Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Neural Image Re-Exposure

Xinyu Zhang†[a], Hefei Huang†[a], Xu Jia[a,**], Dong Wang[a], Lihe Zhang[a], Bolun Zheng[b], Wei Zhou[c], Huchuan Lu[a]

[a] *Dalian University of Technology*
[b] *Hangzhou Dianzi University*
[c] *Cardiff University*

ABSTRACT

Images and videos often suffer from issues such as motion blur, video discontinuity, or rolling shutter artifacts. Prior studies typically focus on designing specific algorithms to address individual issues. In this paper, we highlight that these issues, albeit differently manifested, fundamentally stem from sub-optimal exposure processes. With this insight, we propose a paradigm termed re-exposure, which resolves the aforementioned issues by performing exposure simulation. Following this paradigm, we design a new architecture, which constructs visual content representation from images and event camera data, and performs exposure simulation in a controllable manner. Experiments demonstrate that, using only a single model, the proposed architecture can effectively address multiple visual issues, including motion blur, video discontinuity, and rolling shutter artifacts, even when these issues co-occur.

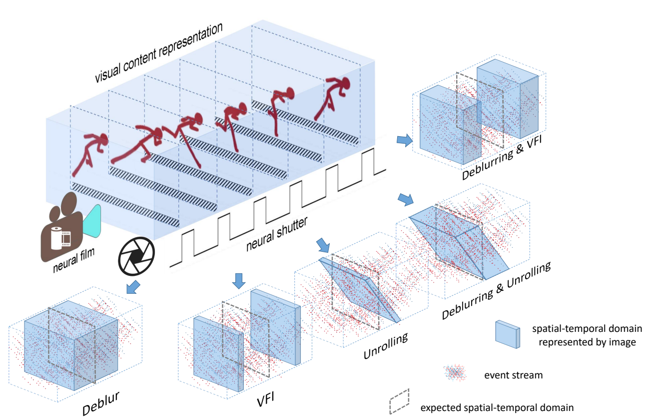

Fig. 1. The image re-exposure method adjusts the spatial-temporal domain represented by the image to an expected states, which is able to address issues including deblur, VFI, unrolling, and their combinations with a single model.

1. Introduction

An image is generated through an exposure process. The exposure process determines a portion of visual content within a spatial-temporal domain, and the image can be regarded as the representation of this portion of visual content. When the amount of visual content exceeds the image's representational capacity or the spatial-temporal domain is distorted, the quality of the image will degrade.

For instance, when the exposure period is too long or there is too much motion, the spatial-temporal domain represented by the image will contain an excessive amount of visual content, resulting in an image exhibiting noticeable blur. In the case of a rolling shutter that adopts a row-by-row readout scheme, the represented spatial-temporal domain becomes tilted, leading to the distortion commonly referred to as the "jello effect". A video is a sequence of images that carries a stream of visual content over a long period of time. If the framerate is low, there are not enough images to carry the visual content, resulting in a jerky and unstable effect. Furthermore, it is common for these issues to co-occur, producing images or videos with complex degradation.

To address these issues, methods for blur removal Zhang et al. (2022); Kupyn et al. (2018); Nah et al. (2017a), rolling shutter correction Liu et al. (2020); Zhong et al. (2021), and video frame interpolation Zhang et al. (2022); Jiang et al. (2018); Bao et al. (2019) have been explored. These methods deal with individual issues separately. When it comes to the combination of these issues, these methods are typically applied in succession.

**Corresponding author, †Equal contributions.
e-mail: xjia@dlut.edu.cn (Xu Jia)

However, since the spatial-temporal domain is determined by the exposure process, all aforementioned issues can be attributed to sub-optimal exposure. This suggests the possibility of a unified paradigm that can address all these issues. Therefore, from the perspective of exposure simulation, we propose a paradigm we refer to as *re-exposure*. This paradigm involves constructing representation of the visual content from sensor data, and simulating an optimal exposure process where the spatial-temporal domain represented by the image is in a desired state. As illustrated in Fig. 1, re-exposure is a flexible paradigm that is able to address all aforementioned problems and their combinations with a single model.

Following the proposed paradigm, we designed our method, which is called *neural image re-exposure* (NIRE), as follows.

First, we design a visual content constructor that builds a representation of visual content from images and event camera data. In this process, event cameras Lichtsteiner et al. (2008); Posch et al. (2011), also known as dynamic vision sensors, produce a stream of records about brightness change in microsecond temporal resolution, complementing the temporal information of the degraded images.

Following that, we simulate the exposure process as a succession of adaptive information exchanges based on a stack of specially designed operation called *temporalized attention*. Through a manually specified time encoding called *neural shutter*, we can control the exposure process to a desired state.

Akin to the film in a traditional camera, we design a structure referred to as *neural film* as the carrier for visual content. The neural film together with the visual content representation goes through several rounds of attention-based information exchange, retrieving the visual content specified by the neural shutter. By appropriately adjusting the neural shutter, we can manipulate the visual content of the resulting image, optimizing it to suit various applications.

Through the proposed architecture, we can address visual issues such as motion blur, video discontinuity, rolling shutter artifacts, and even their combinations, with a single, unified model.

2. Related Works

2.1. Motion Deblur

Motion blur occurs when the object or camera moves at high speed during the exposure period. To deblur the images, some methods Ren et al. (2020); Kaufman and Fattal (2020) model make estimation about blur kernel first and conduct deconvolution with the estimated kernel. Some methods Nah et al. (2017a); Cho et al. (2021); Chen et al. (2022) adopt the encoder-decoder architectures to deblur images with neural network. Due to the complexity of blur patterns and lack of motion information within the exposure period, the performance of these methods is still limited especially when it comes to scenes with complex motion.

Benefiting from the rich temporal information with the events, event-based methods Pan et al. (2019); Jiang et al. (2020); Lin et al. (2020); Zhang and Yu (2022); Song et al. (2022); Xu et al. (2021) have achieved significant progress. Pan *et al.* Pan et al. (2019) proposed the Event-based Double Integral (EDI) model by exploring the relationship between events, blurry images, and the latent sharp image to deblur the image by optimizing an energy function. Considering the impact of noise and the unknown threshold of events, some methods Jiang et al. (2020); Lin et al. (2020); Zhang and Yu (2022) use deep learning networks to predict the sharp image based on the same principle. Song *et al.* Song et al. (2022) model the motion by means of per-pixel parametric polynomials with a deep learning model. REDNet *et al.* Xu et al. (2021) estimates the optical flow with the event to supervise the deblurring model with blurry consistency and photometric consistency. By investigating the impact of light on event noise, Zhou *et al.* Zhou et al. (2021) attempted to estimate the blur kernel with events to deblur images by deconvolution. Sun *et al.* Sun et al. (2022) proposed a cross-modality channel-wise attention module to fuse event features and image features at multiple levels.

2.2. Video Frame Interpolation

Most frame-only methods Jiang et al. (2018); Lee et al. (2020); Bao et al. (2019); Huang et al. (2022) are based on linear motion assumption. These methods estimate the optical flow according to the difference between two frames, and linearly calculate the displacement from the key frames to the target timestamp. Because of lack of motion information between frames.

Compared with frame-only interpolation, event-based interpolation methods are more effective due to the power of events in motion modeling. This makes them competent for scenarios with more complex motion patterns. Xu *et al.* Xu et al. (2021) proposed to predict optical flow between output frames to simulate nonlinear motion within exposure duration. He *et al.* He et al. (2022) proposed an unsupervised event-assisted video frame interpolation framework by cycling the predicted intermediate frames in extra rounds of frame interpolation. Tulyakov *et al.* Tulyakov et al. (2021) designed a frame interpolation framework by combining a warping-based branch and a synthesis-based branch to fully exploit the advantage of fusion of frames and events.

2.3. Rolling Shutter Correction

Rolling shutter effect is caused by the row-by-row readout scheme, in which each row of pixels is exposed at a different time. Frame-only unrolling is mostly based on the motion flow and linear motion assumption. Fan *et al.* Fan and Dai (2021); Fan et al. (2022) proposed to estimate the motion field between two adjacent input rolling shutter images, and predict the global shutter image based on that. In SUNet Fan et al. (2021) and DSUN Liu et al. (2020) pyramidal cost volume is computed to predict motion field and global shutter image is predicted by warping features of key frames based on that. Zhou *et al.* Zhou et al. (2022) introduced the event data to the unrolling task, and designed a two-branch structure which fully leverages information with frames and events to correct the rolling shutter effect.

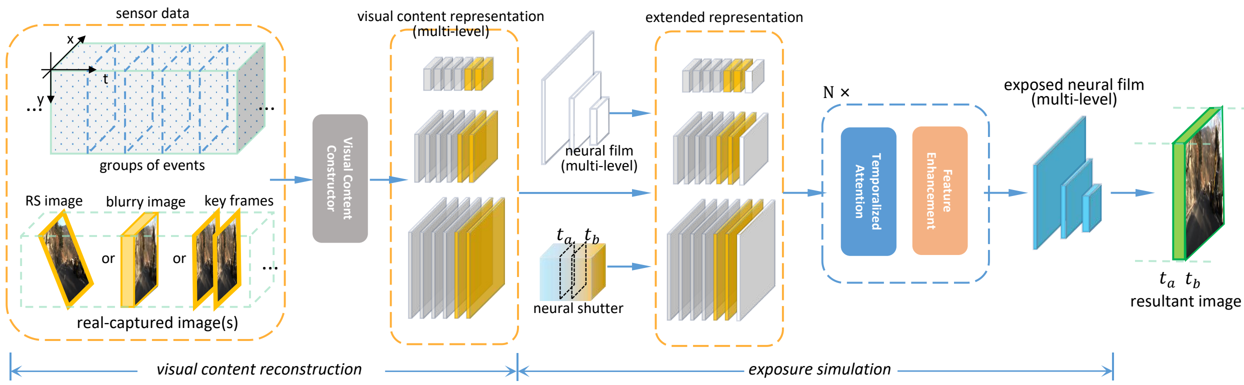

Fig. 2. Overall pipeline. A multi-level representation of the visual content is constructed from the sensor data by the visual content constructor. Then, together with the visual content, a multi-level neural film is fed into the exposure simulator. By specifying a desired neural shutter, a desired re-exposed image can be produced.

2.4. Joint Tasks

There have already been some efforts in dealing with multiple tasks simultaneously. Some methods Zhang and Yu (2022); Lin et al. (2020); Oh and Kim (2022) deal with image deblur and frame interpolation simultaneously. DeMFI Oh and Kim (2022) takes blurry key frames as input, deblurring the image with a flow-guided module and interpolating sharp frames with a recursive boosting module. Zhang *et al.* Zhang and Yu (2022) and Lin *et al.* Lin et al. (2020) unifid the image deblur and frame interpolation with the help of events. EVDI Zhang and Yu (2022) predicts sharp images of a given timestamp by leveraging blurry images and corresponding events, which are then fused as interpolation results. Lin *et al.* Lin et al. (2020) proposed to use events to estimate the residuals for the sharp frame restoration, and the restored frames compose a video of higher framerate.

Zhong *et al.* Zhong et al. (2021) and Zhou *et al.* Zhou et al. (2022) proposed methods to convert blurry rolling shutter images into sharp global shutter images. JCD Zhong et al. (2021) joint address motion blur and rolling shutter effect with a bi-directional warping stream and a middle deblurring stream. EvUnroll Zhou et al. (2022) is an event-based method that deblurs the blurry rolling shutter image first, then corrects the rolling shutter effects in a two-branch structure.

It is worth noting that, although above methods address multiple issues in a single model, they handle each aspect of the joint task with a corresponding module in a multi-stage manner. In this work, we propose a unified framework to deal with all shutter-related problems. By re-exposing the captured image with a desired shutter, all aspects of the joint task can be addressed in a unified way.

3. Re-exposure Paradigm

In this section, we derive a symbolic expression to illustrate the re-exposure paradigm.

The re-exposure paradigm is derived from the relationship between the visual content, the spatial-temporal domain determined by the exposure process, and the resulting image. For an image $I(x, y)$, the pixel at (x, y) is determined by integrating the visual content $V(x, y, t)$ over the exposure period $[t_a(x, y), t_b(x, y)]$. Mathematically, this can be expressed as:

$$I(x, y) = \int_{t_a(x,y)}^{t_b(x,y)} V(x, y, t)\, dt, \tag{1}$$

It is worth noting that the exposure period $[t_a(x, y), t_b(x, y)]$ may vary with the position (x, y). This flexibility is to accommodate scenarios such as the rolling shutter camera, where the exposure period varies across different positions.

It can be observed that each image represents visual content within a certain spatial-temporal domain, which can be denoted as $\Omega = [0, W] \times [0, H] \times [t_a(x, y), t_b(x, y)]$. By introducing a shutter function corresponding to the spatial-temporal domain, we can decouple an image into the visual content and a shutter function, leading to the equation as follows:

$$I(x, y) = \int_0^T V(x, y, t)\, S(x, y, t)\, dt, \tag{2}$$

Here, $S(\cdot)$ represents the shutter function, defined as:

$$\begin{gathered} S(x, y, t) = \mathbb{1}_{t>0}(t - t_a(x, y)) - \mathbb{1}_{t>0}(t - t_b(x, y)), \\ \text{s.t. } 0 < t_a(x, y) < t_b(x, y) < T, \end{gathered} \tag{3}$$

where $\mathbb{1}_{t>0}(\cdot)$ is the unit step function. Notice the integral limits have been extended to $[0, T]$ which encompasses Ω, indicating the visual content of interest distributes within a larger time span than any shutter function.

Under this framework, different types of images correspond to different shutter functions. For example, a global shutter image corresponds to a shutter function with $t_a(x, y) = t_1$ and $t_b(x, y) = t_2$, where t_1 and t_2 are constant for all position. An image captured by a rolling shutter camera corresponds to a shutter function with $t_a(x, y) = t_1 + \alpha y$ and $t_b(x, y) = t_2 + \alpha y$, where α represents the readout delay between adjacent rows. And for the blurry image, $|t_a(x, y) - t_b(x, y)|$ is typically large.

However, there remains an issue: the overall intensity of $I(x, y)$ is positively related to $|t_a(x, y) - t_b(x, y)|$—the smaller it is, the darker the resulting image will be. In particular, for an image representing a specific moment, $|t_a(x, y) - t_b(x, y)| = 0$ will lead to an entirely black image.

To address this problem, we introduce a normalized shutter function to better reflect the relationship.

$$\bar{S}(x, y, t) = \frac{S(x, y, t)}{|S(x, y, t)|}, \tag{4}$$

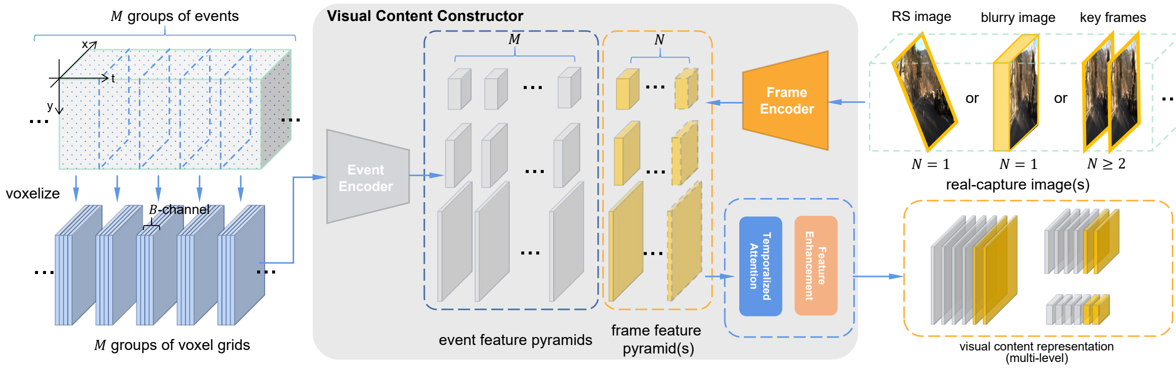

Fig. 3. Illustration of the visual content constructor. It first extracts features from events and images, produces a set of multi-level feature pyramids. The feature pyrmaids are then processed by temporalized attention and feature enhancement, resulting in a multi-level representation of visual content.

In particular, when $|t_a(x,y)-t_b(x,y)| = 0$ and $t_a(x,y) = t_b(x,y) = t_0$, where t_0 is a constant timestamp, we define $\bar{S}(x,y,t) = \delta(t - t_0)$. We then obtain the following exposure formula:

$$I(x,y) = \int_0^T V(x,y,t)\,\bar{S}(x,y,t)\,dt, \tag{5}$$

Now, we express the relationship between the visual content, the shutter function, and the resulting image in an exposure process. According to Eq. 5, given the visual content and the shutter function, we can derive the corresponding image, which is the core of the re-exposure paradigm. For any specific task, we can specify the shutter function on the requirement to address corresponding visual issues.

Following the re-exposure paradigm, we expect to approximate the relationship reflected by Eq. 5 with a neural network, which can be abstracted as:

$$\mathbf{I}_{\Omega|\mathbf{V}} = f(F(\mathbf{V}), \Omega) \tag{6}$$

Here, $\mathbf{V} = \{V(x,y,t)|(x,y,t) \in [0,H] \times [0,W] \times [0,T]\}$ can be regarded as a tensor sampled from V(x, y, t), which serves as the input of the neural network. $\mathbf{I}_{\Omega|\mathbf{V}} = \{I(x,y)|(x,y) \in [0,H] \times [0,W]\}$ is the image corresponding to the given visual content and shutter function. $F(\cdot)$ serves as the feature extractor, mapping the visual content to the feature domain, while $f(\cdot)$ simulates the exposure process, retrieving a subset of the visual content to produce the desired images.

However, in practical applications, $V(x,y,t)$ is not initially provided, and the given images $\hat{\mathbf{I}}$ are typically degraded. Therefore, we need to construct the visual content representation from the sensor data. Considering the degradation of $\hat{\mathbf{I}}$, we incorporate event camera data $E = \{(x,y,p,t)|t \in [0,T]\}$ as a supplement, which is a stream of records about brightness change in microsecond temporal resolution. In this way, we can approximate the visual content V in feature domain with the events E and the given degraded image $\hat{\mathbf{I}}$:

$$F(\mathbf{V}) = g(\hat{\mathbf{I}}, E) \tag{7}$$

Finally, we derive the following expression representing our method:

$$\mathbf{I}_{\Omega|\hat{I},E} = f(g(\hat{\mathbf{I}}, E), \Omega) \tag{8}$$

This suggests that given degraded images $\hat{\mathbf{I}}$ and a chunk of events E, we can obtain desired image by manipulating the spatial-temporal domain Ω.

4. Method

In this section, we approximate Eq. 8 with a neural network, which is an architecture we term **N**eural **I**mage **Re**-**E**xposure (**NIRE** for short). The overall architecture is shown in Fig. 2. NIRE first constructs a visual content representation from the sensor data, including images and events. It then simulates the exposure process under the control of a neural shutter mechanism. The neural film retrieves the visual content specified by the neural shutter in this process, producing an image with desired content and quality.

4.1. Feature Extraction

As shown in Fig. 3, to obtain the visual content representation from the degraded image $\hat{I}$ and events E, the visual content constructor first extract their features respectively.

To process events with convolutional neural network, we split the events E into M segments by time. Each segment is converted to a voxel grid Zhu et al. (2018) with B bins, which is fed into a bi-directional LSTM Hochreiter and Schmidhuber (1997), obtaining M feature pyramids, $\{\mathcal{E}_1^l, \mathcal{E}_2^l, \cdots, \mathcal{E}_M^l\}_{l=1}^L$, with each feature pyramid $\mathcal{E}_i^l \in \mathbb{R}^{C_l \times \frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}}}$ and L is the total number of levels, and C_l is the number of channels of the l-th level.

As for the degraded images, each of them is processed by a fully convolutional multi-scale encoder, producing a feature pyramid $\mathcal{I}_i^l \in \mathbb{R}^{C_l \times \frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}}}$, composing a set of feature pyrmids $\{\mathcal{I}_1^l, \cdots, \mathcal{I}_N^l\}_{l=1}^L$. Here the number of images N depends on the task, *e.g.*$N = 2$ for the VFI task and $N = 1$ for the image deblur task.

Through the feature extraction process, we can obtain a set of feature pyramids $\{\mathcal{E}_1^l, \mathcal{E}_2^l, \cdots, \mathcal{E}_M^l, \mathcal{I}_1^l, \cdots, \mathcal{I}_N^l\}_{l=1}^L$, which will be used in the construction of visual content representation (to be illustrated in Sec. 4.3).

4.2. Temporalized Attention

Before we proceed to the construction of the visual content representation, we need to introduce an operation termed as *temporalized attention*, which plays a critical role in both the construction of the visual content representation and in the exposure simulation process.

It should be noted that each extracted feature pyramid corresponds to specific spatial-temporal domains. To pinpoint their spatial-temporal position accurately and process their relationships, we have designed the temporalized attention.

Following the standard vision transformer Dosovitskiy et al. (2021), the feature tokens are initially projected to d-dimension queries Q, keys K and values V with three linear layers f_Q, f_K, and f_V respectively, as illustrated in Eq. 9.

$$[Q, K, V] = [f_Q(Z), f_K(Z), f_V(Z)] \tag{9}$$

Different from vision transformer Dosovitskiy et al. (2021), the proposed operation works with our specially designed time-related positional encodings. For a timestamp t, we can encode it into a sinusoidal positional encoding:

$$\gamma(t) = \left(\sin\left(2^0 \pi t\right), \cos\left(2^0 \pi t\right), \cdots, \sin\left(2^{K-1} \pi t\right), \cos\left(2^{K-1} \pi t\right)\right) \tag{10}$$

where $t \in [0, 1]$ represents a normalized timestamp, with $t = 0$ and $t = 1$ indicating the temporal boundaries of the visual content of interest.

By concatenating the encodings of the start and end timestamps of a certain range, we can describe the time range with:

$$\mathcal{T}(t_a, t_b) = [\gamma(t_a(x, y)), \gamma(t_b(x, y))]. \tag{11}$$

Then, the encodings are also projected to d-dimension by a linear layer f_T. And we can obtain the time-aware queries $\widetilde{Q}$ and keys $\widetilde{K}$ through the following *temporalize* operation

$$\widetilde{Q} = Q + f_T(\mathcal{T}), \widetilde{K} = K + f_T(\mathcal{T}). \tag{12}$$

Ultimately, the temporalized attention can be denoted as:

$$Attention(\widetilde{Q}, \widetilde{K}, V) = softmax(\widetilde{Q}\widetilde{K}^T / \sqrt{d})V. \tag{13}$$

Following vision transformer Dosovitskiy et al. (2021), temporlizaed attention adopts the multi-head design, and the usage of LayerNorm and FFN are kept unchanged.

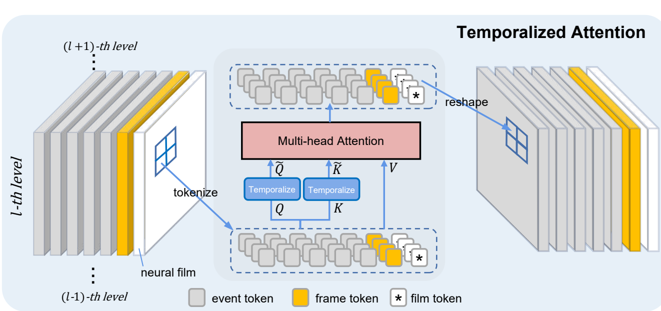

Fig. 4. Illustration of the temporalized attention module.

As shown in Fig. 4, temporalized attention takes n feature maps at a certain level as input, resulting in n feature maps at the same level, where n is the total number of the feature maps [1]. To mitigate the computational burden, the feature maps are divided into non-overlapping $r \times r$ windows, and the attention operation is applied to the $n \times r \times r$ tokens within each window.

[1] $n = N + M$ for visual content representation, where N and M are the numbers of event and image based feature maps respectively; $n = N + M + 1$ for the extend visual content representation, where the additional one feature map is the neural film.

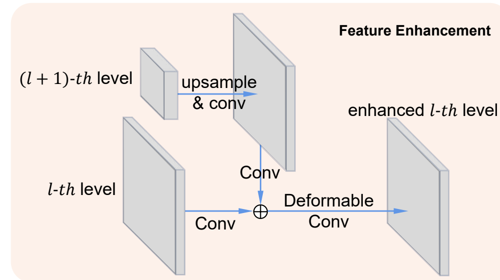

Fig. 5. Illustration of the feature enhancement module. The l-th feature level is fused with the upsampled $(l + 1)$-th level by addition, and the fused feature is processed by a deformable convolution.

Following each temporalized attention, we apply a feature enhancement module to promote the interaction across different feature levels and windows. As shown in Fig. 5, it fuses the features with their coarser level by upsampling and addition, and processes the fused features with a deformable convolution Dai et al. (2017) to get the enhanced feature.

4.3. Visual Content Representation

As shown in Fig. 3, a set of feature pyramids is obtained after the feature extraction. We intend to unify these features with temporalized attention to represent the visual content.

Notice that each token in the temporalized attention requires a time-related positional encoding to pinpoint its temporal position. For the tokens originating from image features, their time encodings encode the start and end timestamps of their exposure period. For the tokens derived from event features, their time encodings represent the start and end timestamps of their corresponding event segments. After applying temporalized attention and feature enhancement, the set of feature pyramids interact with each other, producing an updated feature pyramid set denoted as $\{\hat{\mathcal{E}}^l_1, \hat{\mathcal{E}}^l_2, \cdots, \hat{\mathcal{E}}^l_M, \hat{\mathcal{I}}^l_1, \cdots, \hat{\mathcal{I}}^l_N\}^L_{l=1}$, among which each feature pyramid represents part of the whole visual content within a certain spatial-temporal domain.

4.4. Neural Film, Neural Shutter, and Exposure Simulation

To retrieve the visual content of a certain spatial-temporal domain from the whole visual content representation, we design structures termed *neural film* and *neural shutter*.

The neural film serves as the carrier of visual content, akin to the film in a camera. A neural film is a predefined multi-level feature pyramid, each level is initialized by replicating a learnable vector throughout spatial dimensions. Symbolically, the neural film can be denoted as $\{\mathcal{X}^l_0\}^L_{l=1}$, where each level $\mathcal{X}^l_0 \in \mathbb{R}^{C_l \times \frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}}}$ has the same shape as the feature levels in the visual content representation.

The neural shutter is a manually specified time encoding, pinpointing a spatial-temporal domain whose visual content is expected to be represented by the resulting image. In the exposure simulation, the neural shutter serves as the positional encoding for the neural film in the temporalized attention.

In the exposure simulation, we append the neural film to the feature pyramid set representing visual content, obtaining an extended representation denoted as

$\{\hat{\mathcal{E}}_1^l, \hat{\mathcal{E}}_2^l, \cdots, \hat{\mathcal{E}}_M^l, \hat{\mathcal{I}}_1^l, \cdots, \hat{\mathcal{I}}_N^l, \mathcal{X}_0^l\}_{l=1}^L$. We feed the extended representation to the temporalized attention, where neural film retrieves the visual content from the spatial-temporal domain specified by the neural shutter, resulting in the exposed film which is a feature pyramid encoding our desired image. The exposed neural film is then sent to a convolutional decoder level by level in a top-down manner similar to the FPN Lin et al. (2016) structure, where each feature level is processed by convolutional layers and upsampled to fuse with a finer level. Finally, the finest feature level is decoded into a normalized sRGB image that retains the desired content and meets the standard of quality that we require.

5. Experiments

Since the proposed NIRE method is able to deal with several image/video quality issues within a unified framework, we evaluate it on multiple tasks including image deblur, video frame interpolation (VFI), rolling shutter (RS) correction, and jointly deblurring and frame interpolation.

5.1. Datasets

Two datasets, GoPro Nah et al. (2017b) and Gev-RS Zhou et al. (2022), are used for training and quantitative evaluation in our experiments. GoPro Nah et al. (2017b) is a dataset consisting of sequences shot by a GoPro camera with a frame rate of 240 FPS and a resolution of 1,280×720. It can provide training and testing samples for tasks including image deblur Sun et al. (2022) Kupyn et al. (2018) Tao et al. (2018), frame interpolation Tulyakov et al. (2021) Bao et al. (2019) Jiang et al. (2018), and jointly deblurring and frame interpolation Oh and Kim (2022) Jin et al. (2019). Gev-RS Zhou et al. (2022) is a dataset collected for event-base rolling shutter correction. It is composed of 5,700 FPS video sequences recorded by Phantom VEO 640 high-speed camera such that high-quality RS images and event streams can be simulated. For each task, we follow its common evaluation protocol for fair comparison.

5.2. Training Strategy

In the tasks of interest, the degraded images for training are synthesized, while the original high quality images serve as the groundtruths. For example, a blurry image is synthesized through averaging several sharp frames, a low-framerate video is synthesized by subsampled high-framerate ones, a rolling shutter image is created by composing scanlines from a series of frames. And considering the scarce of calibrated events and images, we adopt the widely used event simulator Hu et al. (2021a,b) for generating the events.

NIRE takes arbitrary types of low-quality images/frames and events as inputs, while original, high-quality images/frames serve as the ground truths. In the forward pass, we first feed the degraded image of random type (e.g. blurry image, sharp image, RS image, blurry RS image, etc.) accompanied with a segment of events that temporally encompasses the degraded image Here 'temporally encompass' suggests that the temporal range of the events should exceed that of the given image. Then we set the neural shutter to encode the timestamp of an available ground truth [2]. This instructs NIRE to predict an image similar to the given ground truth as much as possible, therefore the output image is then compared with the ground truth with a combination of Charbonnier loss Charbonnier et al. (1994) and perceptual loss Johnson et al. (2016), providing supervision in the backward pass. During training, the input images are randomly cropped into 128×128 patches, and we train our model for 60,000 iterations with a batch size of 32 on a Tesla A100 GPU.

5.3. Deblur

Following the experiment setting in Pan et al. (2019); Sun et al. (2022), the 3,214 blurry-sharp image pairs in GoPro dataset are split into 2,103 pairs for training and 1,111 pairs for testing. The blurred images are synthesized by averaging consecutive high-framerate sharp frames.

Table 1. Performance on image deblur.

Methods	event	PSNR	SSIM
E2VID Rebecq et al. (2019)	✓	15.22	0.651
DeblurGAN Kupyn et al. (2018)	✗	28.70	0.858
EDI Pan et al. (2019)	✓	29.06	0.940
DeepDeblur Nah et al. (2017a)	✗	29.08	0.914
DeblurGAN-v2 Kupyn et al. (2019)	✗	29.55	0.934
SRN Tao et al. (2018)	✗	30.26	0.934
SRN+ Tao et al. (2018)	✓	31.02	0.936
DMPHN Zhang et al. (2019)	✗	31.20	0.940
D^{2}Nets Shang et al. (2021)	✓	31.60	0.940
LEMD Jiang et al. (2020)	✓	31.79	0.949
Suin et al. Suin et al. (2020)	✗	31.85	0.948
SPAIR Purohit et al. (2021)	✗	32.06	0.953
MPRNet Zamir et al. (2021)	✗	32.66	0.959
HINet Chen et al. (2021)	✗	32.71	0.959
ERDNet Chen et al. (2020)	✓	32.99	0.935
HINet+ Chen et al. (2021)	✓	33.69	0.961
NAFNet Chen et al. (2022)	✗	33.69	0.967
DFFN Kong et al. (2023)	✗	34.21	0.969
DSTN Pan et al. (2023)	✗	35.05	**0.973**
EFNet Sun et al. (2022)	✓	**35.46**	0.972
NIRE	✓	35.03	**0.973**

As shown in Tab. 1 and Fig. 6, the proposed NIRE outperforms most frame-only methods, and achieves comparable performance with the competitive event-based method EFNet. This demonstrates the effectiveness of our proposed method. Most existing methods restore the sharp frame of a fixed timestamp (*e.g.*middle of exposure time). In contrast, NIRE is able to derive sharp images of arbitrary specified timestamps. Furthermore, by specifying the neural shutter to differet width, the sharpness of the output image can be controlled, as shown in Fig. 7(a)(b).

5.4. Video Frame Interpolation

To validate the effectiveness of our method on VFI task, we evaluate the proposed NIRE method following the same setting as event-based VFI method Tulyakov et al. (2021) on GoPro. As shown in Tab. 2, NIRE achieves much better performance than conventional frame-only methods and is on par

[2]An available ground truth refers to a high quality image involved in the synthesis of the degraded images

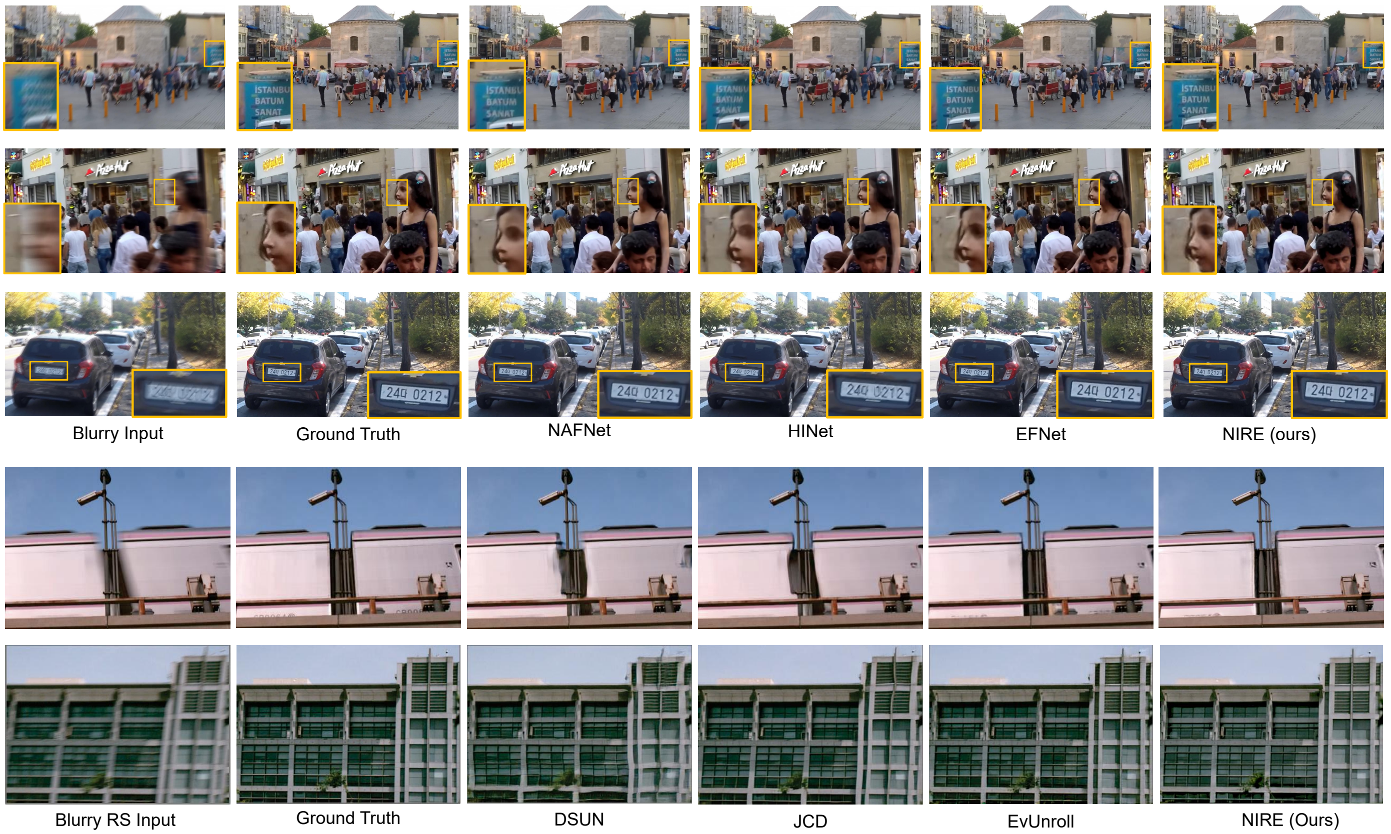

Fig. 6. Qualitative result of NIRE dealing with degraded images. For optimal viewing, please zoom in.

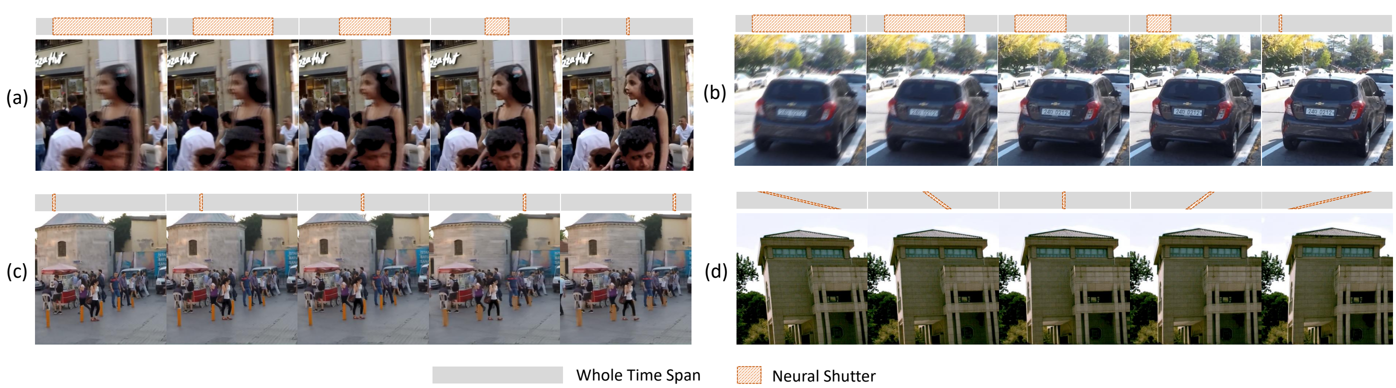

Fig. 7. Illustration of the neural shutter and the resulted images. For optimal viewing, please zoom in.

Fig. 8. Illustration of the NIRE recovering sharp images of arbitrary specified timestamps.

with the specially designed event-based VFI method TimeLens Tulyakov et al. (2021). Fig. 7(c) and Fig. 8 gives illustration about intermediate frames predicted at arbitrary normalized timestamp.

5.5. *Joint Deblur and Rolling Shutter Correction*

The proposed NIRE method is also validated on the RS correction task, following the experiment setting of EvUnroll Zhou et al. (2022).

Benefit from the visual content constructor, NIRE is able to

Table 2. **Performance on video frame interpolation**

Method	frames	events	7 frames skip		15 frames skip	
			PSNR	SSIM	PSNR	SSIM
DAIN Bao et al. (2019)	✓	✗	28.81	0.876	24.39	0.736
SuperSloMo Jiang et al. (2018)	✓	✗	28.98	0.875	24.38	0.747
RRIN Li et al. (2020)	✓	✗	28.96	0.876	24.32	0.749
BMBC Park et al. (2020)	✓	✗	29.08	0.875	23.68	0.736
EMA-VFI Zhang et al. (2023)	✓	✗	32.79	0.942	29.70	0.904
E2VID Rebecq et al. (2019)	✗	✓	9.74	0.549	9.75	0.549
EDI Pan et al. (2019)	✓	✓	18.79	0.670	17.45	0.603
TimeLens Tulyakov et al. (2021)	✓	✓	34.81	0.959	**33.21**	0.942
NIRE	✓	✓	**34.97**	**0.964**	32.85	**0.945**

Table 3. Performance on joint deblur and RS correction. ***unroll+deblur*** **indicates using blurry RS images as input and performing both deblur and unroll tasks simultaneously, while** ***unroll*** **indicates using sharp RS image as input and only performing the unroll task.**

Methods	events	PSNR	SSIM
DSUN Liu et al. (2020)(unroll)	✗	23.10	0.70
JCD Zhong et al. (2021)(unroll)	✗	24.90	0.82
EvUnroll Zhou et al. (2022)(unroll+deblur)	✓	**30.14**	**0.91**
EvUnroll Zhou et al. (2022)(unroll)	✓	**32.16**	**0.91**
NIRE(unroll+deblur)	✓	29.86	**0.91**
NIRE(unroll)	✓	31.75	**0.91**

Table 4. Performance on joint deblur and frame interpolation.

Methods	unified	events	PSNR	SSIM
SRN Tao et al. (2018) + SloMo Jiang et al. (2018)	✗	✗	24.72	0.7604
SRN + MEMC-Net Bao et al. (2021)	✗	✗	25.70	0.7792
SRN + DAIN Bao et al. (2019)	✗	✗	25.17	0.7708
EDVR Wang et al. (2019) + SloMo	✗	✗	24.85	0.7762
EDVR + MEMC-Net	✗	✗	27.12	0.8301
EDVR + DAIN	✗	✗	29.01	0.8981
UTI-VFI	✓	✗	25.63	0.8148
EVDI Zhang and Yu (2022)	✓	✓	25.89	0.7922
PRF Shen et al. (2021)	✓	✗	25.68	0.8053
TNTT Jin et al. (2019)	✓	✗	26.68	0.8148
DeMFI-Net Oh and Kim (2022)	✓	✗	31.25	0.9102
NIRE-cascade	✗	✓	30.18	0.8923
NIRE	✓	✓	**33.43**	**0.9477**

construct the visual content representation from the images with motion blur and rolling shutter effect. Once the visual content representation is constructed, we can retrieve arbitrary desired global shutter image free of motion blur.

As shown in Tab. 3 and Fig. 6, NIRE outperforms the frame-only methods and achieves comparable performance with the SOTA event-based method EvUnroll Zhou et al. (2022), demonstrating the effectiveness of NIRE on jointly removing rolling shutter artifact and blur.

5.6. *Joint Deblur and Frame Interpolation*

In addition, the proposed method is also validated on the task of joint deblur and frame interpolation following the same setting as DeMFI Oh and Kim (2022). The conventional VFI task usually assumes the given key frames are sharp. Nonetheless, videos that require interpolation are often degraded by blur induced by either camera motion or object movement, which degrades the interpolation results.

Simply cascading an image deblur model and a VFI model is a direct solution, but it will lead to error accumulation and suboptimal performance. In contrast, NIRE inherently resolves all visual quality issues simultaneously. As shown in Tab. 4, NIRE outperforms existing frame-only methods by a large margin, showing its advantage in handling the joint task. We also try to apply NIRE twice, one for deblur and one for frame interpolation, resulting in a pipeline denoted as NIRE-cascade. It achieves significantly worse performance than addressing them in the unified manner, showing the advantage of re-exposure paradigm.

5.7. *Ablation Study*

Ablation study is conducted to investigate importance of components of the proposed framework. In Tab. 5, ‘NIRE w/o event’ represents the baseline with the visual content representation is construct only based on the frame, without incorporating the events. ‘NIRE w/o TimEnc’ denotes the NIRE by simply disabling the time encodings. ‘NIRE w/o FeatEnhance‘ denotes the NIRE without feature enhancement module. The results show all these components are necessary for our proposed architecture.

Table 5. Ablation study of NIRE (in PSNR/SSIM and Flops/Params).

Tasks	VFI	Deblur	Unroll	Deblur+VFI	Flops(G)/Params(M)
NIRE	34.97/0.964	35.03/0.973	29.86/0.908	33.43/0.948	438.8/33.2
w/o Event	30.40/0.886	29.53/0.928	24.08/0.803	26.46/0.815	321.7/25.6
w/o TimEnc	31.23/0.921	33.44/0.955	20.38/0.584	29.76/0.874	437.8/33.2
w/o FeatEnhance	32.83/0.928	33.78/0.952	26.42/0.835	30.62/0.894	435.2/33.0

In addition, we compare specialized and versatile NIRE models by restricting the training data. Specifically, when we restrict the training data to blurry-sharp pairs, the NIRE model is specialized for deblur. When we restrict the training data to RS-GS pairs, the NIRE model is specialized for Unrolling task. When we restrict the training data to keyframe and intermediate frames, the NIRE model is specialized for VFI task. As shown in Tab. 6, the re-exposure pardigm is not only versatile, but also performs on-par with or even better than specilized counterparts, demonstrating different tasks are naturally unified, without conflicting with each other.

Table 6. Comparison of specialized and versatile NIRE (in PSNR/SSIM).

strategy / task	MT	VFI	Deblur	Unroll
VFI	34.97/0.964	34.44/0.955	-	-
Deblur	35.03/0.973	-	34.72/0.966	-
Unroll	30.08/0.909	-	-	30.04/0.909

6. Conclusion

In this work, we highlight that a variety of visual issues can be attributed to sub-optimal exposure. Through a paradigm called re-exposure, the degraded images can be restored in a controllable way. Following the re-exposure paradigm, a novel architecture called NIRE is proposed, which constructs representation of visual content from images and events and performs exposure simulation under the control of a neural shutter. By adjusting the simulated exposure to a desired state, the proposed method can be used to address multiple tasks, including deblur, rolling shutter correction, and joint deblur and frame interpolation.

Acknowledgments

The research was partially supported by the National Natural Science Foundation of China, (grants No. 62106036, U23B2010, 62206040, 62293540, 62293542), the Dalian Science and Technology Innovation Fund (grant No. 2023JJ11CG001).

References

Bao, W., Lai, W., Ma, C., Zhang, X., Gao, Z., Yang, M., 2019. Depth-aware video frame interpolation, in: CVPR.

Bao, W., Lai, W., Zhang, X., Gao, Z., Yang, M., 2021. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. TPAMI 43, 933–948.

Charbonnier, P., Blanc-Féraud, L., Aubert, G., Barlaud, M., 1994. Two deterministic half-quadratic regularization algorithms for computed imaging, in: ICIP, pp. 168–172.

Chen, H., Teng, M., Shi, B., Wang, Y., Huang, T., 2020. Learning to deblur and generate high frame rate video with an event camera. CoRR abs/2003.00847.

Chen, L., Chu, X., Zhang, X., Sun, J., 2022. Simple baselines for image restoration, in: ECCV.

Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C., 2021. Hinet: Half instance normalization network for image restoration, in: CVPRW.

Cho, S., Ji, S., Hong, J., Jung, S., Ko, S., 2021. Rethinking coarse-to-fine approach in single image deblurring, in: ICCV.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks, in: ICCV.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image at scale, in: ICLR.

Fan, B., Dai, Y., 2021. Inverting a rolling shutter camera: Bring rolling shutter images to high framerate global shutter video, in: ICCV.

Fan, B., Dai, Y., He, M., 2021. Sunet: Symmetric undistortion network for rolling shutter correction, in: ICCV.

Fan, B., Dai, Y., Zhang, Z., Liu, Q., He, M., 2022. Context-aware video reconstruction for rolling shutter cameras, in: CVPR.

He, W., You, K., Qiao, Z., Jia, X., Zhang, Z., Wang, W., Lu, H., Wang, Y., Liao, J., 2022. Timereplayer: Unlocking the potential of event cameras for video interpolation, in: CVPR.

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

Hu, Y., Liu, S., Delbrück, T., 2021a. v2e: From video frames to realistic DVS events, in: CVPRW.

Hu, Y., Liu, S., Delbrück, T., 2021b. v2e: From video frames to realistic DVS events, in: CVPRW.

Huang, Z., Zhang, T., Heng, W., Shi, B., Zhou, S., 2022. Real-time intermediate flow estimation for video frame interpolation, in: ECCV.

Jiang, H., Sun, D., Jampani, V., Yang, M., Learned-Miller, E.G., Kautz, J., 2018. Super slomo: High quality estimation of multiple intermediate frames for video interpolation, in: CVPR.

Jiang, Z., Zhang, Y., Zou, D., Ren, J.S.J., Lv, J., Liu, Y., 2020. Learning event-based motion deblurring, in: CVPR.

Jin, M., Hu, Z., Favaro, P., 2019. Learning to extract flawless slow motion from blurry videos, in: CVPR.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution, in: ECCV.

Kaufman, A., Fattal, R., 2020. Deblurring using analysis-synthesis networks pair, in: CVPR.

Kong, L., Dong, J., Ge, J., Li, M., Pan, J., 2023. Efficient frequency domain-based transformers for high-quality image deblurring, in: CVPR.

Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J., 2018. Deblurgan: Blind motion deblurring using conditional adversarial networks, in: CVPR.

Kupyn, O., Martyniuk, T., Wu, J., Wang, Z., 2019. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better, in: ICCV.

Lee, H., Kim, T., Chung, T., Pak, D., Ban, Y., Lee, S., 2020. Adacof: Adaptive collaboration of flows for video frame interpolation, in: CVPR.

Li, H., Yuan, Y., Wang, Q., 2020. Video frame interpolation via residue refinement, in: ICASSP.

Lichtsteiner, P., Posch, C., Delbruck, T., 2008. A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. JCCS 43, 566–576.

Lin, S., Zhang, J., Pan, J., Jiang, Z., Zou, D., Wang, Y., Chen, J., Ren, J.S.J., 2020. Learning event-driven video deblurring and interpolation, in: ECCV.

Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J., 2016. Feature pyramid networks for object detection. CoRR abs/1612.03144.

Liu, P., Cui, Z., Larsson, V., Pollefeys, M., 2020. Deep shutter unrolling network, in: CVPR.

Nah, S., Kim, T.H., Lee, K.M., 2017a. Deep multi-scale convolutional neural network for dynamic scene deblurring, in: CVPR.

Nah, S., Kim, T.H., Lee, K.M., 2017b. Deep multi-scale convolutional neural network for dynamic scene deblurring, in: CVPR.

Oh, J., Kim, M., 2022. Demfi: Deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting, in: ECCV.

Pan, J., Xu, B., Dong, J., Ge, J., Tang, J., 2023. Deep discriminative spatial and temporal network for efficient video deblurring, in: CVPR.

Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y., 2019. Bringing a blurry frame alive at high frame-rate with an event camera, in: CVPR.

Park, J., Ko, K., Lee, C., Kim, C., 2020. BMBC: bilateral motion estimation with bilateral cost volume for video interpolation, in: ECCV.

Posch, C., Matolin, D., Wohlgenannt, R., 2011. A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. JSSC 46, 259–275.

Purohit, K., Suin, M., Rajagopalan, A.N., Boddeti, V.N., 2021. Spatially-adaptive image restoration using distortion-guided networks, in: ICCV.

Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D., 2019. Events-to-video: Bringing modern computer vision to event cameras, in: CVPR.

Ren, D., Zhang, K., Wang, Q., Hu, Q., Zuo, W., 2020. Neural blind deconvolution using deep priors, in: CVPR.

Shang, W., Ren, D., Zou, D., Ren, J.S., Luo, P., Zuo, W., 2021. Bringing events into video deblurring with non-consecutively blurry frames, in: ICCV.

Shen, W., Bao, W., Zhai, G., Chen, L., Min, X., Gao, Z., 2021. Video frame interpolation and enhancement via pyramid recurrent framework. TIP 30, 277–292.

Song, C., Huang, Q., Bajaj, C., 2022. E-CIR: event-enhanced continuous intensity recovery, in: CVPR.

Suin, M., Purohit, K., Rajagopalan, A.N., 2020. Spatially-attentive patch-hierarchical network for adaptive motion deblurring, in: CVPR.

Sun, L., Sakaridis, C., Liang, J., Jiang, Q., Yang, K., Sun, P., Ye, Y., Wang, K., Gool, L.V., 2022. Event-based fusion for motion deblurring with cross-modal attention, in: ECCV.

Tao, X., Gao, H., Shen, X., Wang, J., Jia, J., 2018. Scale-recurrent network for deep image deblurring, in: CVPR.

Tulyakov, S., Gehrig, D., Georgoulis, S., Erbach, J., Gehrig, M., Li, Y., Scaramuzza, D., 2021. Time lens: Event-based video frame interpolation, in: CVPR.

Wang, X., Chan, K.C.K., Yu, K., Dong, C., Loy, C.C., 2019. EDVR: video restoration with enhanced deformable convolutional networks, in: CVPRW.

Xu, F., Yu, L., Wang, B., Yang, W., Xia, G., Jia, X., Qiao, Z., Liu, J., 2021. Motion deblurring with real events, in: ICCV.

Zamir, S.W., Arora, A., Khan, S.H., Hayat, M., Khan, F.S., Yang, M., Shao, L., 2021. Multi-stage progressive image restoration, in: CVPR.

Zhang, G., Zhu, Y., Wang, H., Chen, Y., Wu, G., Wang, L., 2023. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, IEEE. pp. 5682–5692. URL: `https://doi.org/10.1109/CVPR52729.2023.00550`, doi:`10.1109/CVPR52729.2023.00550`.

Zhang, H., Dai, Y., Li, H., Koniusz, P., 2019. Deep stacked hierarchical multi-patch network for image deblurring, in: CVPR.

Zhang, K., Ren, W., Luo, W., Lai, W., Stenger, B., Yang, M., Li, H., 2022. Deep image deblurring: A survey. IJCV 130, 2103–2130.

Zhang, X., Yu, L., 2022. Unifying motion deblurring and frame interpolation with events, in: CVPR.

Zhong, Z., Zheng, Y., Sato, I., 2021. Towards rolling shutter correction and deblurring in dynamic scenes, in: CVPR.

Zhou, C., Teng, M., Han, J., Xu, C., Shi, B., 2021. Delieve-net: Deblurring low-light images with light streaks and local events, in: ICCVW.

Zhou, X., Duan, P., Ma, Y., Shi, B., 2022. Evunroll: Neuromorphic events

based rolling shutter image correction, in: CVPR.
Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K., 2018. Unsupervised event-based optical flow using motion compensation, in: ECCVW.