communications physics

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42005-025-02142-w

Event-guided temporally super-resolved synchrotron X-ray imaging

Check for updates

Hongjian Wang [1,2], Alexander Hadjiivanov [3], Emmanuel Blazquez[3], Christian M. Schlepütz [1], Marco Stampanoni[1,2] & Goran Lovric [1]

Event cameras, as novel bio-inspired neuromorphic sensors, detect per-pixel brightness changes asynchronously. Despite their growing popularity in various applications, their potential in X-ray imaging remains largely unexplored. Synchrotron-based X-ray imaging plays a significant role in various fields of science, technology and medicine. However, time-resolved imaging still faces several challenges in achieving higher sampling rates and managing the substantial data volume. Here, we introduce an inline dual-camera setup, which leverages a high-speed CMOS camera and an event camera, aiming to temporally super-resolve the sampled frame data using sparse events. To process the data, frames and events are first aligned pixel-by-pixel using feature matching, and then used to train a deep-learning neural network. This network effectively integrates the two modalities to reconstruct the intermediate frames, achieving up to a 6-fold temporal upsampling. Our work demonstrates an event-guided temporal super-resolution approach in the X-ray imaging domain, which unlocks possibilities for future time-resolved experiments.

Event cameras are bio-inspired neuromorphic sensors and, unlike conventional frame-based cameras, they detect per-pixel brightness changes asynchronously by generating a stream of events encoding the triggered time, location, and polarity[1]. This adaptive and asynchronous feature offers several advantages, including high temporal resolution (few µs), high dynamic range (up to 140 dB), and low power consumption[2]. When combined with conventional frame-based cameras, the high-frequency changes recorded by event cameras have been reported to provide complementary information for inferring objects' nonlinear motions[3,4] and reconstructing video frames[5–7]. Recently, event-based vision is gaining popularity for a variety of applications that require fast response time or operation in low-light and dynamic environments, ranging from single-molecule localization[8], through real-time monitoring[9], to space situational awareness[10]. Moreover, recent works[11,12] have investigated different perovskite materials as novel materials for X-ray scintillating films, and coupled with an event camera to achieve high light output with minimal ghosting artifacts. However, the application of event cameras in X-ray imaging remains largely unexplored.

X-ray imaging has experienced remarkable advancements in recent decades, driven by both improvements in X-ray lab technology[13,14] as well as the emergence of third-generation and fourth-generation synchrotron radiation sources. Particularly, sources that produce partially coherent X-rays with high photon flux, such as synchrotrons, enable in situ, *operando*, and in vivo studies with unmatched spatial and temporal resolutions[15]. Thus, synchrotron-based X-ray imaging has been established as the primary technique for studying fundamental properties of various materials[16,17], such as dynamic flow in evolving liquid metallic foams[18], fast motion of the middle ear structures during sound transmission[19] and many more. To fulfill the demands of upcoming scientific applications, the pursuit of higher spatio-temporal resolution in time-resolved experiments necessitates continual innovation of detectors. Though lots of specialized cameras have been developed for scientific and industrial high-speed X-ray imaging experiments with different purposes, there is always a trade-off between spatial and temporal resolutions[20,21].

With the emergence of CMOS technology, high-speed X-ray radiography has first been demonstrated to achieve the temporal sampling rate of 5400 frames per second (FPS) in full-frame mode with 1024 × 1024 pixels and 20 µm in pixel size[22], and more than 10,000 FPS is possible with a reduced field of view (FOV)[18]. However, these high frame rates have typically been limited by onboard memories where the sampled frames were stored, and that would prevent a higher temporal sampling rate and/or under-extended scanning times. On the other side, so-called burst CMOS cameras utilize on-chip memories[23], and can achieve a sampling rate of 5 MHz with a detector size of 400 × 256 pixels and 30 µm × 21.3 µm in pixel size[24], but only 128 frames can be recorded per burst. Continuous data streaming for achieving sustainable high-speed sampling has been realized on a lower temporal sampling scale in the kHz range by several fast readout systems[25,26]. However, with current detectors, achieving a higher sampling rate while maintaining the same FOV still remains a challenging task. Also,

[1]Swiss Light Source, Paul Scherrer Institute, Villigen, Switzerland. [2]Institute for Biomedical Engineering, ETH Zurich, Zurich, Switzerland. [3]Advanced Concepts Team, European Space Research & Technology Centre, Noordwijk, The Netherlands. e-mail: goran.lovric@psi.ch

high-speed imaging experiments produce extensive data volumes (such as GBs of data in a matter of seconds), which poses further challenges in data storage and processing[27]. Therefore, it is becoming increasingly imperative to mitigate these limitations by refining data acquisition methods.

In this work, an inline dual-camera setup is designed and evaluated at the TOMCAT beamline X02DA, Swiss Light Source (SLS). As a first example, we characterize the system by dynamically imaging a sandclock in radiographic mode and present all necessary steps for achieving temporal super-resolution. The proposed system demonstrates an event-guided approach to temporally super-resolve the sampled frame data, marking an initial exploration of the potential of event cameras in X-ray imaging. We report the current performance and limitations, showcasing the potential of utilizing the event camera to reduce the data generation rate, and facilitate more efficient data transmission, processing, and visualization in future time-resolved X-ray imaging experiments.

Results

In the following, we first present results affiliated with the experimental setup and alignment. Subsequently, we evaluate the temporal super-resolution capabilities qualitatively and quantitatively.

Data acquisition and alignment

We have developed an inline dual-camera platform, as shown in Fig. 1a and b, which incorporates two distinct types of detectors: a frame-based high-speed CMOS camera[25] (called "GigaFRoST", denoted as GF hereinafter) and an event camera. As we describe in detail in the Methods section, the sample is positioned upstream of the first detector. As X-rays pass through the sample, they impinge on the first scintillator that converts them to visible light. This visible light is then captured by the GF camera, producing frame-based images. Since not all X-rays are absorbed by the first scintillator, the transmitted ones continue to the second scintillator screen that by the same principle is coupled to the event camera that captures scene changes as events. The GF and the event camera are synchronized with an external trigger signal, and equipped with different magnifying visible-light optics to achieve similar effective pixel sizes, *i.e.*, 1.1 μm and 1.2 μm, respectively (as shown in Supplementary Movie 1). In the Fresnel diffraction regime, phase contrast effects become visible under sufficiently high spatial coherence and can differ for varying propagation distances of the two scintillator positions. Although we didn't observe these effects in our setup, we have employed a Fresnel diffraction simulation framework[28] to simulate potential effects under different experimental conditions, see Supplementary Note 1.

The second part of our imaging platform is shown in Fig. 1c and represents an event-guided data processing pipeline for temporal super-resolution that processes the data from the two detectors. Since the different pixel sizes of these two cameras inevitably lead to imperfect spatial alignment between frames and events, the two modalities need to be precisely aligned before further processing. Previous works have reported that the misalignment may negatively influence the reconstruction performance[29]. As is depicted in Fig. 2a, the intermediate events between two frames are overlaid on top of the frame, and the events (red and blue dots) appear randomly scattered throughout the frame, indicating the misalignment of the two modalities. For achieving a pixel-by-pixel events-to-frame spatial alignment, we have developed two methods for alignment and validation.

Considering that the outer contour of the sandclock sample is always static, an events-to-frame spatial alignment pipeline is designed to retrieve

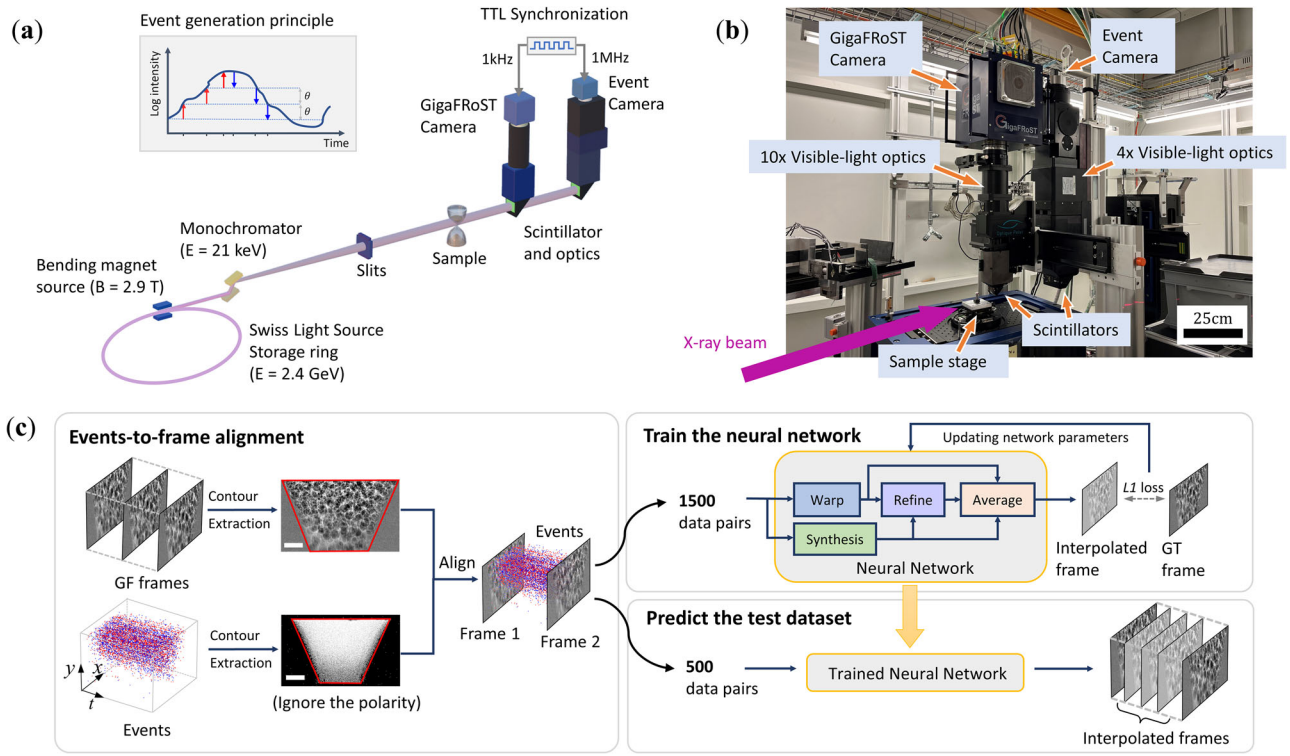

Fig. 1 | Inline dual-camera X-ray imaging platform. a Schematic diagram of the dual-camera experimental setup. Monochromatic X-rays are employed during the calibration of the detectors, while a filtered white beam is applied during the imaging experiment. Before impinging on the sample, slits are used to adjust the beam size; after passing the sample, the transmitted X-ray light is absorbed by two scintillators that transform the X-rays to visible light, which by means of two different optics are focused on the GigaFRoST (GF) and the event camera, respectively. The inset illustrates how the event camera detects changes in brightness: it triggers a positive or a negative event when the per-pixel intensity change exceeds a threshold θ. Red and blue arrows represent positive and negative events, respectively. **b** Image of the accompanying components at the beamline. **c** Flowchart of the data processing pipeline, which first conducts an alignment of GigaFRoST (GF) frames and events to obtain a pixel-by-pixel correspondence of the imaged sample, followed by neural network training and prediction. The scale bar, shown as a white line in the frame (in **c**), represents 200 μm.

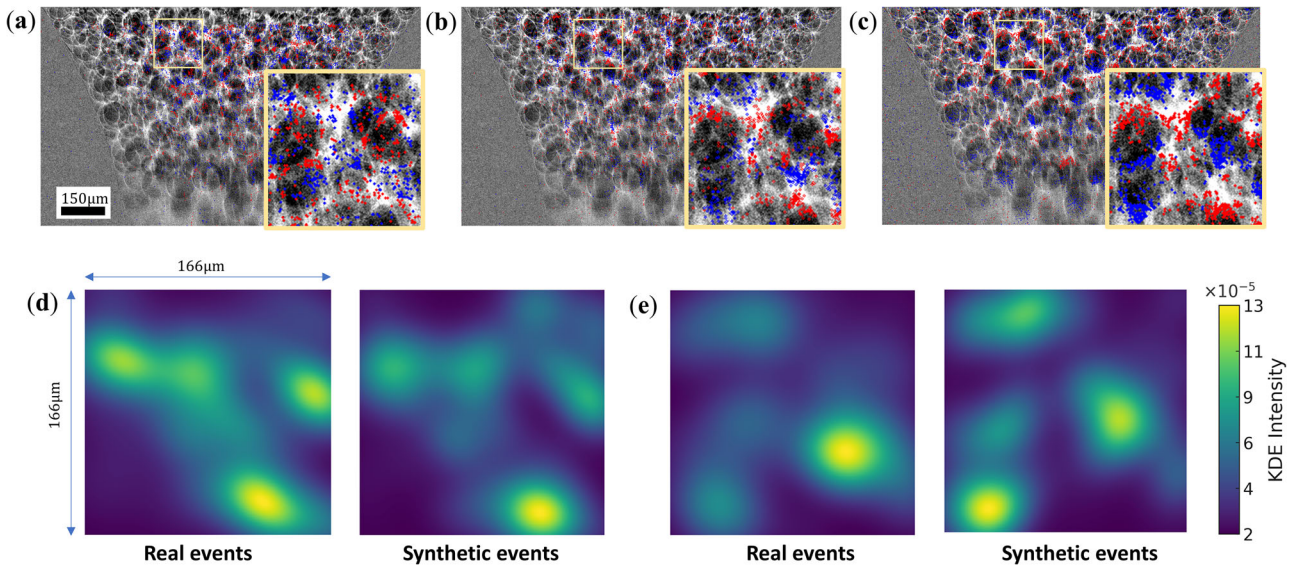

Fig. 2 | Visualization and validation of spatial alignment. a A frame from the GigaFRoST (GF) camera overlaid with "real" events before alignment. The red and blue dots represent positive and negative events, respectively. A local region is selected and magnified by a factor of 3 to better visualize the two modalities; the same zoomed-in area is used in (**b**) and (**c**). **b** The same GF frame with events after the alignment pipeline. **c** The GF frame overlaid with "synthetic" events used as ground truth for comparison. **d** Two-dimensional kernel density estimation (KDE) plots of positive events in the zoomed-in region from (**b**) and (**c**). **e** KDE plots of negative events in the same region.

the elements of an affine image transform matrix (for details, see the Methods section). The overall processing pipeline includes two steps: 1) events stacking and denoising; 2) alignment and refinement. Before the alignment, to reconstruct *quasi* frames, events captured by the event camera were accumulated over a given period, disregarding polarity, which resulted in a grayscale image. We refer to these events captured by the event camera in the experiment as "real events". Subsequently, the stacked *quasi* event frame is denoised and used to match the contour in the GF frame, to obtain the rough alignment parameters. For the further refinement of the alignment parameters, we generated so-called "synthetic events", which originate from the GF frame data. As explained in the Methods section, we produced these using the event emulator from the Prophesee SDK, which detects intensity changes of each pixel to mimic the behavior of a real event camera responding to dynamic scene changes. These synthetic events were used as ground truth for the refinement step.

After running the whole pre-processing and alignment pipeline, we obtained the final scale factor s between the GF and event-based *quasi* frames with $s = 1.1$, and the centroid displacement in the x-axis direction was found to be +11 pixels, while the one in the y-axis direction was -7 pixels. With these, the original coordinates of all "real" events were transformed, for which the visual result is shown in Fig. 2b. It can be seen that the positive events tend to cluster near the white regions in the GF frame, indicating these pixels are going to become brighter, while the negative ones indicate the pixels that are becoming darker.

The second method served the purpose of validating these results as well as devising a metric that is independent of the shape of the imaged sample. As visible in Fig. 2c, the synthetic events exhibited a similar distribution pattern as the aligned real events (Fig. 2b), but for a quantitative comparison, the positive and negative events were further input to calculate kernel density estimates (KDEs). The KDEs of the positive and negative events in these zoomed-in regions of Fig. 2b and c are shown in Fig. 2d and e, demonstrating that clustering centers of real events mostly correspond to their synthetic counterparts, though some centers differ minimally. This difference can primarily be attributed to the complexity of real motion, which makes it challenging to replicate through the simple interpolation used in generating synthetic events[30]. Overall, however, the distribution of the aligned real events could be very well matched with the synthetic events, demonstrating the effectiveness and validity of our alignment approach.

Event-guided temporal interpolation

After spatial alignment, we could analyze the inner dynamics based on the frames and events captured by these two cameras. The sand grains within the sample exhibit different velocities and hence allowed for characterizing the system under different conditions. To integrate the sparse intermediate events into several interpolated frames, both the frame and event datasets were used as input to train a deep-learning neural network, whose architecture follows the *Time Lens*[5] framework. The latter consists of four dedicated modules that enable a complementary interpolation scheme, combining a warping-based and synthesis-based approach. Moreover, it has been previously validated on highly dynamic scenes, such as spinning an umbrella and bursting a balloon[5]. Here, as shown in Fig. 1c, our training dataset contained 1500 data pairs, and then the trained model of the network was used to predict the test dataset with 500 data pairs. To assess the benefit of integrating the event data stream into the interpolation, we further conducted a comparative analysis between the event-guided method and a frame-based interpolation method, represented by the *Super SloMo*[31] algorithm. This one only uses frames as input for reconstructing intermediate frames, as opposed to *Time Lens* that integrates both the frames and events.

The results of this comparison are illustrated in Fig. 3, where the interpolation results for different temporal upsampling factors are shown. Since the GF camera was operated at a constant sampling frequency (1 kHz), the upsampling was achieved by leaving out a given number of frames and utilizing the events to reconstruct these, while using the skipped frames as ground truth data. Here, the middle interpolated frames across different upsampling factors, as well as three different regions from top to bottom are selected and visually compared. Due to the shape of the sandclock, the three magnified regions correspond to different amounts of stacked sand grains and different flowing velocities. In the first row (Fig. 3a), the interpolation frame is shown that has been reconstructed purely from frames; in the second row (Fig. 3b) and Supplementary Movies 2 and 3, event-guided interpolation frames from "real" events are presented; and in the last row (Fig. 3c) the same is shown for "synthetic" ones.

As can be seen for the 2× upsampling, the frame-based interpolation (Fig. 3a) reconstructed most features from the upper region, but the edges in the middle and lower regions were mostly lost, indicating that the method is suitable only for interpolating slowly varying frames. Consequently, the 6×

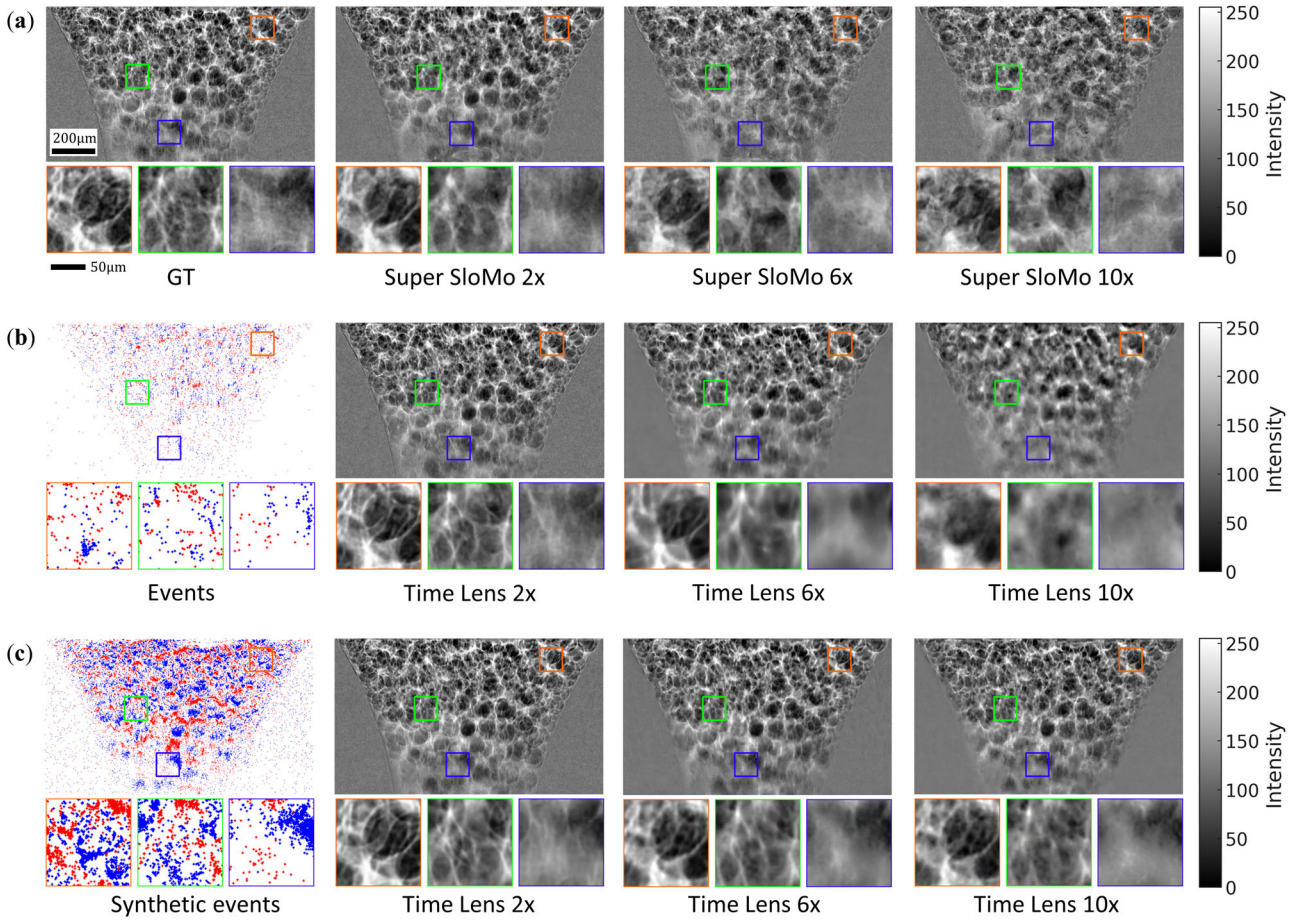

Fig. 3 | Comparison of the interpolated frames. a Reconstructed frames from a pure frame-based (*Super SloMo*) and **b** from the event-guided interpolation method (*Time Lens*), after 2×, 6×, and 10× temporal upsampling, compared to ground-truth data (GT). To illustrate the interpolation performance under different motion dynamics, three vertical regions corresponding to slow, moderate, and fast movement of sand grains are selected and zoomed in by a factor of 3. **c** The synthetic events and the reconstructed interpolated frames. The scale bars shown in the top-left panel apply to all panels.

and 10× upsampling resulted in blurriness and incorrect motion estimation in the interpolated frames, which is evident from the intertwined edges and mis-positioning of black shadows. On the other hand, these lost edges could effectively be retrieved using the event-guided method (Fig. 3b). The motion estimation became more accurate when events were incorporated, leading to the fact that some edge information could be recovered for 6× upsampling. However, for the 10× event-guided upsampling, the interpolation result became very blurry as well, even though the main shadow feature in the upper region was retained.

In the analysis of real events, we observed that the limited number of events posed challenges in achieving higher image quality and upsampling factors, particularly in the lower regions. In the sandclock sample, however, despite faster grain movement in the lower part, fewer events were measured compared to the upper part. With the shape of the sandclock, in the lower part, the nature of the event generation principle with respect to the Beer-Lambert law led to a relatively small event generation rate. Namely, an X-ray image I follows the relation: $I = I_{in}e^{-\sum \mu_i d_i}$, where I_{in} is the fluence of the incident beam, μ is the attenuation coefficient and d_i is the thickness of each grain. Hence, events are triggered by changes in the logarithmic scale, *i.e.*, the brightness difference $\Delta L = \log\left(\frac{I(t_1)}{I(t_0)}\right)$ is proportional to the thickness difference Δd. That is why, larger changes in grain dynamics of the upper layer part resulted in more events and higher contrast. In addition to these effects, we also found that the lower part of our field of view had a 25% lower photon intensity than the upper part (due to a misalignment of the X-ray beam).

To evaluate the impact of an increased event rate on the overall interpolation performance, we generated synthetic events using the event emulator, resulting in six times more synthetic events than in the (real) experimental case. We then investigated the interpolation performance in combination with 2× , 6× , and 10× upsampling. The results are shown in Fig. 3c and it can be seen that the increase in event rate enhanced both the image quality and motion estimation, with only slight blurriness observed, while most features were preserved even up to a 10× upsampling.

Quantitative performance evaluation

The previous demonstration of the upsampling results showed the superiority of the event-guided approach. However, the visual comparison of the event-guided interpolated frames with the ground-truth ones may be challenging due to marginal differences such as blurriness and small structural variances. To quantitatively evaluate the interpolation performance, two metrics, the PSNR and SSIM, were used to measure the similarity of the images in terms of noise levels and structural features. To ensure that the reconstructed frames correspond closely to their respective ground-truth (GT) frames, rather than neighboring frames, particularly at very high upsampling factors, we introduced correlation matrices. These matrices were used to evaluate the PSNR and SSIM scores between different indices of the interpolated frames and GT frames, respectively.

The correlation matrices are illustrated in Fig. 4. For instance, in the case of 10× upsampling with real events, each interpolated frame with index i is compared with each GT frame with index j. Here, the bottom-left to

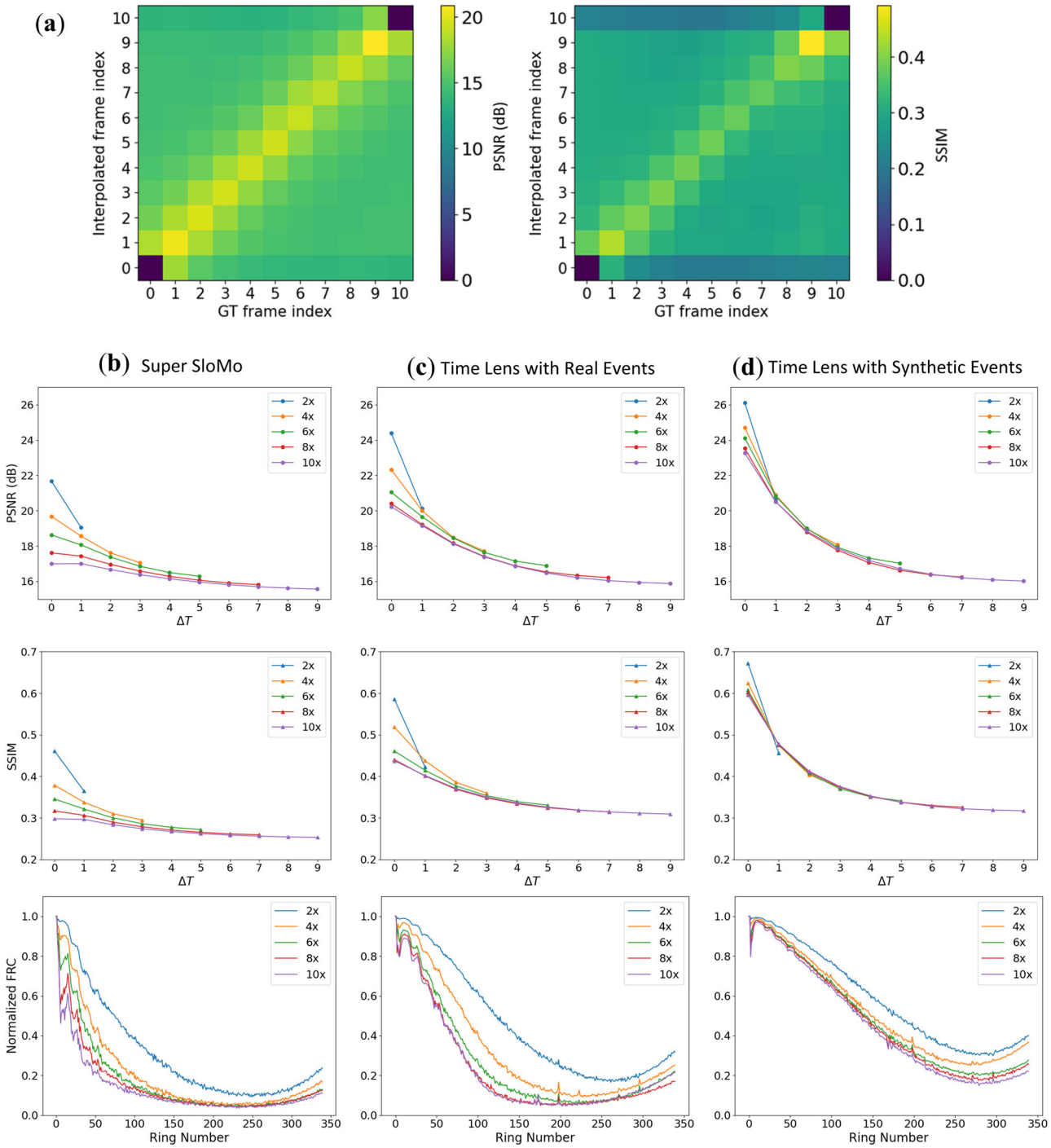

Fig. 4 | Quantitative evaluation of the interpolation performances. a Correlation matrices of the PSNR and SSIM scores between the interpolated frames (indexes 1 to 9) and all the ground-truth (GT) frames, with *Time Lens*. **b** Comparison of interpolation using *Super SloMo* and *Time Lens* with **c** real events and **d** synthetic events, respectively. PSNR and SSIM curves with the frame index difference ΔT when interpolating for 2× , 4× , 6× , 8× and 10× temporal upsampling. Each data point indicates the averaged value with a certain ΔT. Normalized Fourier ring correlation (FRC) curves [Eq. (8)][43] for different ring numbers and upsampling factors, and the interpolated frames are compared with their corresponding GT frames. The FRC value is 1 when the interpolated frames are identical to the GT frames.

top-right diagonal in Fig. 4a shows the SSIM and PSNR scores when $i = j$, respectively. It is visible that the interpolated frames close to the two input boundary frames (0 and 10) exhibit higher PSNR and SSIM values. When the index difference $\Delta T = |i - j|$ between two compared frames is higher, lower PSNR and SSIM values are expected as displayed in Fig. 4a, which demonstrates lower similarity compared with their neighboring frames.

To evaluate the variations in image quality with increased ΔT, and by performing interpolation with only frames as well as incorporating event, the PSNR and SSIM values were calculated under different upsampling factors: 2× , 4× , 6× , 8× , and 10× . The results are presented in Fig. 4b for interpolation using only frames and incorporating both frames and events. It can be seen that all the PSNR and SSIM scores gradually decrease as ΔT increases, indicating that the interpolated frames align with their

corresponding GT frames, confirming the accuracy of the overall motion reconstruction. In the case of frame-only interpolation, errors in motion reconstruction can lead to frames that are barely distinguishable from their neighboring frames, as exemplified by the 10× upsampling. When the event data were integrated (Fig. 4c), the PSNR improved by more than 2 dB and the SSIM increased by over 0.1, indicating that more details are reconstructed. For 10× upsampling, when ΔT increases to 9, the PSNR converges to 16, while the SSIM converges to 0.30. These converged values suggest a comparison between two significantly different states and can be interpreted as an approximate noise level. The difference between these converged values and the values at $\Delta T = 0$ can be used to assess the current interpolation performance.

Finally, the overall image quality of the interpolated frames was evaluated in the frequency domain using the Fourier ring correlation (FRC) metric, which quantifies image correlations across different frequencies. When upsampling with only frames, the FRC scores decrease rapidly, indicating a loss of corresponding frequency components. For the 2× case, using only frames results in an FRC score of 0.2 at a ring number of ~200, whereas incorporating both frames and real events yields a minimum FRC of around 0.3 at a ring number of 250. In the 6× scenario, the FRC converges to 0.1 at a ring number of 150 with only frames, compared to ~0.2 at a ring number of 200 when both data types are used. For 10× upsampling, the FRC quickly converges to 0.1 at an approximate ring number of 100 in both cases, reflecting a significant loss of high-frequency information and corresponding to the observed blurriness at higher upsampling factors. However, in the synthetic event case where six times more events are present (Fig. 4d), most high-frequency details are still preserved and the information loss is not as significant. This performance difference between real and synthetic events can be attributed to the varying numbers of events in each case; the more events generated, the better the motion information can be encoded, allowing higher frequency information to be restored in the interpolation result.

Discussion

In our work, we have demonstrated an event-guided platform for time-resolved synchrotron X-ray imaging based on an inline dual-camera setup. The main advantage of our setup design is its efficient utilization of the X-ray beam, achieved by optimizing both the scintillator thickness and the camera placement. By using a high-magnification objective paired with a thinner scintillator upstream, we maximize photon efficiency and reduce visible-light diffusion, while a thicker scintillator downstream compensates for the reduced photon flux to ensure adequate event detection. However, this configuration inherently reduces photon counts for the event camera, leading to the capture of fewer, noisier events. Future optimization such as using alternative scintillator materials[11,12] or adjusting the scintillator thickness, may further enhance imaging performance under low-light conditions.

With our data processing pipeline, we have shown that events, as a novel modality, can provide complementary motion information to enhance the interpolation process[7,32], offering a straightforward and versatile solution tailored for dynamic X-ray imaging experiment setups. In particular, we have demonstrated promising reconstruction results under 6× temporal upsampling with real events captured during the experiments, enabling us to use events to super-resolve the skipped-frame 167 FPS video to a 1000 FPS counterpart, indicating that these sparse events have the potential to replace the frames. In the experimental results presented above, it shall be highlighted that upsampling factors up to 10× were tested and validated against ground-truth data recorded at 1 kHz. As demonstrated in the quantitative evaluation of interpolation performance for 2× to 10× upsampling (Fig. 4), increasing the upsampling factor beyond 10× would cause the PSNR and SSIM metrics to further approach the noise level.

For the present example, a dense frame with dimensions of 1180 × 680 pixels and 12-bit grayscale depth is being replaced by around 7000 events, each represented by 4 elements at 16 bits per element. The original frame size is 9,614,400 bits, while the event data size is 448,000 bits, comprising only about 4.66% of the original dense frame size. Already this replacement bears the potential of a significant data reduction by ~95%. Nevertheless, we also conclude that by comparing the interpolation results with real and synthetic events, a certain data loss in the form of blurring is noticeable. Hence, for future experiments, it will be necessary to further tune and optimize the parameters of the event camera, such as lowering threshold to guarantee a sufficient event generation rate[30,33], while mitigating the trade-offs with noise levels. For the event camera used in our work, the Prophesee EVK4, a maximum sensor event readout speed of 3 Gigaevents per second is ideally possible, which already allows for a certain increase in event rate and hence a higher sampling rate with better image quality.

Finally, we have also applied our trained network to the original frame data without skipping frames. In doing so, we were able to upsample the original dataset to 2000 FPS, which is shown in Fig. 5a. This demonstrates that even with a limited number of intermediate events (around 7000), it was possible to increase the information content. However, since no ground-truth data was available, we used an optical flow metric based on Gunnar Farneback's algorithm to visualize the velocity field of the sand grains, indicating the reconstruction of some nonlinear motion patterns (as shown in Fig. 5b). Therein, different colors represent varying motion speeds across the image. The optical flow revealed overlapping regions between frames, and the interpolated frames helped separate these regions and made it possible to track smaller changes that occurred between the consecutive frames.

In conclusion, our data processing pipeline enabled the reconstruction of a low-frequency video into a high-frequency counterpart. This can help reduce data storage volume and speed up data readout. Additionally, our method also shows the potential for enhancing temporal resolution without compromising the field of view, paving the way for ultra-fast continuous dynamic imaging applications, e.g., event-guided 3D and 4D X-ray imaging.

Methods

Experimental setup

The experiment was conducted at the X02DA TOMCAT beamline of the Swiss Light Source (Paul Scherrer Institute, Villigen, Switzerland), where the X-ray beam is produced by a 2.9 T superbending magnet from a 2.4 GeV storage ring with 400 mA ring current, in top-up mode. The double-multilayer monochromator (DMM) was first used to generate a monochromatic beam with X-ray energy $E = 21$ keV for calibration of the two cameras (including adjustment of the camera tilts and focusing). For focusing the frame camera, a routine based on a fractal checkerboard pattern was used[34]. For focusing and aligning the event camera, we first created pseudo frames from the event stream. First, events were accumulated over a duration of 100 ms and a grayscale image was created where each pixel represented the total number of recorded events. Then this process was repeated 100 times, resulting in 100 gray value images, which were then all input to calculate the "mean" gray value image. At the same time, a fast X-ray shutter was operated at 2 Hz to artificially create events. This resulting "mean" image resembled a pseudo grayvalue image which was then input through the same focusing routine[34].

During the measurement, the beam was operated in a filtered white beam configuration by utilizing a 5 mm glassy carbon plate (Sigradur) in combination with a 1 mm thick silicon plate. Compared with monochromatic beam, white beam can provide higher flux density, to improve the resolution and signal-to-noise ratio[35].

The proposed experimental setup is shown in Fig. 1. In our experiment, the sample was placed at 28.2 m downstream of the source. Two detectors were used for detection, *i.e.*, a high-speed frame-based GigaFRoST detector (in-house developed, GF), and an event camera (Prophesee, Metavision Evaluation Kit 4, EVK). They were coupled with different scintillators and high numerical aperture microscopes. Specifically, the GF was paired with a 100 μm thick LuAG:Ce scintillator and 10× visible-light optics, and the EVK was coupled with a 150 μm thick LuAG:Ce scintillator and a 4× visible-light optics. Then the GF and the EVK achieve similar effective pixel sizes of 1.1 μm and 1.2 μm, respectively.

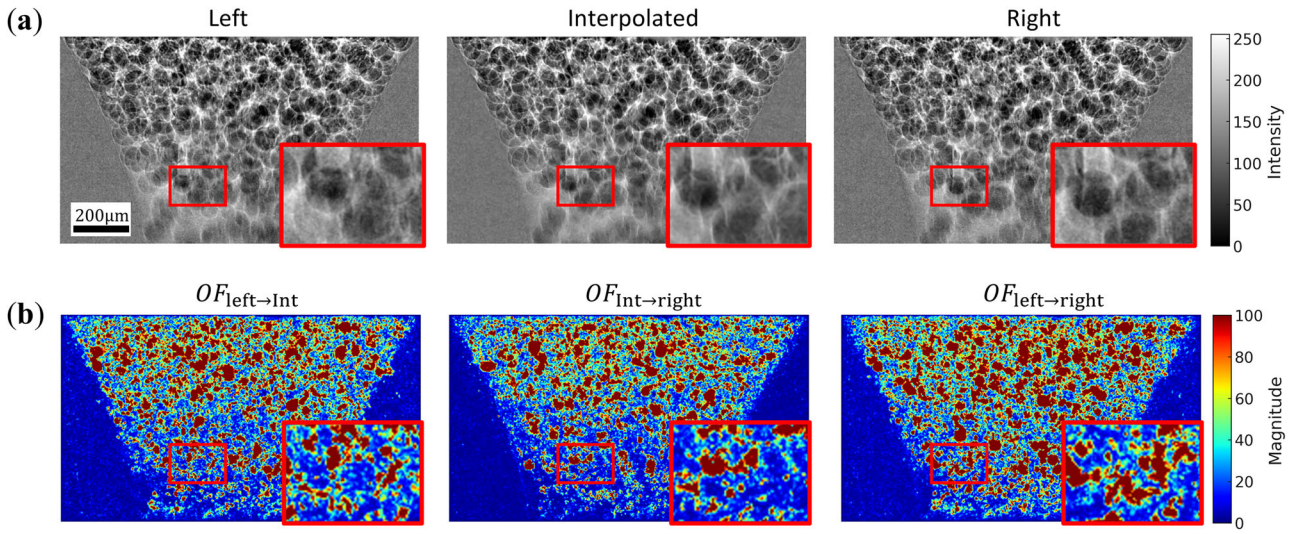

Fig. 5 | Interpolation results to achieve the frame rate of 2000 PFS and the related optical fields. a One middle frame is interpolated between two frames from the original video frames sampled at 1 kHz, the nonlinear motion can be tracked, as shown in the zoomed-in region. **b** The optical flow (*OF*) to visualize the magnitude of the speed field, between the left to right frames $OF_{left\rightarrow right}$, left to interpolated frames $OF_{left\rightarrow int}$, and interpolated to right frames $OF_{int\rightarrow right}$. All of these images have dimensions of 1160 × 680 pixels and a size of 1.1 μm. The scale bar shown in the top-left panel applies to all panels.

Table 1 | Specifications of the system with GigaFRoST camera (GF) and event camera (Metavision Evaluation Kit 4, EVK)

Camera	Scintillator	Optics	Pixel size	Sampling rate	To-sample distance
GF	100 μm LuAG:Ce	×10	11 μm	1 kHz	5.8 cm
EVK	150 μm LuAG:Ce	×4	4.86 μm	1 MHz	35 cm

A position and acquisition box (PandABox) was employed to generate the trigger signals to synchronize these two cameras[36]. The frame-based camera operates at a frame sampling frequency of 1 kHz with an exposure time of 0.99 ms, where the remaining 0.01 ms is allocated for readout. Finally, the event camera operates at a sampling frequency of 1 MHz which produces a relatively continuous and smooth event stream. These specifications are summarized in Table 1, and the corresponding experimental recording is available in Supplementary Movie 1.

Principle of event generation

Different from conventional frame-based cameras, event cameras detect the brightness changes at each pixel asynchronously[37]. The principle is illustrated in the schematic diagram in Fig. 1a, where the brightness of a pixel changes over time, generating three positive events (red upward arrows) and three negative events (blue downward arrows), respectively. For a pixel at a spatial location (x_k, y_k), the formula of the event triggering can be formulated as follows:

$$\Delta L = L(x_k, y_k, t_k) - L(x_k, y_k, t_k - \Delta t_k) = p_k\theta, \quad (1)$$

where L is the logarithmic brightness, x_k, y_k are the detector's coordinates, t_k is a particular timestamp and Δt_k is the time difference between which the brightness change occurs for firing an event. Once the absolute change of ΔL is greater than a set threshold θ, an event with the polarity p_k will be triggered, and they are recorded as a tuple $e_k = (x_k, y_k, t_k, p_k)$. Specifically, p_k is $+1$ if the brightness increases and p_k equals -1 if the brightness decreases. In our work, the event camera was operated at a sampling frequency of 1 MHz, allowing Δt ideally as short as 1 μs. The events were recorded as a set of e_k as $E = \{e_k | k = 1, \ldots, N\}$, and the average count N between two frames captured at 1 ms was about 7000 events.

To study the interpolation performance with different event counts, synthetic events can be simulated in silico. In our work, we utilized the emulator from the Metavision SDK package by Prophesee to mimic the characteristics of the event camera used in our experiments. During the simulation, the original video frames are first temporally resolved using the deep learning-based interpolation method, *Super SloMo*[31], to obtain temporally resolved frames. Subsequently, brightness changes are calculated using Eq. (1) with a manually set triggering threshold, and synthetic events are generated by comparing the brightness changes to this threshold. To simulate the event camera more accurately, stochastic behaviors of electronic devices, such as latency and noise, are further incorporated[30]. By tuning the threshold, we generated synthetic events with an average count approximately six times higher than our recorded real events from the experiment, ~42,000 events/ms.

Alignment between frames and events

Alignment of the two datasets (frames and events) is the critical step before our further interpolation, and it needs to be tailored specifically for each system[5,6,38]. Before the experiment, the two detectors with slightly different pixel sizes were calibrated separately, which led to misalignment in the horizontal and vertical directions (centroid displacements), and the slightly different effective pixel sizes led to the scale difference. To align the sparse events to the frames, an alignment pipeline is proposed as outlined in Fig. 6.

The events from timestamp t_0 to t_1 are firstly stacked to a two-dimensional event count image $E(x, y) = \int_{t_0}^{t_1} |e_t| dt$. Since the events are inherently noisy, a significant amount of noise accumulates in the background, shown as sparsely distributed low-density regions. In contrast, the targeted events caused by the sand flowing inside the sandclock form high-density regions. Then we utilize the Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[39] algorithm to eliminate the background noise and obtain the denoised image $E'(x, y)$ with higher contrast.

Subsequently, the GF image and $E'(x, y)$ are input for contour matching. The OpenCV library is applied to extract the contour, find the centroids and calculate the mapping relationships, yielding the centroid displacements Δx, Δy and the scale factor s. To refine the factors, as demonstrated in step 2 of Fig. 6, the set of synthetic events is used as the

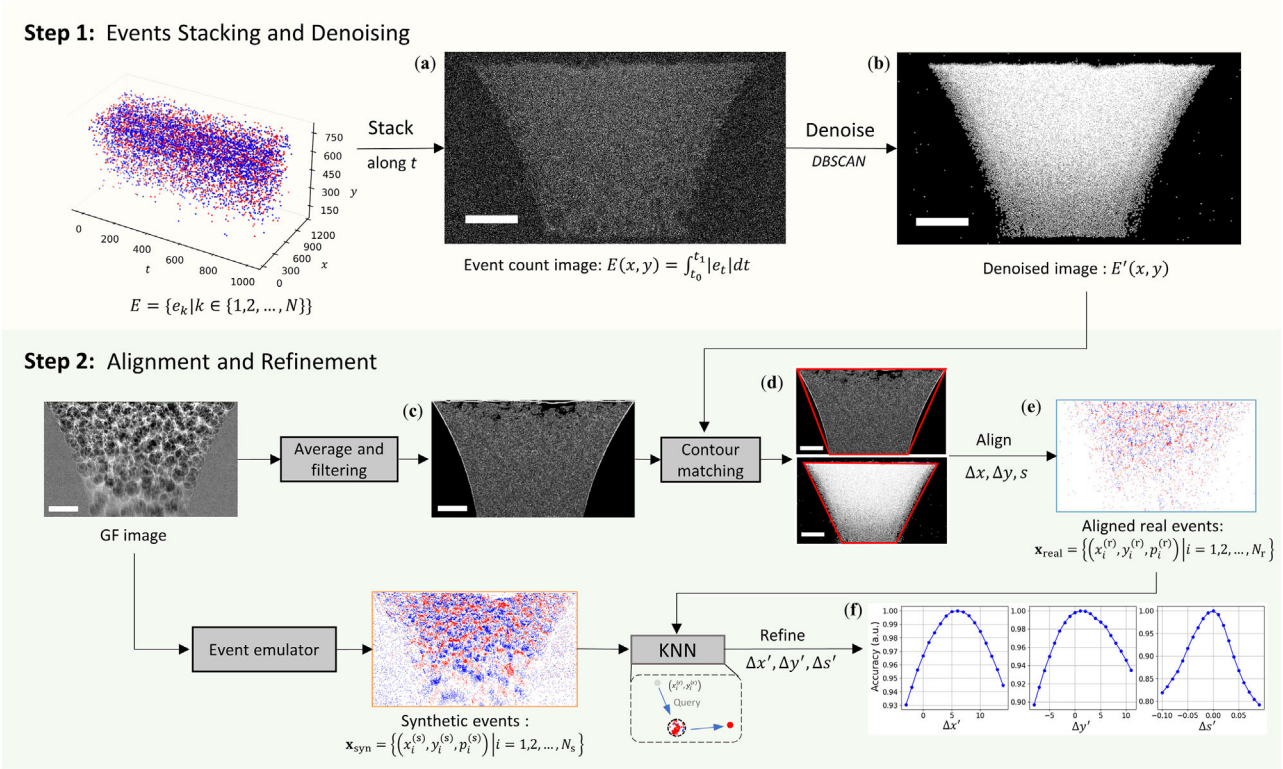

Fig. 6 | The proposed events-to-frame alignment pipeline. a The sparse events E are stacked into a 2-D frame $E(x, y)$, and **b** denoised to $E'(x, y)$, which is used for contour matching with the GF image. For the GF images, **c** several images are averaged together to remove the sand grain features, and a filter is used to enhance the contour. After **d** contour matching with $E'(x, y)$, the centroid displacements $\Delta x = +5$, $\Delta y = -8$ and the scale factor $s = 1.1$ are retrieved, and applied to (**e**) spatially transform the (raw) real event data. To refine these alignment parameters, synthetic events are generated with an event emulator, and they serve as the GT to tune the parameters using the K-Nearest Neighbors (KNN) algorithm. The values in (**f**) refinement curves indicate the impact on the accuracy when the shifts, $\Delta x' = +6$, $\Delta y' = +1$, $\Delta s' = 0$, are applied. All of these images have dimensions of 1160×680 pixels and a size of 1.1 μm. The scale bars, shown as white lines in the frames, represent 200 μm.

labeled dataset $\mathbf{x}_{\text{syn}} = \{(x_i^{(s)}, y_i^{(s)}, p_i^{(s)})|i = 1, \dots, N_s\}$, where the pixel locations serve as features and the polarity values as labels.

The K-Nearest Neighbors (KNN) algorithm is employed to predict the label $p_i^{(r)}$ of each real event from $\mathbf{x}_{\text{real}}$ at location $(x_i^{(r)}, y_i^{(r)})$ based on the majority vote among the K nearest neighbors. Specifically, the location $(x_i^{(r)}, y_i^{(r)})$ of one real event serves as a query for the synthetic events (dataset with labels). The labels of the nearest K synthetic events around the location $(x_i^{(r)}, y_i^{(r)})$ are collected, and the most common label corresponds to the predicted label.

The predicted labels are compared with the labels from real events in order to calculate the prediction accuracy. An optimization loop of the overall prediction accuracy is conducted, to search for the best parameters $\Delta x'$, $\Delta y'$, $\Delta s'$ for coordinate transformation and scaling, which is formulated as

$$(\Delta x'^*, \Delta y'^*, \Delta s'^*) = \text{argmax}_{\Delta x', \Delta y', \Delta s'} \text{KNN}\,(\mathbf{x}_{\text{syn}}, \mathbf{x}_{\text{real}}(\Delta x', \Delta y', \Delta s'), K), \tag{2}$$

where the amount of selected neighbors K is set to 50, and KNN(·) computes the prediction accuracy with the current parameters. The best parameters' combination $(\Delta x'^*, \Delta y'^*, \Delta s'^*)$ after the optimization are then used to refine the alignment parameters.

Interpolation methods

To reconstruct the frames, an analytical filter is proposed to directly integrate the high-frequency information of events into intensity frames[40]. However, this analytical approach was not suitable for our work due to the sparsity of events, which introduced unexpected noises and posed challenges in generating sharp images. Therefore, more effective deep learning-based methods are preferred for frame reconstruction.

Some deep-learning neural networks have been proposed to synthesize or warp the boundary frames and estimate intermediate frames with only frames[41]. For example, *Super SloMo* calculates the bi-directional optical flow between two input boundary frames[31]. These optical flows are employed to warp the boundary frames, followed by a linear fusion process to generate the intermediate frame. Nonetheless, due to the inherent difficulty in recovering the missing motion information between two frames, video frame interpolation remains a challenging and ill-posed problem. Therefore, guiding the interpolation of the intermediate frames with events is becoming a non-trivial task.

Recently, several neural networks have been designed to integrate the frames and events to interpolate the frames[5–7], where *Time Lens* neural network stands as a robust baseline method[5]. *Time Lens* combines synthesis-based and warping-based networks to estimate the non-linear intermediate motion between two frames, and the same network architecture is utilized for interpolation in our work. The overall data processing pipeline is shown in Fig. 1c. The events are firstly aligned with the left and boundary frames before inputting into the neural network. During the training or testing of the neural network, the frames and events are processed by the four modules progressively, i.e., *warp*, *refine*, *synthesis* and *average* modules, to obtain the output interpolated frames, and the network's input and output are defined as

$$I_t = \text{Net}(t, I_1, I_2, E_{1\to 2}, \text{Params}), \tag{3}$$

where the input to the network includes the timestamp of interpolated frame t, the left frame I_1, the right frame I_2, and their intermediate events $E_{1\to 2}$, to

obtain the output interpolated frame I_t. To optimize the parameters of the network, the L_1 loss function is calculated with

$$L1 = \| I_{\text{int}} - I_{\text{gt}} \|, \tag{4}$$

to evaluate the difference between the interpolated frame I_{int} and the GT frame I_{gt}. The error is then back-propagated to update the parameters of the network.

The overall dataset was split into two parts: 1500 paired GF frames and their corresponding intermediate events were used as the training dataset, while 500 pairs were used as the testing dataset. Even though there are available pre-trained models on natural scenes for both *Time Lens* and *Super SloMo*, they were not directly applicable, most likely due to different acquisition settings, which results in variations in noise level and dynamic range and hinders model generalization across different datasets. Therefore, we further trained (*i.e.*, fine-tuned) the pre-trained models on our dataset to adjust their parameters to better fit the specific features of our dataset. During the training, for *Time Lens* with real and synthetic events, the pre-trained model was fine-tuned using an L1 loss over 50 epochs with a learning rate of 10^{-4} until convergence. For *Super SloMo*, the fine-tuning with the L1 loss did not converge on our dataset, so we progressively fine-tuned the model using the original loss for 200 epochs at the same learning rate of 10^{-4}. Detailed training information is provided in Supplementary Note 2.

Metrics

To meet the goal of data reduction, in our setting, we skip several frames between two frames during the experiments, and replace the skipped frames with the intermediate sparse events to reconstruct the skipped frames. The skipped frames are used as GT for evaluating the quality of interpolated frames on different metrics in terms of structural similarity, signal-to-noise ratio, and frequency.

The Structural Similarity Index Measure (SSIM) evaluates the similarity between two images in terms of luminance, contrast, and structural features, and is computed according to

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{5}$$

where x and y represent two compared images, μ_x and μ_y denote their averaged values, σ_x and σ_y are the standard deviations, σ_{xy} is the covariance of x and y, C_1 and C_2 are small constants to stabilize the division operations.

The PSNR score provides a quantitative assessment of image fidelity, and a higher score indicates higher signal and noise similarity. Between the image x and image y, the score is calculated as

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right), \tag{6}$$

where both images have the same size $m \times n$, and it measures the ratio between the maximum possible pixel value of the image (MAX, *e.g.*, 255 for an 8-bit image) and the Mean Squared Error (MSE) between the image x and image y. The MSE is calculated according to the following formula:

$$\text{MSE} = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} \| x(i,j) - y(i,j) \|^2. \tag{7}$$

The Fourier Ring Correlation (FRC) measures the correlation level between the spatial frequencies of two images[42,43]. First, the images are transformed using Fourier transformation to produce frequency images. With the center of each frequency image set as the origin, a sequence of concentric rings $r \in R$ with progressively larger radii, corresponding to higher spatial frequencies, are chosen to compute the correlations. The FRC is then computed for each ring, utilizing its normalized form as defined by:

$$\text{FRC}(R) = \frac{\sum_{r\in R} F_x(r) \cdot F_y(r)^*}{\sqrt{\sum_{r\in R}|F_x(r)|^2 \cdot \sum_{r\in R}|F_y(r)|^2}}, \tag{8}$$

where $F_x(r)$ and $F_y(r)$ are the Fourier components of two images, whose magnitude r equals the radius R. In our work, the size of the frames is 1180×680 pixels, and we set a bin width of 2 for the calculation of the FRC. As the number of rings increases from 1 to 340, higher frequencies in the image are incorporated for analysis.

To quantitatively evaluate the interpolation performance, for $n\times$ upsampling, we introduce $(n+1)\times(n+1)$ correlation matrices to evaluate the PSNR and SSIM scores when comparing the interpolated frames with different GT frames. Here, we denote the absolute index difference between one interpolated frame and one frame from the GT dataset as $\Delta T = |\text{Index}_{\text{int}} - \text{Index}_{\text{gt}}|$, *e.g.*, $\Delta T = 0$ indicates that the interpolated frame is compared with the GT frame with the same index, when the interpolated frame 1 is compared with the GT frames with indexes 3 and 10, it corresponds to $\Delta T = 2$ and $\Delta T = 9$ respectively.

Data availability

The dataset in the current study are available from the corresponding authors upon request.

Code availability

The code developed to analyze the results is available from the corresponding author upon reasonable request.

Received: 3 December 2024; Accepted: 15 May 2025;
Published online: 28 May 2025

References

1. Gallego, G. et al. Event-based vision: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 154–180 (2020).
2. Delbrück, T., Linares-Barranco, B., Culurciello, E. & Posch, C. Activity-driven, event-based vision sensors. In *Proc. IEEE International Symposium on Circuits and Systems*, 2426–2429 (2010).
3. Bardow, P., Davison, A. J. & Leutenegger, S. Simultaneous optical flow and intensity estimation from an event camera. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 884–892 (2016).
4. Gallego, G. & Scaramuzza, D. Accurate angular velocity estimation with an event camera. *IEEE Robot. Autom. Lett.* **2**, 632–639 (2017).
5. Tulyakov, S. et al. Time lens: event-based video frame interpolation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16155–16164 (2021).
6. Tulyakov, S. et al. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17755–17764 (2022).
7. Kim, T., Chae, Y., Jang, H.-K. & Yoon, K.-J. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18032–18042 (2023).
8. Cabriel, C., Monfort, T., Specht, C. G. & Izeddin, I. Event-based vision sensor for fast and dense single-molecule localization microscopy. *Nat. Photonics* **17**, 1–9 (2023).
9. Ryan, C. et al. Real-time multi-task facial analytics with event cameras. *IEEE Access* **11**, 76964–76976 (2023).
10. Cohen, G. et al. Event-based sensing for space situational awareness. *J. Astronaut. Sci.* **66**, 125–141 (2019).
11. Sleator, C., Christophersen, M., Cheung, C., Qadri, S. N. & Santiago, F. X-ray tracking using a perovskite scintillator with an event-based

sensor. In *Hard X-Ray, Gamma-Ray, and Neutron Detector Physics XXV*, 104–115 (2023).
12. Zhang, A. et al. Event-based X-ray imager with ghosting-free scintillator film. *Optica* **11**, 606–611 (2024).
13. Hemberg, O., Otendal, M. & Hertz, H. M. Liquid-metal-jet anode electron-impact X-ray source. *Appl. Phys. Lett.* **83**, 1483–1485 (2003).
14. Cole, J. M. et al. High-resolution μCT of a mouse embryo using a compact laser-driven X-ray betatron source. *Proc. Natl Acad. Sci.* **115**, 6335–6340 (2018).
15. Willmott, P. *An Introduction to Synchrotron Radiation: Techniques and Applications* (John Wiley & Sons, 2019).
16. Sedigh Rahimabadi, P., Khodaei, M. & Koswattage, K. R. Review on applications of synchrotron-based X-ray techniques in materials characterization. *X-Ray Spectrom.* **49**, 348–373 (2020).
17. Zhang, K. et al. Pore evolution mechanisms during directed energy deposition additive manufacturing. *Nat. Commun.* **15**, 1715 (2024).
18. García-Moreno, F. et al. Using x-ray tomoscopy to explore the dynamics of foaming metal. *Nat. Commun.* **10**, 3762 (2019).
19. Schmeltz, M. et al. The human middle ear in motion: 3d visualization and quantification using dynamic synchrotron-based X-ray imaging. *Commun. Biol.* **7**, 157 (2024).
20. Porter, J. L., Looker, Q. & Claus, L. Hybrid cmos detectors for high-speed x-ray imaging. *Rev. Sci. Instrum.* **94**, 061101 (2023).
21. García-Moreno, F. et al. Tomoscopy: time-resolved tomography for dynamic processes in materials. *Adv. Mater.* **33**, 2104659 (2021).
22. Rack, A., García-Moreno, F., Baumbach, T. & Banhart, J. Synchrotron-based radioscopy employing spatio-temporal micro-resolution for studying fast phenomena in liquid metal foams. *J. Synchrotron Radiat.* **16**, 432–434 (2009).
23. Rack, A. et al. Recent developments in mhz radioscopy: towards the ultimate temporal resolution using storage ring-based light sources. *Nucl. Instrum. Methods Phys. Res. Sect. A: Accel. Spectrom. Detect. Assoc. Equip.* **1058**, 168812 (2024).
24. Olbinado, M. P. et al. Mhz frame rate hard x-ray phase-contrast imaging using synchrotron radiation. *Opt. Express* **25**, 13857–13871 (2017).
25. Mokso, R. et al. Gigafrost: the gigabit fast readout system for tomography. *J. Synchrotron Radiat.* **24**, 1250–1259 (2017).
26. Leonarski, F. et al. Jungfraujoch: hardware-accelerated data-acquisition system for kilohertz pixel-array X-ray detectors. *J. Synchrotron Radiat.* **30**, 227–234 (2023).
27. Wang, C., Steiner, U. & Sepe, A. Synchrotron big data science. *Small* **14**, 1802291 (2018).
28. Spindler, S. et al. Simulation framework for x-ray grating interferometry optimization. *Opt. Express* **33**, 1345–1358 (2025).
29. Cho, H., Jeong, Y., Kim, T. & Yoon, K.-J. Non-coaxial event-guided motion deblurring with spatial alignment. In *Proc. IEEE/CVF International Conference on Computer Vision*, 12492–12503 (2023).
30. Hu, Y., Liu, S.-C. & Delbruck, T. v2e: From video frames to realistic dvs events. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1312–1321 (2021).
31. Jiang, H. et al. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9000–9008 (2018).
32. Chen, J. et al. Revisiting event-based video frame interpolation. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1292–1299 (2023).
33. Afshar, S. et al. Event-based feature extraction using adaptive selection thresholds. *Sensors* **20**, 1600 (2020).
34. Shi, Z. et al. Fabrication of a fractal pattern device for focus characterizations of x-ray imaging systems by si deep reactive ion etching and bottom-up au electroplating. *Appl. Opt.* **61**, 3850–3854 (2022).
35. Agrawal, A. K., Singh, B., Singhai, P., Kashyap, Y. & Shukla, M. The white beam station at imaging beamline bl-4, indus-2. *J. Synchrotron Radiat.* **28**, 1639–1648 (2021).
36. Zhang, S. et al. Pandabox: A multipurpose platform for multi-technique scanning and feedback applications. In *Proc. 16th Int. Conf. on Accelerator and Large Experimental Physics Control Systems*, 143–150 (2017).
37. Lichtsteiner, P., Posch, C. & Delbruck, T. A 128 × 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE J. Solid-State Circuits* **43**, 566–576 (2008).
38. Gao, Y., Li, S., Li, Y., Guo, Y. & Dai, Q. Superfast: 200× video frame interpolation via event camera. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 7764–7780 (2023).
39. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Trans. Database Syst.* **42**, 1–21 (2017).
40. Scheerlinck, C., Barnes, N. & Mahony, R. Continuous-time intensity estimation using event cameras. In *Proc. Asian Conference on Computer Vision*, 308–324 (2018).
41. Dong, J., Ota, K. & Dong, M. Video frame interpolation: a comprehensive survey. *ACM Trans. Multimed. Comput., Commun. Appl.* **19**, 1–31 (2023).
42. Banterle, N., Bui, K. H., Lemke, E. A. & Beck, M. Fourier ring correlation as a resolution criterion for super-resolution microscopy. *J. Struct. Biol.* **183**, 363–367 (2013).
43. Koho, S. et al. Fourier ring correlation simplifies image restoration in fluorescence microscopy. *Nat. Commun.* **10**, 3103 (2019).

Acknowledgements

We would like to thank P. Zuppiger (PSI) for his great support in setting up the experiment. We further acknowledge F. Marone (PSI) for help with the high-NA microscopes, L. Azzalini (ESTEC) for support in data acquisition as well as J. Grover (ESTEC) and C. Hempel (PSI) for fruitful discussions. H.W. acknowledges the China Scholarship Council. We acknowledge the Paul Scherrer Institut, Villigen, Switzerland for the provision of synchrotron radiation beamtime at the TOMCAT beamline X02DA of the SLS. The study was supported by the Swiss National Science Foundation (SNF), grant no. 200021_219704.

Author contributions

G.L., A.H., and E.B. conceived the project and performed the experiments together with C.M.S. H.W. was in charge of the data analysis, developed all the methods and wrote the first version of the manuscript. M.S. provided additional funding for the project and contributed to the discussions. All authors contributed to and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42005-025-02142-w.

Correspondence and requests for materials should be addressed to Goran Lovric.

Peer review information *Communications Physics* thanks Guangda Niu, David Pennicard and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025