Retinex-Theory (Land, 1977), which decomposes low-light images into reflection and illumination components via priors or regularization. Later, benefiting from the deep-learning models, more researches pay attention to design data-driven methods in various ways. LLNet (Lore et al., 2017) introduces an autoencoder to simultaneously brighten and denoise the low-light images. Successively, DeepLPF (Moran et al., 2020) proposes different learnable spatially local filters to enhance low-light images in an end-to-end way. DeepUPE (Wang et al., 2019b) explores both the global and local features to learn the mapping relationships between images and illuminance. Retinex-Net (Wei et al., 2018) then divides the network into decomposition, adjustment and reconstruction model, to ultimately restore the normal-light images. LightenNet (Li et al., 2018a) later presents a light-weight convolutional network to reduce the computational burden. Wang et al. (2023) further improved the dark area enhancement effect by introducing deep separable convolution based on the Retinex theory. More recently, some unsupervised methods like Enlighten-GAN (Jiang et al., 2021) try to adopt Generative Adversarial Network (GAN) (Goodfellow et al., 2014) to train from unpaired images and Zero-DCE (Guo et al., 2020) formulates light enhancement as a task of image-specific curve estimation with a deep network, respectively. In terms of Low-Light Video Enhancement (LLVE) fields, such kinds of the LLIE approaches (Wei et al., 2018; Jiang et al., 2021; Guo et al., 2020; Zhang et al., 2019) can be readily applied to LLVE tasks. However, they fail to model the spatial-temporal information lying in video, which cannot yield satisfactory performance. MBLLEN (Lv et al., 2018) uses multiple sub-networks of CNNs, and finally the output enhanced image/video by multi-branch fusion. Zhang et al. (2021b) proposes to use optical flow information to learn the temporal consistency in video from single image. SMOID (Jiang and Zheng, 2019b) utilizes a U-Net architecture with fully convolutional network operations to perform video enhancement. Recent, LLVE-SEG (Liu et al., 2023) proposes an event-based workflow, but requires multi-step training and relies more on supervised signals. Nevertheless, all these methods fail to capture the long-range dependencies among different video frames and remain some limitations.

Feature Alignment. In order to preserve the spatial details (*i.e.*, edges, shapes) in low-level vision tasks, feature alignment cannot be ignored. Generally, it can be divided into two strategies: flow-based warping and deformable convolutional. Chan et al. (2021b) and Liang et al. (2022) both use the bidirectional flow guidance generated by Ranjan and Black (2017) to align the features in video super-resolution. Some other video inpainting (Zou et al., 2021; Li et al., 2022) works also adopt flow-guided measure to find correspondences from neighboring frames. However, we argue that obtaining such flow cues is expensive and sometimes they are poor in low-light situations, which may harm the network performance. An alternative is to learn the offsets in DCN (Dai et al., 2017). FaPN (Huang et al., 2021a) and AlignSeg (Huang et al., 2021b) both perform alignment using deformable convolution and achieve performance gains in several image segmentation backbones. In contrast, our proposed transformer block is more

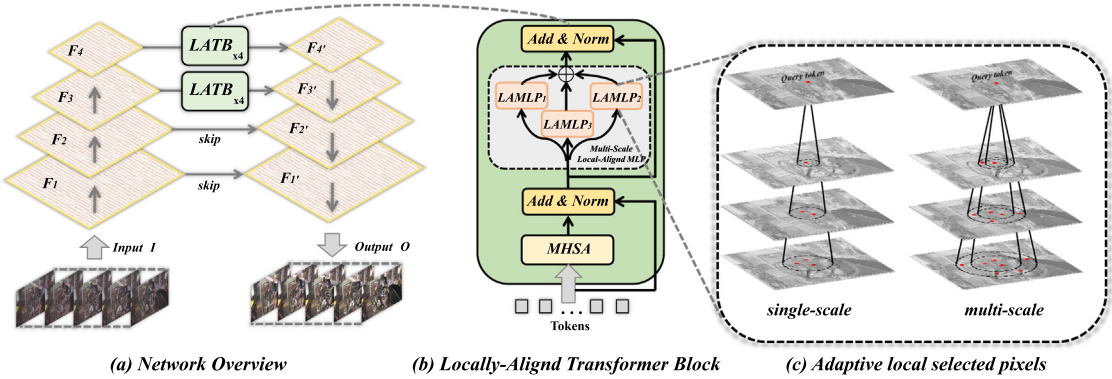

Fig. 2. (a) The overview of our Adaptive Locally-Aligned Transformer network (ALAT). (b) The structural illustration of the proposed locally-aligned transformer block introduced for capturing video long-range spatial-temporal information and for aligning the local contextual details. (c) Single-scale adaptive local selected pixels and Multi-scale adaptive local regions. Note that this is just a visualization and not necessarily the exact set of matches.

suitable for video since the DCN only aligns the local patterns within single images, while ours can model adaptive local spatial information across different frames.

Low-Light Video Dataset. Collecting low-light image and video dataset (Wei et al., 2018; Loh and Chan, 2019) is hard, and thus, previous LLVE methods (Zhang et al., 2021b, 2020) evaluate on the synthetic (*e.g.*, darken the frames by using gamma correction (Bradski, 2000)) low-light videos like DAVIS (Pont-Tuset et al., 2017), YoutubeVOS (Xu et al., 2018) or self-collected videos, *etc*. Later, SMID (Chen et al., 2019) and SMOID (Jiang and Zheng, 2019b) propose the LLVE datasets. However, the former is a static video dataset while the latter is not released to use. Recently, SDSD (Wang et al., 2021) conducts a new dataset by using a mechatronic system to capture pair of low- and normal-light dynamic videos. However, there is no moving foreground object in all these videos and the low-light videos are captured through Neutral-Density filter, whose light distribution is still different from the natural low-light real scenes. In contrast, our *ExpressWay* is much closer to the real-world light distribution since it is collected by surveillance cameras at midnight and during the day. Besides, it is more challenging, which contains different foreground moving objects in dynamic scenes.

3. Methodology

3.1. Network overview

Given T frames of the low-light video sequences, our target is to enhance them and make their light distribution closed to the normal-light videos. To achieve this goal, we propose a novel **A**daptive **L**ocally-**A**ligned **T**ransformer (ALAT) network, whose overall architecture is shown in Fig. 2(a). Concretely, the input $\mathcal{I} \in \mathbb{R}^{B*T\times 3\times H\times W}$ is fed into an encoder–decoder like network, where H and W denote the spatial size of the feature map, B and T denote the batch size and frame number, respectively. Inspired by (Su et al., 2023), the encoder uses several convolution layers to hierarchically extract shallow features of the input video frames so as to preserve low-frequency information. Afterward, these different layers' low-level features are transmitted to the proposed locally-aligned transformer block (LATB), as shown in Fig. 2(b). Specifically, LATB is responsible for modeling the global long-range dependencies among different video frames and aligning their local patterns. We reformulate the original transformer and adopt the locally-aligned MLP operation to overcome the transformation offsets (Dosovitskiy et al., 2020). We use 4 locally-aligned transformer blocks after the shallow CNN features at different levels, which can produce the updated global-local features. And the decoder is a pyramid-like structure that is fused by upsampling with low-level features in skip connection. Finally, we can obtain the enhanced output of video frames $\mathcal{O} \in \mathbb{R}^{B*T\times 3\times H\times W}$.

3.2. Adaptive locally-aligned transformer

Preliminaries. In vision transformer, the given input is split into a sequence of flattened 2D patch tokens $T_P \in \mathbb{R}^{N\times D}$, where D is the embedding dimension and N is the tokens number. Each transformer block consists of a Multi-Head Self-Attention (MHSA) module and a feed forward Multi-layer Perceptron (MLP) module in residual form. LayerNorm (Ba et al., 2016) layers (LN) are operated on the MHSA and the MLP, respectively. As mentioned before, the feed forward MLP ignores the locality that is vital in low-level tasks. To this end, we adopt the alignment MLP transformer with a reformulation to overcome this drawback.

Approach. Specifically, the feature $F \in \mathbb{R}^{B*T\times D\times h\times w}$ from the encoder is then reshaped into the flattened tokens $X \in \mathbb{R}^{B\times D\times N}$, where $N = T \times h \times w$ and D is the number of feature channels. To leverage the self-attention layer to capture the long-range dependencies between tokens, we first obtain the query, key and value matrices, which are computed as follows:

$$Q = \mathbf{W}_Q X + \mathbf{B}_Q, \quad K = \mathbf{W}_K X + \mathbf{B}_K, \quad V = \mathbf{W}_V X + \mathbf{B}_V, \tag{1}$$

where $\mathbf{W}_Q, \mathbf{W}_K$ and $\mathbf{W}_V$ are three learnable linear weight matrices, while $\mathbf{B}_Q, \mathbf{B}_K$ and $\mathbf{B}_V$ are weight vectors. After having the Q, K, V, the global self-attention mechanism can be formulated as:

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{D})V, \tag{2}$$

where we can perform the attention function parallelly and aggregate the results for multi-head self-attention (MHSA).

Subsequently, we adopt our proposed adaptive locally-aligned MLP, named **LAMLP**, after the multi-head self-attention function, whose illustrative diagram is shown in Fig. 2(c) left. We view the current i-th frame as query set, and the left (T-1) frames as support set. For each video pixel in the query set, it will match with its most relevant pixels from the support set. This step enables the query pixel to learn the offsets adaptively by aligning the semantically local features from the different frames. The reason why we do not match the pixels in the same frame are two folds: (*i*) the encoder has already captured the regional information due to the 2D local convolutional operation. (*ii*) the current pixel matching with too many other pixels in the same frame will lack collaborative temporal communication across different frames. And we consider that learning the spatial-temporal consistency from different frames is important for video tasks. Mathematically, we can achieve this feature alignment by first constructing an affinity matrix $\mathcal{A}$ that represents the relationships between tokens. We here use ℓ_2-metric to measure the distance between the two arbitrary tokens: $\mathcal{D}_{ij} = \|x_i - x_j\|^2$. Feature channel normalization is used, which guarantees $\|x_i\|^2 = 1$. Thus, by removing the constant, we can formulate the affinity matrix $\mathcal{A}$ as:

$$\mathcal{A} = X'^T X', \quad \mathcal{A} \in \mathbb{R}^{B\times N\times N}, \tag{3}$$

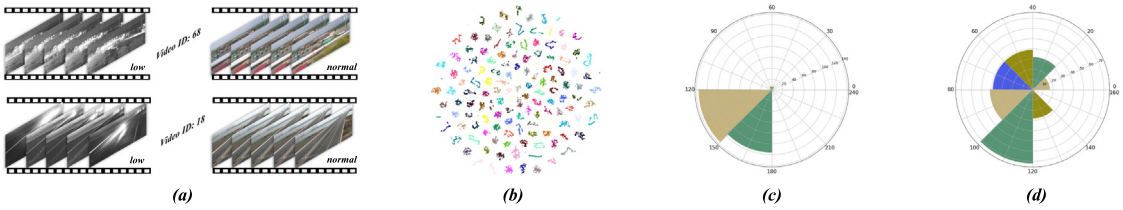

Fig. 3. (a) Some pair video samples from our proposed dataset (zoom in for more details). (b) Video t-SNE (Van der Maaten and Hinton, 2008) embedding distribution. The diversity is helpful in evaluating the ability of the model. (c) Video frame length distribution. The majority of the videos are in the range of 120 ∼ 180 s (c) Video frame brightness distribution. The different light intensities can help to assess the robustness of the algorithms.

where X' is the output from MHSA module. The T diagonal block matrices of shape $hw \times hw$ are set to $-INF$ to avoid the pixels matching others in the same frame. Then, we execute the following step to select and group the regional pixels:

$$\hat{X}_i = X' \circ Rank_{(k_i)}(\mathcal{A}), \quad \hat{X} \in \mathbb{R}^{B\times D\times N\times k_i}, \tag{4}$$

where $Rank(\cdot)$ is a sorting function that operates on the affinity matrix and yields the top-k semantically correspondence local region's index matrix in the shape of $B \times N \times k$ via the distance metric. $\circ$ is the combining function, denotes the retrieval of k feature vectors for each pixel from X' based on the index matrix, followed by aggregation and stacking. The output size is $B{\times}D{\times}N{\times}k$. This means the query pixel with its k-corresponding ones are grouped into the neighborhood region. Compared to other alignment approaches mentioned in Sec 2, our core advantage is to adaptively select the potential pixels of the current query, which avoids redundant and uncorrelated information between the consecutive frames. The subscript i denotes the cardinality of k is scaleable for multi-scale operation, which is shown as in Fig. 2 right. When k becomes larger, it contains more relevant tokens from other frames. In order to align the regional features, the local Multi-layer Perceptron embedding operation is performed as:

$$X = \sum_i \mathrm{Max}(\underset{x\in\hat{X}_i}{\varphi}(x_1, x_2, \ldots, x_{k_i})), \quad X \in \mathbb{R}^{B\times D\times N}, \tag{5}$$

where $\varphi(\cdot)$ is the local feature modeling function and we here use two 1×1 convolution layers followed by ReLU activation function, and we fuse the multi-scale local alignment features by tensor addition. Ultimately, the whole process of one of our Adaptive Locally-Aligned Transformer block can be formulated as:

$$\begin{aligned} X' &= \mathrm{MHSA}(\mathrm{LN}(X)) + X, \\ X &= \mathrm{LAMLP}(\mathrm{LN}(X')) + X', \end{aligned} \tag{6}$$

where four of these identical blocks are stacked end-to-end to model the video information. Then the updated features are fed into the decoder to reconstruct the high-quality video enhancement results.

4. The proposed dataset

Construction guidelines. We collect the videos using three types of cameras including SANTACHI ST-NT89R-EL-AAL, Hikvision iDS-2DF7C445IXR, and DAHUATECH DH-SD-6A9240. We set these pylon cameras in different places to capture real-world scenarios like cityscapes, highways, and other natural scenes, *etc*. These surveillance videos can verify the practicability of the algorithms, such as roadway low-light enhancement to promote vehicle tracking and object detection issues. Specifically, we treat the videos captured during midnight as low-light conditions while those captured in the daytime as normal-light conditions. Some of the video samples are shown in Fig. 3(a). Dataset is available at https://github.com/y1wencao/LLVE-ALAT. More visualization examples can be referred to the supplementary materials.

Dataset statistics. Our dataset consists of 134 paired videos covering diverse scenes, whose overall data feature distribution can be seen in Fig. 3(b). Among them, 109 are for training and 25 are for testing. Moreover, the frame length in each video varies from a minimum of 14 s to a maximum of 215 s, and the frame brightness (Bezryadin et al., 2007) is ranged from 8 to 116, whose statistics are depicted in Fig. 3(c) and Fig. 3(d) respectively.

Usages. Note that it is infeasible to capture the low-light input with its corresponding normal-light ground-truth for dynamic scenes as was done for single-image (Chen et al., 2019). Therefore, our dataset can be for testing only or training in an unsupervised way. (**i**) *Test Only*: We train the model on the non-object moving video dataset (Wang et al., 2021) by adopting $\mathcal{L}_{sup}$ loss (Eq. (7)) in a supervised way, and then use the pre-trained weight testing on our dataset. This can help to verify the generalization performance of the models. (**ii**) *Unsupervised Training*: We adopt the same $\mathcal{L}_{unsup}$ loss (Eq. (7)) as in Jiang et al. (2021) (*i.e.*, the global (G) and local (L) content loss Johnson et al., 2016 and adversarial loss Mao et al., 2017) to train our network on the train set of our dataset, and test on the test set.

$$\begin{aligned} \mathcal{L}_{sup} &= \|\hat{Y} - Y\|_1 + \mathcal{L}_{adv}(\hat{Y}, Y), \\ \mathcal{L}_{unsup} &= \mathcal{L}^G_{cont}(\hat{Y}, I_{low}) + \mathcal{L}^L_{cont}(\hat{Y}, I_{low}) \\ &\quad + \mathcal{L}^G_{adv}(\hat{Y}, I_{normal}) + \mathcal{L}^L_{adv}(\hat{Y}, I_{normal}), \end{aligned} \tag{7}$$

where $\hat{Y}$ and Y denote the prediction and ground-truth, respectively. I_{low} and I_{normal} denote the low-light input and normal-light input.

5. Experiments

Datasets and Metrics. To validate our network, we conduct experiments on 5 benchmarks including 2 synthesized low-light video datasets on DAVIS (Pont-Tuset et al., 2017) and YouTube (Xu et al., 2018), 2 non-object moving low-light video datasets on SDSD (Wang et al., 2021) and SMID (Chen et al., 2019), and our proposed dynamic *ExpressWay* datasets. For synthesized datasets, following (Zhang et al., 2021b), we darken the video frame using gamma correction as: $x = \beta \times (\alpha + y)^{\gamma}$, where $\gamma \sim U(2, 3.5)$, $\alpha \sim U(0.9, 1)$ and $\beta \sim U(0.5, 1)$. We strictly follow the previous work (Zhang et al., 2021b; Wang et al., 2021; Chen et al., 2019) to train our network on the train set and test on the split test set on all the benchmarks for fair comparisons. We adopt two widely-used metrics of peak signal-to-noise ratio (*PSNR*) and structural similarity (*SSIM*) as other similar works (Wang et al., 2021; Chan et al., 2021b).

Implementation Details. We use Adam (Kingma and Ba, 2014) as the optimizer with $\beta_1 - 0.9$ and $\beta_2 - 0.999$. We train our network for 500k iterations. During each iteration, 5 random frames from one video are sampled with mini-batch size of 8. Random flipping is leveraged for data augmentation. The learning rate is initially set to 1e-2 and is reduced at 400 k and 450 k iterations respectively. We use 4 adoptive locally-aligned transformer blocks with 4 multi-heads and the k number is set to 8 by default. We employ VGG (Simonyan and Zisserman, 2014) as our shallow CNN encoder backbone, where the $MaxPool3\_1$ and 4_1 (*i.e.*, F_3 and F_4) layers are fed into the LATB, respectively. We train our network in an end-to-end manner, and more training details can be referred to the supplementary material.

Table 1
Exploration of different components in our proposed network on SDSD (Wang et al., 2021) test set.

(a) Different local alignments.

Method	*PSNR*	*SSIM*
Vanilla MLP (Dosovitskiy et al., 2020)	22.02	0.855
Flow-Based (Chan et al., 2021b)	23.61	0.861
Swin-T (Liu et al., 2021a)	23.29	0.848
2D-DCN (Dai et al., 2017)	22.92	0.851
3D-DCN (Ying et al., 2020)	21.03	0.787
LAMLP (*Ours*)	**24.15**	**0.874**

(b) Top-k number

k	*PSNR*	*SSIM*	GFLOPs
16	24.23	0.870	771.5
14	24.12	0.869	766.4
12	24.09	0.874	761.2
8	**24.15**	**0.874**	**750.9**
6	23.97	0.866	745.8
4	23.84	0.854	740.7

(c) Block and head numbers

LATB		*PSNR*	*SSIM*
B = 2	H = 2	23.88	0.848
	H = 4	24.02	0.865
B = 4	H = 2	23.97	0.870
	H = 4	**24.15**	**0.874**
B = 6	H = 2	24.07	0.870
	H = 4	24.15	0.873

(d) Multi-scale k local regions

k = 8	k = 6	k = 4	*PSNR*	*SSIM*
✔			24.15	0.874
✔	✔		25.11	0.882
✔	✔	✔	**25.46**	**0.888**

(e) LATB in different CNN shallow layers

Baseline	LATB$_{_F3}$	LATB$_{_F4}$	*PSNR*	*SSIM*
✔			18.67	0.626
✔	✔		23.86	0.849
✔	✔	✔	**25.46**	**0.888**

(f) Distance metrics

Distance metrics	*PSNR*	*SSIM*
ℓ_1-distance	25.22	0.87
EMD (Rubner et al., 2000)	25.13	0.86
ℓ_2-distance	**25.46**	**0.89**

Table 2
Comparisons with SOTA models on DAVIS (Pont-Tuset et al., 2017) and YouTube (Xu et al., 2018) datasets.

Method	DAVIS		YouTube	
	PSNR	*SSIM*	*PSNR*	*SSIM*
LIME (Guo et al., 2016)$_{\text{TIP'17}}$	16.83	0.457	16.33	0.541
RetinexNet (Wei et al., 2018)$_{\text{BMVC'18}}$	19.56	0.748	18.76	0.773
BLIND (Lai et al., 2018)$_{\text{ECCV'18}}$	22.94	0.917	21.37	0.884
SID (Chen et al., 2018a)$_{\text{CVPR'18}}$	22.93	0.925	20.10	0.903
MBLLEN (Lv et al., 2018)$_{\text{BMVC'18}}$	18.38	0.798	19.97	0.836
SMOID (Jiang and Zheng, 2019a)$_{\text{ICCV'19}}$	23.42	0.921	22.52	0.917
DVP (Lei et al., 2020)$_{\text{NIPS'20}}$	22.98	0.921	21.49	0.926
BasicVSR (Chan et al., 2021a)$_{\text{CVPR'21}}$	24.17	0.935	25.33	0.930
Zhang et al. (Zhang et al., 2021b)$_{\text{CVPR'21}}$	24.01	0.931	24.68	0.908
BasicVSR++ (Chan et al., 2021b)$_{\text{CVPR'22}}$	24.95	0.942	27.28	0.944
ALAT (*Ours*)	**25.75**	**0.954**	**28.31**	**0.948**

Table 3
Comparisons with SOTA models on SDSD (Wang et al., 2021) and SMID (Chen et al., 2019) datasets.

Method	SDSD		SMID	
	PSNR	*SSIM*	*PSNR*	*SSIM*
MBLLEN (Lv et al., 2018)$_{\text{BMVC'18}}$	21.79	0.65	22.67	0.68
DeepUPE (Wang et al., 2019b)$_{\text{CVPR'19}}$	21.82	0.68	23.91	0.69
SMID (Chen et al., 2019)$_{\text{ICCV'19}}$	24.09	0.69	24.78	0.72
DRBN (Yang et al., 2020)$_{\text{CVPR'20}}$	22.31	0.65	24.42	0.69
ZeroDCE (Guo et al., 2020)$_{\text{CVPR'20}}$	20.06	0.61	22.62	0.67
DVP (Lei et al., 2020)$_{\text{NIPS'20}}$	20.84	0.68	22.39	0.66
BasicVSR (Chan et al., 2021a)$_{\text{CVPR'21}}$	23.21	0.72	24.62	0.71
SDSD (Wang et al., 2021)$_{\text{ICCV'21}}$	24.92	0.73	26.03	0.75
BasicVSR++ (Chan et al., 2021b)$_{\text{CVPR'22}}$	24.07	0.76	25.94	0.73
LLVE-SEG (Liu et al., 2023)$_{\text{AAAI'23}}$	**25.81**	0.80	–	–
ALAT (*Ours*)	25.46	**0.89**	**26.86**	**0.79**

5.1. Ablation analysis

In this section, we will analyze each proposed component of our proposed model and explore their impact on the SDSD dataset (Wang et al., 2021) in Table 1.

(i) Different local alignments: Table 1(a) shows that by using the original MLP in VIT (Dosovitskiy et al., 2020), we fail to obtain satisfactory performance. When we replace the vanilla MLP by using a flow-based alignment method as in Chan et al. (2021b), it can boost the performance to some degree but it increases the number of parameters by nearly 1.5 times. Moreover, by using the latest proposed local shift window mechanism in Swin-T (Liu et al., 2021a) also cannot achieve the best metrics (−0.86% worse than our LAMLP in *PSNR*). Besides, 2D (Dai et al., 2017) and 3D (Ying et al., 2020) deformable convolutional operations are also experimentally adopted in our transformer block. However, they both cannot bring us significant gains or even get worse. By contrast, our LAMLP is able to yield the best performance, which demonstrates its necessity and effectiveness. Furthermore, Fig. 4 visualizes the middle upsampling features and the optical flow output from LATB $w/$ or w/o alignment. It reveals that the aligned features are able to preserve more detailed information like boundaries and persevere the cross-frame motion consistency.

(ii) Different k number. As shown in Table 1(b), larger k indicates involving more semantically local tokens, which may contain more useful spatial-temporal cues. Here we choose $k-8$ since it can successfully achieve competitive performance with moderate GFLOPs.

(iii) Block and head numbers in LATB. Likewise, as seen in Table 1(c), we set block and head numbers both to 4. The small numbers of blocks and heads may fail to sufficiently capture useful information while the too large numbers may learn some redundant information that harms the network.

(iv) Multi-scale k local regions. As can be seen in Table 1(d), by aggregating different large and small scales of local information, it can make our network more robust and yield more satisfactory results.

(v) LATB in different CNN shallow layers. Table 1(d) analyzes the effect of LATB in hierarchical encoder layers. It shows that without using LATB, the *baseline* model can merely achieve low $PSNR$:18.67. This is because the conventional CNNs cannot well model the spatial-temporal information in videos. By applying our proposed LATB, we can boost the network significantly. In this paper, we adopt LATB following behind both F_3 and F_4 features from the encoder layers to pursue higher accuracy.

(vi) Distance metrics. We tried different distance metrics, including ℓ_1-distance and EMD (Rubner et al., 2000) distance. As shown in Table 1(f), we found that the differences in performance using various distance functions were not significant. In the end, we chose the ℓ_2-distance, which performed better.

5.2. Compare with the state-of-the-art methods

As shown in Table 2, we first report quantitative results on the synthesized DAVIS (Pont-Tuset et al., 2017) and YouTube (Xu et al., 2018) low-light video datasets. Compared to the image-based methods (*i.e.*, LIME (Guo et al., 2016), RetinexNet (Wei et al., 2018)), other video-based approaches have obvious advantages and can make significant progress. As for the methods (*i.e.*, Zhang *et al.* (Zhang et al., 2021b) and BasicVSR++ (Chan et al., 2021b)) that employ optical flow guidance, they can achieve better performance. By contrast, our ALAT can produce the top results on both the large datasets without using additional motion-guided cues. Likewise, as shown in Table 3 in terms of two self-captured low-light video SDSD (Wang et al., 2021) and SMID (Chen et al., 2019) datasets, our ALAT again outperforms other either image- or video-based methods by a large margin (*SSIM:* +**0.09**

Table 4
Comparisons with SOTA models on *ExpressWay* dataset with user study (US)↑/NIQE (Mittal et al., 2012)↓ scores. † indicates our re-implement results since the authors do not provide codes.

Method	CycleGAN (Zhu et al., 2017)$_{ICCV'17}$	Zero-DCE (Guo et al., 2020)$_{CVPR'20}$	EnlightenGAN (Jiang et al., 2021)$_{TIP'21}$	RAUS (Liu et al., 2021b)$_{CVPR'21}$	BasicVSR++ (Chan et al., 2021b)$_{CVPR'22}$	LLVE-SEG† (Liu et al., 2023)$_{AAAI'23}$	***Ours***
Test only	5.6/7.83	7.5/7.52	19.5/6.92	20.6/6.51	15.6/6.88	–/–	**31.2/6.44**
Unsup. train	4.5/7.54	9.4/7.48	8.2/7.91	12.5/6.73	12.4/6.63	14.4/7.24	**38.6/6.13**

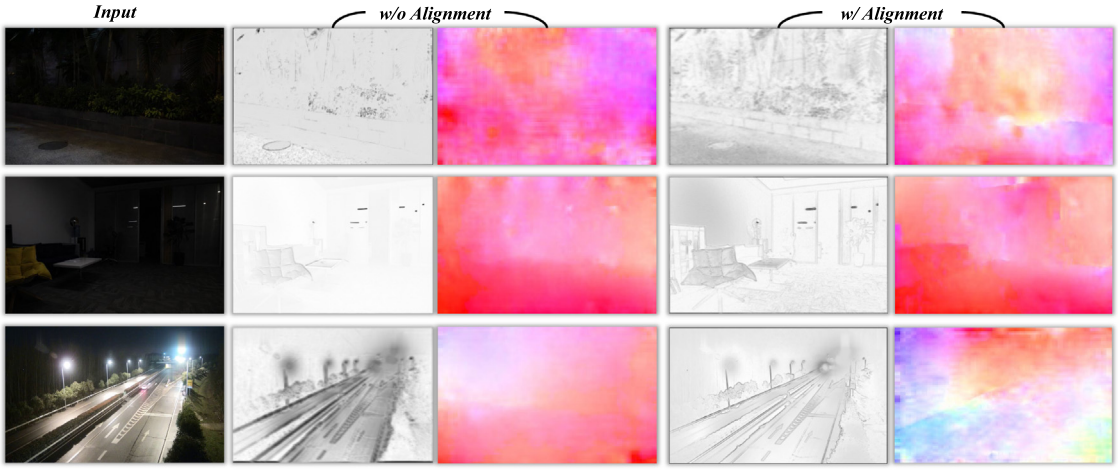

Fig. 4. Visualization of the features from the middle layer and the output flow field by using Raft (Teed and Deng, 2020). Compared with the results without adaptive locally-aligned operation, our patterns are more detailed the flows are much smaller and hierarchical, indicating that the proposed alignment module can be aware of the local feature and well preserve the motion spatial-temporal consistency.

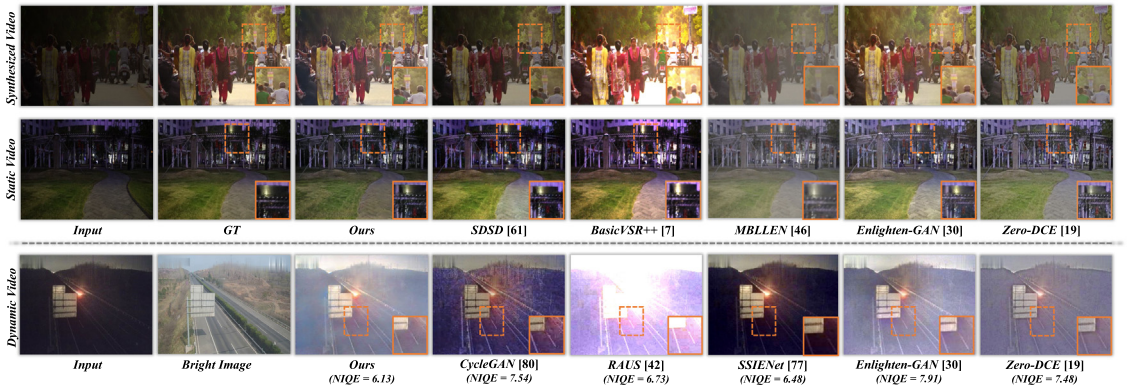

Fig. 5. Visual comparison. Our proposed ALAT can perform high-quality and spatial-temporal video enhancement. The first and second rows are derived from the synthesized (*i.e.*, DAVIS (Pont-Tuset et al., 2017)) and static non-object videos (Wang et al., 2021) datasets, and thus they have ground-truths. The third row is derived from our proposed dynamic real-world *ExpressWay* dataset, where we can yield smoother and brighter outputs and achieve lower NIQE (Mittal et al., 2012). Best viewed by zooming.

on SDSD; *PSNR*: +**0.83**, *SSIM:* +**0.04** on SMID compared to the second best approach). On the SDSD dataset, our method exhibits a slightly lower PSNR but a substantially higher SSIM. It is worth noting that SSIM is a more robust metric for assessing image quality (Wang and Bovik, 2009), as it better corresponds to human perception, giving our method a competitive edge.

In order to validate the robustness of the algorithms in dynamic real-world scenarios, we further compare with the SOTA methods on *ExpressWay* dataset. For fair comparisons, under *Test Only* setting, we train all the algorithms on SDSD (Wang et al., 2021). We use NIQE (Mittal et al., 2012) metric to evaluate the non-reference image quality and perform a user study to quantify the subjective visual quality (Guo et al., 2020; Jiang et al., 2021) of different methods. As can be seen in Table 4, when directly testing on it, ALAT can obtain lower NIQE and the participants prefer our predicted videos. This reflects that our network is not over-fitting on the former datasets, which can learn useful potential information. When training on *ExpressWay* by using the unsupervised loss as in Jiang et al. (2021), we can further produce high-quality results and lower NIQE (**-0.31**) compared to the model trained under the *Test Only* setting. It is notable that some of the unsupervised LLIE methods (*i.e.*, RAUS (Liu et al., 2021b) and Enlighten (Jiang et al., 2021)) get worse performance, we conjecture that they fail to model the multi-frame motion information in a more challenging dynamic dataset. Compared to the LLVE method (*i.e.*, BasicVSR++ (Chan et al., 2021b) and LLVE-SEG (Liu et al., 2023)), ALAT does not need to consider the flow factor, and thus, it is more flexible for both supervised/unsupervised training and it can yield the better performance. Note that we use the same $\mathcal{L}_{unsup}$ in Eq. (7) for BasicVSR++ (Chan et al., 2021b) and LLVE-SEG (Liu et al., 2023) training. Generally, all the above experiments well reflect the superiority and effectiveness of our network and Fig. 5 show some qualitative results. More qualitative visualizations can be referred to the supplementary material.

Table 5
Comparisons with SOTA models on other video enhancement tasks.

(a) Derain on RainSynAll100 (Yang et al., 2021).

Method	*PSNR*	*SSIM*
SE (Wei et al., 2017)$_{ICCV'17}$	15.29	0.505
SpacCNN (Chen et al., 2018b)$_{ICCV'18}$	18.39	0.646
FastDerain Jiang et al. (2018)$_{TIP'18}$	17.09	0.582
J4RNet-P (Liu et al., 2018b)$_{CVPR'18}$	19.26	0.624
FCRVD (Yang et al., 2019)$_{CVPR'19}$	21.06	0.741
RMFD (Yang et al., 2021)$_{TPAMI'21}$	25.14	0.917
BasicVSR++ (Chan et al., 2021b)$_{CVPR'22}$	27.67	0.913
NCBF (Huang et al., 2022)$_{CVPR'22}$	28.11	0.923
ALAT (*Ours*)	**29.22**	**0.935**

(b) Dehaze on REVIDE (Zhang et al., 2021a)

Method	*PSNR*	*SSIM*
VDN (Ren et al., 2018)$_{TIP'18}$	16.64	0.813
EDVR (Wang et al., 2019a)$_{CVPR'19}$	21.22	0.871
FFA (Qin et al., 2020)$_{AAAI'20}$	16.65	0.813
MSBDN (Dong et al., 2020)$_{CVPR'20}$	22.01	0.876
KDDN (Hong et al., 2020)$_{CVPR'20}$	16.32	0.773
CG-IDN (Zhang et al., 2021a)$_{CVPR'21}$	23.21	0.884
BasicVSR++ (Chan et al., 2021b)$_{CVPR'22}$	21.68	0.873
NCBF (Huang et al., 2022)$_{CVPR'22}$	23.63	**0.892**
ALAT (*Ours*)	**23.90**	0.857

(c) Desnow on Videvo (Lai et al., 2018).

Method	*PSNR*	*SSIM*
CycleGAN (Zhu et al., 2017)$_{ICCV'17}$	16.84	0.765
Desnow (Liu et al., 2018a)$_{TIP'18}$	18.77	0.796
DAD (Zou et al., 2020)$_{CVPR'20}$	20.43	0.817
JSTAR (Chen et al., 2020)$_{ECCV'20}$	21.11	0.825
BasicVSR (Chan et al., 2021a)$_{CVPR'21}$	25.48	0.898
HDCWNet (Chen et al., 2021)$_{ICCV'21}$	19.40	0.828
BasicVSR++ (Chan et al., 2021b)$_{CVPR'22}$	27.94	0.919
TSAN (Xu et al., 2022)$_{AAAI'22}$	25.89	0.886
ALAT (*Ours*)	**28.17**	**0.933**

5.3. Extensions

ALAT can be extended to many video enhancement applications. Following the similar work (Huang et al., 2022), we conduct experiments on the challenging RainSynAll100 (Yang et al., 2021) and REVIDE (Zhang et al., 2021a) for derain and dehaze, respectively. Since there is currently no dataset for video snow removal, we establish a new video desnow dataset by applying the snow masks from Chen et al. (2021) to the Videvo (Lai et al., 2018) dataset. More details on constructing video desnow dataset can be referred to the supplementary material. As shown in Table 5(a) and 5(b) (note that some of the results are derived from Huang et al. (2022)), ALAT can achieve state-of-the-art performance on derain task and get promising results on haze removal. Moreover, Table 5(c) shows that ALAT can also complete the desnow task, which outperforms the SOTA image-based desnow methods (*i.e.*, JSTAR Chen et al., 2020, HDCWNet (Chen et al., 2021)) and some concurrent video restoration works (Xu et al., 2022; Chan et al., 2021b). More visualizations can be referred to the supplementary material.

Limitations. One potential limitation is that our network may consume large GPU resources due to the matrix multiplication operation in Eq. (3). We leave this issue and try to solve it in the future work.

6. Conclusion

In this paper, we propose a strong **A**daptive **L**ocally-**A**ligned **T**ransformer (ALAT) model for low-light video enhancement, which focuses on capturing the video global spatial-temporal information while preserving the local aligned features. Besides, we establish a dynamic real-world low-light video dataset to encourage more follow-up research on this area. Extensive experiments on different benchmarks validate the superiority and effectiveness of our method, which can achieve the new state-of-the-art performance. Furthermore, our network can be extended for other video enhancement applications such as derain, dehaze and desnow, *etc*, and ALAT will serve as a solid baseline.

CRediT authorship contribution statement

Yiwen Cao: Conceptualization, Methodology, Software, Validation, Data curation, Writing – original draft, Writing – review & editing. **Yukun Su:** Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Jingliang Deng:** Conceptualization, Methodology, Software, Visualization. **Yu Zhang:** Data curation, Supervision, Project administration, Funding acquisition. **Qingyao Wu:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62272172), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515012920).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cviu.2023.103916.

References

Ba, J., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv, arXiv:1607.06450.
Bezryadin, S., Bourov, P., Ilinih, D., 2007. Brightness calculation in digital image processing. In: International Symposium on Technologies for Digital Photo Fulfillment, Vol. 2007, no. 1. Society for Imaging Science and Technology, pp. 10–15.
Bradski, G., 2000. The opencv library.. Dr. Dobb's J. Softw. Tools Prof. Program. 25 (11), 120–123.
Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2020. Nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631.
Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp. 213–229.
Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C., 2021a. BasicVSR: The search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4947–4956.
Chan, K.C., Zhou, S., Xu, X., Loy, C.C., 2021b. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. arXiv Preprint arXiv:2104.13371.
Chen, C., Chen, Q., Do, M.N., Koltun, V., 2019. Seeing motion in the dark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3185–3194.
Chen, C., Chen, Q., Xu, J., Koltun, V., 2018a. Learning to see in the dark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3291–3300.
Chen, W.-T., Fang, H.-Y., Ding, J.-J., Tsai, C.-C., Kuo, S.-Y., 2020. JSTASR: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal. In: European Conference on Computer Vision. Springer, pp. 754–770.

Chen, W.-T., Fang, H.-Y., Hsieh, C.-L., Tsai, C.-C., Chen, I., Ding, J.-J., Kuo, S.-Y., et al., 2021. ALL snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4196–4205.

Chen, J., Tan, C.-H., Hou, J., Chau, L.-P., Li, H., 2018b. Robust video content alignment and compensation for rain removal in a cnn framework. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6286–6295.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks. In: 2017 IEEE International Conference on Computer Vision. ICCV, pp. 764–773. http://dx.doi.org/10.1109/ICCV.2017.89.

Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., Yang, M.-H., 2020. Multi-scale boosted dehazing network with dense feature fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2157–2167.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Fu, X., Zeng, D., Huang, Y., Zhang, X.-P., Ding, X., 2016. A weighted variational model for simultaneous reflectance and illumination estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2782–2790.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27.

Goyal, A., Bengio, Y., 2022. Inductive biases for deep learning of higher-level cognition. Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 478 (2266), 20210068.

Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R., 2020. Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1780–1789.

Guo, X., Li, Y., Ling, H., 2016. LIME: Low-light image enhancement via illumination map estimation. IEEE Trans. Image Process. 26 (2), 982–993.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Hong, M., Xie, Y., Li, C., Qu, Y., 2020. Distilling image dehazing with heterogeneous task imitation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3462–3471.

Huang, C., Li, J., Li, B., Liu, D., Lu, Y., 2022. Neural compression-based feature learning for video restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5872–5881.

Huang, S., Lu, Z., Cheng, R., He, C., 2021a. FaPN: Feature-aligned pyramid network for dense image prediction. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 844–853. http://dx.doi.org/10.1109/ICCV48922.2021.00090.

Huang, Z., Wei, Y., Wang, X., Shi, H., Liu, W., Huang, T.S., 2021b. Alignseg: Feature-aligned segmentation networks. IEEE Trans. Pattern Anal. Mach. Intell..

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2462–2470.

Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z., 2021. Enlightengan: Deep light enhancement without paired supervision. IEEE Trans. Image Process. 30, 2340–2349.

Jiang, T.-X., Huang, T.-Z., Zhao, X.-L., Deng, L.-J., Wang, Y., 2018. Fastderain: A novel video rain streak removal method using directional gradient priors. IEEE Trans. Image Process. 28 (4), 2089–2102.

Jiang, H., Zheng, Y., 2019a. Learning to see moving objects in the dark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7324–7333.

Jiang, H., Zheng, Y., 2019b. Learning to see moving objects in the dark. In: 2019 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 7323–7332. http://dx.doi.org/10.1109/ICCV.2019.00742.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision. Springer, pp. 694–711.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Lai, W.-S., Huang, J.-B., Wang, O., Shechtman, E., Yumer, E., Yang, M.-H., 2018. Learning blind video temporal consistency. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 170–185.

Land, E.H., 1977. The retinex theory of color vision. Sci. Am. 237 (6), 108–129.

Lei, C., Xing, Y., Chen, Q., 2020. Blind video temporal consistency via deep video prior. Adv. Neural Inf. Process. Syst. 33, 1083–1093.

Li, C., Guo, J., Porikli, F., Pang, Y., 2018a. LightenNet: A convolutional neural network for weakly illuminated image enhancement. Pattern Recognit. Lett. 104, 15–22.

Li, B., Leroux, S., Simoens, P., 2021. Decoupled appearance and motion learning for efficient anomaly detection in surveillance video. Comput. Vis. Image Underst. 210, 103249.

Li, M., Liu, J., Yang, W., Sun, X., Guo, Z., 2018b. Structure-revealing low-light image enhancement via robust retinex model. IEEE Trans. Image Process. 27 (6), 2828–2841.

Li, Z., Lu, C.-Z., Qin, J., Guo, C.-L., Cheng, M.-M., 2022. Towards an end-to-end framework for flow-guided video inpainting. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR.

Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L., 2022. VRT: A video restoration transformer. arXiv preprint arXiv:2201.12288.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125.

Liu, L., An, J., Liu, J., Yuan, S., Chen, X., Zhou, W., Li, H., Wang, Y.F., Tian, Q., 2023. Low-light video enhancement with synthetic event guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 2. pp. 1692–1700.

Liu, Y.-F., Jaw, D.-W., Huang, S.-C., Hwang, J.-N., 2018a. DesnowNet: Context-aware deep network for snow removal. IEEE Trans. Image Process. 27 (6), 3064–3073.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021a. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, R., Ma, L., Zhang, J., Fan, X., Luo, Z., 2021b. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10561–10570.

Liu, J., Yang, W., Yang, S., Guo, Z., 2018b. Erase or fill? deep joint recurrent rain removal and reconstruction in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3233–3242.

Loh, Y.P., Chan, C.S., 2019. Getting to know low-light images with the exclusively dark dataset. Comput. Vis. Image Underst. 178, 30–42.

Lore, K.G., Akintayo, A., Sarkar, S., 2017. LLNet: A deep autoencoder approach to natural low-light image enhancement. Pattern Recognit. 61, 650–662.

Lv, F., Lu, F., Wu, J., Lim, C., 2018. MBLLEN: Low-light image/video enhancement using cnns.. In: BMVC, Vol. 220, No. 1. p. 4.

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2794–2802.

Mittal, A., Soundararajan, R., Bovik, A.C., 2012. Making a "completely blind" image quality analyzer. IEEE Signal Process. Lett. 20 (3), 209–212.

Moran, S., Marza, P., McDonagh, S., Parisot, S., Slabaugh, G., 2020. Deeplpf: Deep local parametric filters for image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12826–12835.

Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L., 2017. The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675.

Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H., 2020. FFA-net: Feature fusion attention network for single image dehazing. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 07. pp. 11908–11915.

Ranjan, A., Black, M.J., 2017. Optical flow estimation using a spatial pyramid network. In: IEEE Conference on Computer Vision and Pattern Recognition.

Raskar, R., Tumblin, J., 2005. Computational photography. In: ACM SIGGRAPH 2005 Courses. pp. 1–es.

Ren, X., Yang, W., Cheng, W.-H., Liu, J., 2020. LR3M: Robust low-light enhancement via low-rank regularized retinex model. IEEE Trans. Image Process. 29, 5862–5876.

Ren, W., Zhang, J., Xu, X., Ma, L., Cao, X., Meng, G., Liu, W., 2018. Deep video dehazing with semantic segmentation. IEEE Trans. Image Process. 28 (4), 1895–1908.

Rubner, Y., Tomasi, C., Guibas, L.J., 2000. The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. 40, 99–121.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Su, Y., Deng, J., Sun, R., Lin, G., Su, H., Wu, Q., 2023. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. IEEE Trans. Multimed..

Tan, T., Lim, J.M.-Y., Foo, J.J., Muniandy, R., 2023. 3D detection transformer: Set prediction of objects using point clouds. Comput. Vis. Image Underst. 103808.

Teed, Z., Deng, J., 2020. Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision. Springer, pp. 402–419.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4489–4497.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).

Vishwakarma, S., Agrawal, A., 2013. A survey on activity recognition and behavior understanding in video surveillance. Vis. Comput. 29 (10), 983–1009.

Wang, Z., Bovik, A.C., 2009. Mean squared error: Love it or leave it? A new look at signal fidelity measures. IEEE Signal Process. Mag. 26 (1), 98–117.

Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C., 2019a. Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

Wang, M., Li, J., Zhang, C., 2023. Low-light image enhancement by deep learning network for improved illumination map. Comput. Vis. Image Underst. 232, 103681.

Wang, R., Xu, X., Fu, C.-W., Lu, J., Yu, B., Jia, J., 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. pp. 9700–9709.

Wang, R., Zhang, Q., Fu, C.-W., Shen, X., Zheng, W.-S., Jia, J., 2019b. Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6849–6857.

Wei, C., Wang, W., Yang, W., Liu, J., 2018. Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560.

Wei, W., Yi, L., Xie, Q., Zhao, Q., Meng, D., Xu, Z., 2017. Should we encode rain streaks in video as deterministic or stochastic? In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2516–2525.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34.

Xu, L., He, G., Zhou, J., Lei, J., Xie, W., Li, Y., Tai, Y.-W., 2022. Transcoded video restoration by temporal spatial auxiliary network. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 3. pp. 2875–2883.

Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T., 2018. Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327.

Yang, W., Liu, J., Feng, J., 2019. Frame-consistent recurrent video deraining with dual-level flow. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1661–1670.

Yang, W., Tan, R.T., Feng, J., Wang, S., Cheng, B., Liu, J., 2021. Recurrent multi-frame deraining: Combining physics guidance and adversarial learning. IEEE Trans. Pattern Anal. Mach. Intell..

Yang, W., Wang, S., Fang, Y., Wang, Y., Liu, J., 2020. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3063–3072.

Ying, X., Wang, L., Wang, Y., Sheng, W., An, W., Guo, Y., 2020. Deformable 3d convolution for video super-resolution. IEEE Signal Process. Lett. 27, 1500–1504.

Zhang, X., Dong, H., Pan, J., Zhu, C., Tai, Y., Wang, C., Li, J., Huang, F., Wang, F., 2021a. Learning to restore hazy video: A new real-world dataset and a new method. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9239–9248.

Zhang, M., Gao, Q., Wang, J., Turbell, H., Zhao, D., Yu, J., Lu, Y., 2020. RT-venet: A convolutional network for real-time video enhancement. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4088–4097.

Zhang, F., Li, Y., You, S., Fu, Y., 2021b. Learning temporal consistency for low light video enhancement from single images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4967–4976.

Zhang, Y., Zhang, J., Guo, X., 2019. Kindling the darkness: A practical low-light image enhancer. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 1632–1640.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al., 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6881–6890.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232.

Zou, Z., Lei, S., Shi, T., Shi, Z., Ye, J., 2020. Deep adversarial decomposition: A unified framework for separating superimposed images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12806–12816.

Zou, X., Yang, L., Liu, D., Lee, Y.J., 2021. Progressive temporal feature alignment network for video inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16448–16457.