

Overcoming Photon and Spatiotemporal Sparsity in Fluorescence Lifetime Imaging with SparseFLIM

Corresponding Author: Professor Liwei Liu

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Shen et al, SparseFLIM: Overcoming Photon and Spatiotemporal Sparsity in Fluorescence Lifetime Imaging

This manuscript describes a novel network that presumably estimates multi-component lifetimes from sparse photon data much better than others. This is important because it would speed up data collection at least 10-fold and prevent unnecessary photon-damage, bleaching (and, when applied to life-cell imaging) phototoxicity. The authors apply the techniques to a single FLIM modality (2-photon TCSPC) for which it appears to work fine, and apply it to tissue slices with complex decay characteristics. Their network 'replenishes' photons that were missing in the sparse data. This topic is very specialist but if it would indeed present a 10-fold relaxation on the requirement to collect photons, compared to current state-of-the-art, it would still be extremely interesting for Commsbio's rather broad readership.

First some caveats.

I accepted to review this paper on the base of decades of experience in FLIM, lifetime fitting and manners to present lifetime data, but I only have a few years of limited experience with deep learning networks. I cannot judge whether the backward and forward propagation in the network is absolutely groundbreaking, or whether it is just state of the art in other applications (video-super resolution is mentioned), nor can I judge whether the implementation was optimal, or even correct. In part, I secured some help from a trained physicist (image analyzer) and the review reflects our combined remarks.

Second, after putting in over a full working day in this manuscript, I still am not sure I fully grasp all details and am left with remaining questions. So perhaps this should not be considered a full review.

Review.

1) The manuscript is a little hard to read because of the way it is presented. Data on image acquisition, for example, are in part in the methods section, in part in the main text. Terminology is a bit different from standard, for example, the use of 't' on the x-axis instead of the more common terminology ('time bin') and terms like '3D time sequence' which suggested that multiple full FLIM images were collected in a time series (common in life-cell imaging), and details in how the findings are presented. Probably this could be remedied by reformatting and rewriting. I think in particular the first paragraph should be shaped in such a way that there can be no misunderstanding on what exactly is the idea of the algorithm. Also, see 3) for that.

2) The claimed beneficiary effects on photobleaching and bleaching, I would think, are related to the authors choice to restrict themselves to 2-photon imaging because TCSPC by itself is already extremely photon-efficient (compared to gated FLIM or frequency-domain FLIM). In single-photon TCSPC imaging, we routinely get thousands of good FLIM images with just 10 microWatt of excitation power from a single FOV of living cells, and it would be very interesting to test the network on such data as well.

3) To me there are still unclear details on how exactly the procedure works, and I think the manuscript (the main paper) should be written in such a manner that interested non-expert readers can understand the basic principles. For example, Fig. 1h is supposed to reflect the data from a single pixel with a hundred photons. The data are bin counts, and I see at least 20 different Y amplitudes (discrete values of Y coordinate of the circles). If so, than that pixel has WAY more than 100 photons, right? Why is the y-axis in A.U. and not in photon counts? (compare e.g. Fig S7c where only max 5 photons are present per bin).

What exactly happens with the sparseFLIM restoration process? The algorithm generates extra photons to fill the gaps? Or the discrete numbers of photons are replaced with averaged non-discrete numbers to approximate a more smooth curve,

and subsequently rescaled to reflect an increased nr of photons?

Why, if we focus on fig. 1g, is the average lifetime in the large red part of the image (in the centre, above the cross) quite uniformly red (0.4ns) for Photon-rich and for Enriched cases, while in the sparse image it is orange (0.8 ns) without even a single pixel that is red? If it is just a matter of noise improvements, I would expect some very red pixels in the top image. I do not understand the L-shape of this data cloud.

How can it be that in suppl. fig 5a there appears to be NO correlation between the lifetimes calculated from sparse photon and from sufficient photon data. This just implies that the LMA algorithm does an exceptional bad job, so why using it as a standard against which to benchmark improvements? I do not understand the L-shape of this data cloud. Similarly, why is there NO correlation between reference data and sparse data in Fig. 2h?

It is not intuitively clear why photon refill would result in significantly more signal restoration for ~1 ns lifetimes (over 12-fold) than for 3-4 ns lifetimes (just 4 to 5-fold; Fig. 2 i, j). Presented in this way, it looks like short lifetimes are under-represented in LMA fits of sparse data since the sparse FLIM algorithm injects more photons at ~1 ns bins than at ~3 ns bins. If so, then why are lifetimes longer in the sparse data (fig. 1g, top middle picture)? Yet, the close agreement between sparseFLIM and photon-rich data shows the network is right, so this is just something that puzzles readers.

4) Main reservations I have concern the comparison against benchmarks. At the moment, the improvement in fitting is compared against the LMA fit of sparse data (100 photons in 100 bins). LMA is probably one of the worst choices for noisy exponential fitting, and according to literature it favors short lifetimes under these conditions. Everyone would agree that the fitted red line in Fig. S7g, right panel, is just an exceptional poor fit. LMA assumes that each point has symmetric (Gaussian) variance around the expectation value, and for low photon counts the Poisson distribution significantly deviates from a Gaussian.

Later in the paper additional benchmarks are introduced which do not differ so extremely from the sparseFLIM method. Why was the LMA selected rather than e.g. a bayesian method, e.g. maximum likelihood algorithm, which performs much better in sparse data? Or better even, against the best published result until this manuscript?

5) How easy and valid is it to transfer this learned network to data from completely unrelated preparations (non-tissue)? Does application to e.g. multicomponent analysis of fluorescent proteins in cell biology require a complete new training?

6) In suppl. Fig 5, why are the dots color-coded so that their color depicts the distance from the diagonal (ΔT)? this information is already in their position on the x and y axis. Rather, use the color to reflect the relative abundance of the number of times that x,y pair occurs in the image. 1 dot = 1 pixel, right? (cannot find that detail). That would be much more valuable to judge how important some clouds of dots are, compared to the whole image (similar to the representation of FACS plots).

7) on page 21, limitations of the studies are discussed, but it appears to be more on future directions. The paper also needs a real discussion on possible limitations, e.g.:

- How well does it perform on data from different setups? Is it only applicable on your custom two-photon FLIM setup? What if the system has fewer time bins (and thus, higher counts per bin).
- Does it always assume two single-exponential components? What happens if this is not the case? What is the rationale for sticking with two lifetimes?

8) page 7, the Rsquared value was used for estimation of goodness of fit. For non-linear models the Chi-squared value is more appropriate than the R-squared value. R-squared is actually not the goodness-of-fit, but the coefficient of determination.

A high R-squared does not mean that the fit is good, but instead means that the variation is explained by the fit model. Also, R-squared is not valid for non-linear models, to my understanding.

Minor:

page 10, "The Pearson correlation comparison the network reconstruction results and the sparse inputs also exhibit a substantial enhancement, with a...." Insert a comma after comparison, to reflect an enumeration of arguments.

page 16. " We assessed the feasibility of recovering time sparsity by removing multiple time frames (equivalent to reducing the 256 time channels)." We thought there were just 100 time channels?

on page 24, " This facilitates handling misalignments and information..." I can't find out what 'misalignments' means and don't see it introduced in the paper.

In summary, a very difficult paper to fully grasp, even after multiple times reading it. I think every effort should be made to make it more transparent.

(Remarks to the Author)

The manuscript introduces SparseFLIM, a deep learning model claiming to address issues related to low photon counts and spatio-temporal sparsity in fluorescence lifetime imaging (FLIM). While the potential of such deep learning approaches in overcoming challenges in FLIM lifetime estimation is acknowledged, however, the manuscript requires significant technical improvements.

1. The manuscript asserts to handle slow photon counting routines in FLIM, yet fails to provide details regarding the speed gain achieved through this Deep Learning method. Additionally, the authors omitted crucial details regarding sparse photon and sufficient photon acquisition e.g., changes in system or camera parameters, illumination power, etc. Clarification is needed regarding whether the photon count mentioned in line 89 represents cumulative photons collected or photons at a specific frame. Furthermore, details regarding the laser repetition rate are necessary to ascertain if the 5ns data range captures truncated or whole decay.

1. Line 89 states that a 5ns temporal range is sufficient for autofluorescence lifetime estimation, but the basis for this claim is not provided. Moreover, it is unclear how this model can be generalized for other exogenous fluorescence dye lifetimes. The assertion that working within a 5ns range reduces non-semantic information learning repetition requires elaboration.

2. The manuscript mentions bidirectional information flow in the model, utilizing features from both the current and previous frames. However, it lacks clarity on how temporal decay information with long-term dependencies at each pixel is preserved. Typically, temporal dependencies are maintained through recurrent neural networks, yet evidence of long-term temporal dependency preservation in this model is absent.

3. Terms such as "high-level features," "high-quality features," and "reduce cumulative errors" are used ambiguously in lines 124-126. It remains unclear which features the authors aim to capture and whether cumulative errors refer to spatial or temporal reconstruction and what metric is used to compare that.

4. Figure 1i,j displays good non-linear least square fitting with higher photon counts, yet when both normalized graphs would be plotted together, they clearly exhibit different decay rates. Contrary to the claim that both decays provide the same mean lifetime of 1 ns, both decay curves should be plotted in one graph, and metrics should be provided to compare the mismatch in temporal decay.

5. Line 167 suggests that SparseFLIM excels at restoring intrinsic correlation, but evidence supporting this claim is lacking.

6. In Figure 2, the network output fluorescence lifetime map appears significantly filtered (smooth) compared to the ground truth. Utilizing better statistical metrics (for all figures) such as SSIM and t-SNE could facilitate a more comprehensive comparison of feature extraction.

7. The manuscript poorly explains the network architecture, with no mention of equations in the text. Certain terms remain unclarified, such as the significance of "h-bar," the distinction between "h-hat" (estimator) and "h" outputs, the necessity of LeakyReLU, and the functions of PETCP, SU-PETCP, TU-PETCP, U_s , and U_t in Equation (4).

8. In the training options, the rationale behind selecting the 8th, 10th, 12th, and 16th frames as reference key frames is unclear. Additionally, the role of temporal offset and how it is addressed require explanation.

9. Information regarding Spatial Upsampling (SU) and Temporal Upsampling (TU) modules is absent, and Figures 3(a) and 4(a) fail to provide insight into these modules.

Reviewer #3

(Remarks to the Author)

The manuscript presents an intelligent paradigm, SparseFLIM, which employs a coupled bidirectional propagation network to reconstruct high-fidelity fluorescence lifetime imaging microscopy (FLIM) data from sparse photon measurements. This addresses a longstanding limitation in FLIM, where there exists a trade-off between measurement duration and information content. The authors have shown over tenfold photon enrichment, dramatically improving SNR, lifetime accuracy, and correlation compared to the original sparse data, demonstrating that SparseFLIM can effectively address spatial and temporal sparsity through upsampling modules. The proposed SparseFLIM technique represents a significant advancement in fluorescence lifetime imaging, addressing a fundamental limitation and potentially enabling new applications in biomedical imaging and sensing.

Minor Revision Comments:

1. The fluorescence decay produced by applying SparseFLIM exhibits a remarkably higher level of smoothness (Fig 1i) when compared to the high SNR (SNR) reference recording (Fig 1j). This intriguing phenomenon raises questions about the underlying factors responsible for this stark contrast in smoothness. The authors should provide a concise clarification on the mechanisms that why the reference image-guided learning approach employed in SparseFLIM can yield lower-noise, higher R-squared restorations.

2. While the photon number histograms in Fig. 2j and Fig. 2k exhibit a high degree of similarity, the authors should offer an explanation for the slight discrepancies observed between these two distributions. Furthermore, it is crucial to address whether these subtle differences have any potential impact on the reliability and accuracy of the inference results obtained through the SparseFLIM approach.

3. The fluorescence decay curve depicted in Fig. 2l (input trace) displays a low level of noise. Whether this curve still necessitates reconstruction through the SparseFLIM model, despite its seemingly low noise levels?

4. The superiority of the proposed SparseFLIM model over alternative approaches, such as 3D UNet and 3D RCAN, is a critical aspect that warrants an explanation from the authors. A detailed elucidation of the specific reasons and mechanisms that render the SparseFLIM model more effective and accurate in reconstructing high-fidelity FLIM data from sparse photon measurements is desirable.

5. The authors have employed channel expansion as a key strategy to address time sparsity in FLIM data. It would be beneficial for the authors to elucidate the advantages of their reconstruction methodology over traditional interpolation techniques. A discussion of the unique strengths and benefits offered by the reconstruction approach, particularly in terms of accurately recovering missing temporal information and preserving data integrity, would enhance the reader's understanding and appreciation of the proposed method.

In summary, this manuscript presents a new approach to address a longstanding challenge in fluorescence lifetime imaging, with a strong potential for enabling new applications in biomedical imaging and sensing. I recommend its publication in the journal after addressing the minor revision comments outlined above.

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

The explanation and terminology regarding the speed improvement by SparseFLIM are difficult to understand, and I couldn't grasp the rebuttal. Is the author suggesting that the SparseFLIM method can achieve similar efficiency with data captured for 10 seconds as it would with data captured in 150 seconds?

Based on Figure S8, there is a noticeable difference between the lifetime estimated using the full 10 ns decay curve and the 5 ns decay. Specifically, the pixels initially in green (approximately 1.1 ns) shifted towards 0.4 ns (red), while the orange pixels moved to a darker range (red). Hence, the authors' claim that for lifetime estimation only decay is located between the (max, max/e) regions is not correct.

The authors' argument that "in practice, however, to reduce complexity, it is often assumed that the current state is only related to the previous states" is incorrect. Especially in the case of fluorescence lifetime decay, each pixel has long-term dependencies (entire decay), not just on previous and next states. Especially when dealing with a bi-exponential decay profile, the maximum signal part of the temporal point spread function (TPSF) is essential for a short lifetime, and the tail part of the decay is crucial for a long lifetime. That's why the difference in mean lifetime estimation is visible in Fig. S8. However, the model and architecture explanation show dependencies only on previous and next elements, ignoring long-term dependencies such as LSTM or GRU units do. I believe this model architecture is only suitable for processing and improving (up-sampling) video signals, which require only the previous and next frame to enhance resolution, but it is not suitable for lifetime TPSF data. I believe the author could have achieved better results by using a simple bi-directional LSTM model or transformer model.

As mentioned in point 4 of the earlier review, both raw data in 1i and 1j must be normalised, and their fitting (NLSF) must be plotted together (raw and fit for 1i and 1j). The reason for repeating this point is simple: in 1j, multiple points (at the peak of decay) are not part of the fitting, resulting in a poor chi-squared value of 1.53. Additionally, these decay curves should be plotted on semilogarithmic scales for better clarity. This exercise will highlight the error in lifetime estimation between fitting and the SparseFLIM method.

In the explanation of the model, the authors mentioned a module called feature extractor (E), but its function has not been explained in the manuscript or supplementary materials. Simply referencing it without explanation is insufficient. The authors also mentioned temporal and spatial attention mechanisms. Where are the equations or parts of the model architecture where the attention mechanism is applied? Which query (Q), key (K), and value (V) were used for attention? Again, merely referencing it without sufficient explanation is inadequate.

Lastly, the reference key frames selected for this study [8, 10, 12, 16] might not be the same for all samples or other implementations. Does the user need to train the model for each experiment with different endo and exogenous dyes?

In Figures S2 and S3, there are mismatches in the tau_m values of SparseFLIM and Photon-rich; what is the explanation for that?

Figure S5e represents background; what is this background? Which is present in the photon-rich section but missing in SparseFLIM. The authors must highlight the background region using pseudocolor to better understand the processing of the model. How is the model getting rid of the background information?

Chi-squared values are missing in all other fits except Fig. S1, for example, Fig. S6.

In conclusion, even after several reads, the article still lacks clarity in a number of sections, particularly the explanations of the model and architecture. I believe authors need to put greater effort into making them sound.

Reviewer #3

(Remarks to the Author)

The authors have addressed my concerns.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

The authors have addressed all my concerns.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, 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 license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <https://creativecommons.org/licenses/by/4.0/>

Dear Reviewers,

Thank you very much for your careful review of our manuscript and additional feedback. The comments are insightful and highly helpful for revising our manuscript and our next work. We have studied the comments and revised our manuscript carefully. Our reply to the reviewers' comments is shown in blue. Our revised text is shown in red here and highlighted in the manuscript. Please find below our point-by-point responses to the reviewers' comments.

Response 1 (Reviewer #1)

This manuscript describes a novel network that presumably estimates multi-component lifetimes from sparse photon data much better than others. This is important because it would speed up data collection at least 10-fold and prevent unnecessary photon-damage, bleaching (and, when applied to life-cell imaging) phototoxicity. The authors apply the techniques to a single FLIM modality (2-photon TCSPC) for which it appears to work fine, and apply it to tissue slices with complex decay characteristics. Their network 'replenishes' photons that were missing in the sparse data.

This topic is very specialist but if it would indeed present a 10-fold relaxation on the requirement to collect photons, compared to current state-of-the-art, it would still be extremely interesting for Commsbio's rather broad readership.

I accepted to review this paper on the base of decades of experience in FLIM, lifetime fitting and manners to present lifetime data, but I only have a few years of limited experience with deep learning networks. I cannot judge whether the backward and forward propagation in the network is absolutely groundbreaking, or whether it is just state of the art in other applications (video-super resolution is mentioned), nor can I judge whether the implementation was optimal, or even correct. In part, I secured some help from a trained physicist (image analyzer) and the review reflects our combined remarks.

Second, after putting in over a full working day in this manuscript, I still am not sure I fully grasp all details and am left with remaining questions. So perhaps this should not be considered a full review.

[Reply] Thank you for sharing your background and experience with FLIM and deep learning. We appreciate the significant effort you have invested in reviewing our manuscript. We understand that you may have remaining questions or areas of uncertainty after reviewing this manuscript and have made revision to address your concern.

(1) Groundbreaking vs. State of the Art:

The bidirectional propagation network we introduced in this manuscript, though inspired by concepts used in video super-resolution, has been significantly adapted and extended for FLIM applications, which is novel. In video super-resolution, the emphasis is on spatial resolution

enhancement, whereas in FLIM, the critical challenge lies in temporal and spatial photon sparsity. Our approach not only enriches photon counts but recovers spatiotemporal information critical for accurate lifetime estimation. This is particularly groundbreaking as it addresses the inherent limitations of FLIM related to photon sparsity and acquisition speed, which have not been effectively tackled by existing deep learning methods in this field.

(2) Optimality and Correctness of Implementation:

We have rigorously tested the implementation of our network to ensure its correctness and optimality. The network architecture, which includes forward and backward propagation branches, has been validated against multiple datasets that span various FLIM modalities, including multispectral and in vivo endoscopic FLIM. This validation process included quantitative comparisons with traditional FLIM techniques, showing significant improvements in SNR, photon count, and lifetime accuracy. These results affirm that our implementation not only adheres to the theoretical model but also enhances real-world applicability.

(3) Understanding and Details of the Manuscript:

We apologize if some sections of the manuscript were unclear and left some questions unanswered. We are committed to improving the manuscript's clarity. We appreciate for the positive comments on our work and have made revision to address the concern.

1) The manuscript is a little hard to read because of the way it is presented. Data on image acquisition, for example, are in part in the methods section, in part in the main text. Terminology is a bit different from standard, for example, the use of 't' on the x-axis instead of the more common terminology ('time bin') and terms like '3D time sequence' which suggested that multiple full FLIM images were collected in a time series (common in life-cell imaging), and details in how the findings are presented. Probably this could be remedied by reformatting and rewriting. I think in particular the first paragraph should be shaped in such a way that there can be no misunderstanding on what exactly is the idea of the algorithm. Also, see 3) for that.

[Reply] We understand your concern regarding the presentation and organization of the manuscript. We have improved the readability and clarity of the manuscript by reorganizing certain sections and using more standard terminology familiar to the FLIM community. Specifically:

1) Data on image acquisition is currently combined into main text. We have rewritten and reorganized the first paragraph to enhance the manuscript's readability and clarity for the FLIM audience.

Modified Text:

“ FLIM measurements were performed using a synchronized system comprising a femtosecond laser (~100 fs, 80 MHz, Chameleon Discovery, Coherent), galvanometric scanner (LSKGG4, Thorlabs), and high-speed time-resolved detectors (HPM-100-40, Becker & Hickl GmbH). The femtosecond excitation beam was relayed, magnified, and corrected by scan lenses (SL50-2P2, Thorlabs) and tube lenses (TTL200MP, Thorlabs) to match the back aperture of a 20× objective (MRD70200, 0.75 NA, Nikon), as shown in Fig. 1a. Emitted fluorescence was collected by the objective and separated from excitation using a long-pass dichroic (DMLP650R, Thorlabs). The second long-pass dichroic (DMLP490R, Thorlabs) split fluorescence from SHG signals. Fluorescence was then detected by the time-resolved detector connected to TCSPC electronics (SPC-150 and DCC-100, Becker & Hickl GmbH), which was synchronized with the laser signal to facilitate precise calibration of time delays. The data acquisition (DAQ) system generated frame pulses, line pulses, and pixel pulses based on the XY scanning signals using three counters, which were subsequently directed to the TCSPC electronics. ...”

2) We have used standard terminology like “time bin” instead of just “t” on axes of Fig. 1c to align better with FLIM conventions and avoid potential confusion. We have changed “3D time sequence” to “life-cell imaging”. We have used standard terminology of “Time (ns)” instead of “t (ns)” on axes of Fig. 1h-j, Fig. 4f, Fig. S7 as many paper did to avoid vagueness.

2) *The claimed beneficiary effects on photobleaching and bleaching, I would think, are related to the authors choice to restrict themselves to 2-photon imaging because TCSPC by itself is already extremely photon-efficient (compaired to gated FLIM or frequency-domain FLIM). In single-photon TCSPC imaging, we routinely get thousands of good FLIM images with just 10 microWatt of excitation power from a single FOV of living cells, and it would be very interesting to test the network on such data as well.*

[Reply] Thank you for raising this important point regarding the claimed benefits of SparseFLIM on reducing photobleaching and phototoxicity. We mentioned photobleaching in the Introduction: “... prolonged exposure increases photobleaching, phototoxicity, and susceptibility to sample perturbation and motion artifacts”. You make a fair critique that the choice of two-photon excitation, rather than the TCSPC technique itself, may be the primary factor contributing to this low optical conversion efficiency. While the excitation intensities of single-photon are lower than two-photon, the repeated raster scanning involved in single-photon TCSPC may still induce cumulative photobleaching over time.

You make a good suggestion to test our SparseFLIM network on FLIM data from living cells. Currently, our experimental setup is optimized for two-photon excitation using a femtosecond laser source. It lacks the necessary excitation sources (suitable visible wavelength, picosecond pulses) and optical components for effective single-photon excitation of common visible fluorophores

used in live cell imaging. Implementing single-photon FLIM of living cells would require modifying our system to incorporate appropriate optical elements compatible with the visible excitation and emission ranges for the fluorescent labels. To partially address this, we tested our network on single-photon FLIM data of liver tissue, as shown in the new Fig. S6. The distorted lifetime results from sparse photon acquisitions were effectively restored by SparseFLIM to be consistent with the FLIM images from sufficient photon counts.

While this liver tissue data provides initial validation, you make a good point that explicitly testing on FLIM of living cells would be a valuable analysis. However, since our system has been tested on both in vivo endoscopy and multispectral FLIM, we believe it will also work in single-photon FLIM with low photon counts.

The high photon accumulation rate in single-photon TCSPC, as you mentioned, may render deep learning reconstruction less crucial in scenarios with sufficient photon statistics. While the current implementation of SparseFLIM focuses on the two-photon excitation regime, where the benefits of photon enrichment and denoising are more pronounced, we acknowledge your insightful perspective. Evaluating SparseFLIM's utility on single-photon TCSPC data would indeed be an interesting future direction, contingent on access to the required instrumentation.

Modified Text:

“... prolonged exposure (repeated raster scanning involved in TCSPC) may increase photobleaching, phototoxicity, and susceptibility to sample perturbation and motion artifacts, especially for photon-inefficient two-photon FLIM ...”

“... We assessed the adaptability of our SparseFLIM model by applying it to three other distinct imaging modes. One such mode is single-photon FLIM of liver metastasis. The distorted lifetime maps obtained from sparse photon acquisitions, which suffer from limited photon counts and noise, were effectively restored by the pre-trained SparseFLIM model to match the high-quality FLIM images acquired with sufficient photon counts (Fig. S6a,c). We compared the fluorescence decay for the sparse photon, the network-enriched output, and the sufficient photon reference data (Fig. S6b,d). The limited photon counts and noise introduced irregularities and distortions in the sparse photon data, causing it to deviate from the expected trend. However, after the network reconstruction, the decay curve of the sparse photon data became highly consistent with that of the sufficient photon reference, accurately capturing the fluorescence decay dynamics. This highlights ability of the network to recover the true fluorescence decay characteristics from single-photon excited sparse and noisy data, effectively mitigating the detrimental effects of limited photon counts and noise. ...”

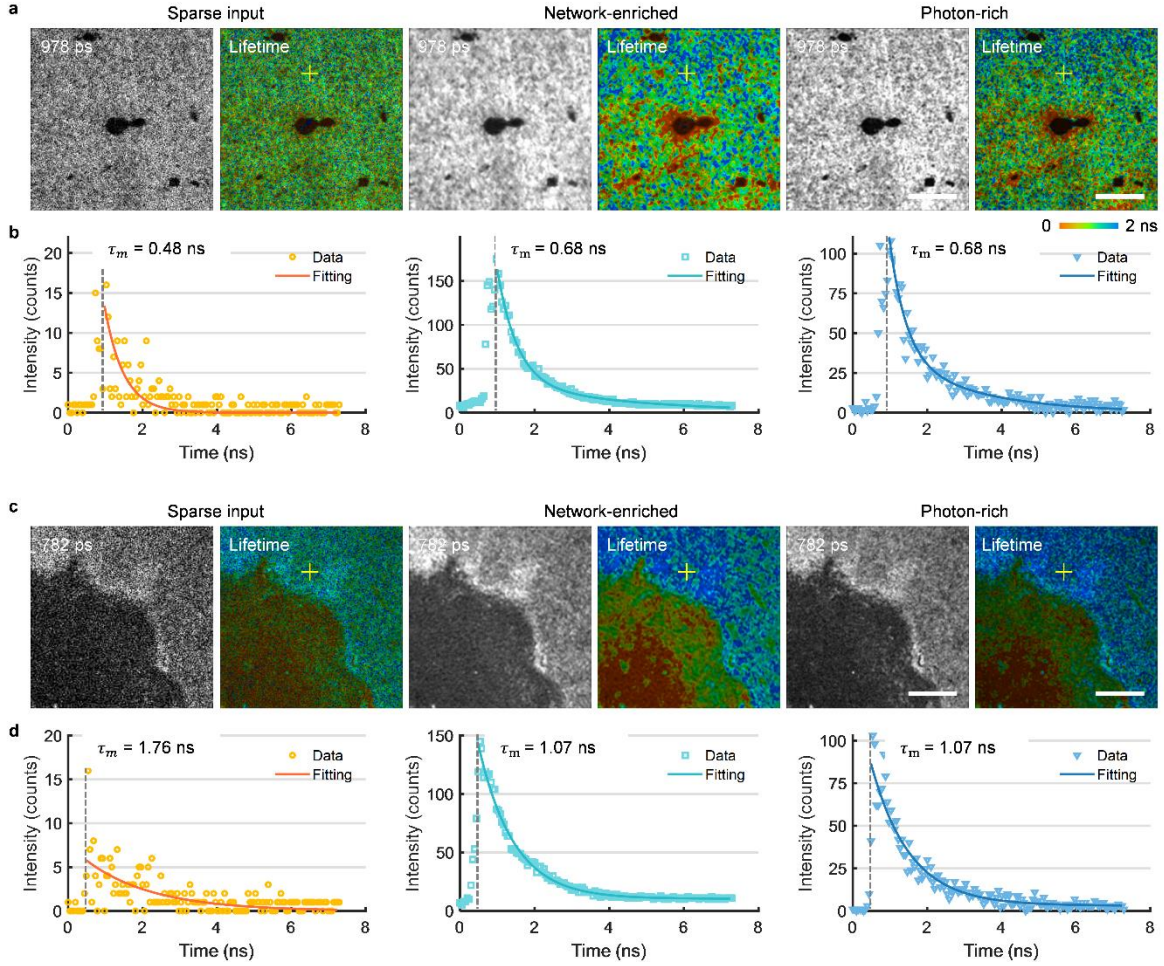


Fig. S1 Network validation on single-photon FLIM.

a, Single-photon FLIM images of normal liver tissue of mouse². **b**, Fluorescence intensity decay over time for the sparse photon, the network-enriched output, and the sufficient photon reference data. **c**, Single-photon FLIM images of liver tissue with boundary between non- and metastasis. **d**, Fluorescence intensity decay over time for the sparse photon, the network-enriched output, and the sufficient photon reference data. Scale bar, 200 μm .

3) To me there are still unclear details on how exactly the procedure works, and I think the manuscript (the main paper) should be written in such a manner that interested non-expert readers can understand the basic principles. For example, Fig. 1h is supposed to reflect the data from a single pixel with a hundred photons. The data are bin counts, and I see at least 20 different Y amplitudes (discrete values of Y coordinate of the circles). If so, than that pixel has WAY more than 100 photons, right? Why is the y-axis in A.U. and not in photon counts? (compare e.g. Fig S7c where only max 5 photons are present per bin).

[Reply] You raised valid points about the lack of clarity in certain details and the need for better explanations targeted towards non-expert readers. We have modified the main text in a more accessible manner to help non-experts understand the basic principles.

Regarding Fig. 1h, you are correct that the data appears to have more than 20 bin counts, which would imply more than the stated 100 photons for a single pixel trace. The reason for this discrepancy is that we had previously normalized the sparse photon FLIM images to better show the indiscernible morphological features (otherwise invisible), which resulted in a confused distribution of the photon counts on the y-axis. This normalization step was an oversight on our part. You provided a helpful suggestion that using labeled axes with clear units (photon counts) instead of arbitrary units would avoid such confusion. We agree and thank you for catching this issue. We have used labeled axes with clear units (intensity counts) instead of arbitrary units would avoid such confusion.

For Fig. S7c, the y-axis units were also not explicitly labeled, though you noted the maximum y-value of ~5 aligned with the stated max 5 photons per bin for that sparse data example. Moving forward, **we have revised all figures to use labeled y-axes with unambiguous photon count units.** Clearing stating the units and avoiding any unnecessary normalization or scalings to prevent confusion about the absolute photon levels present in the data traces.

Thank you again for taking the time to scrutinize these figure details. Identifying and addressing such clarity issues is crucial for accurately communicating our data and findings.

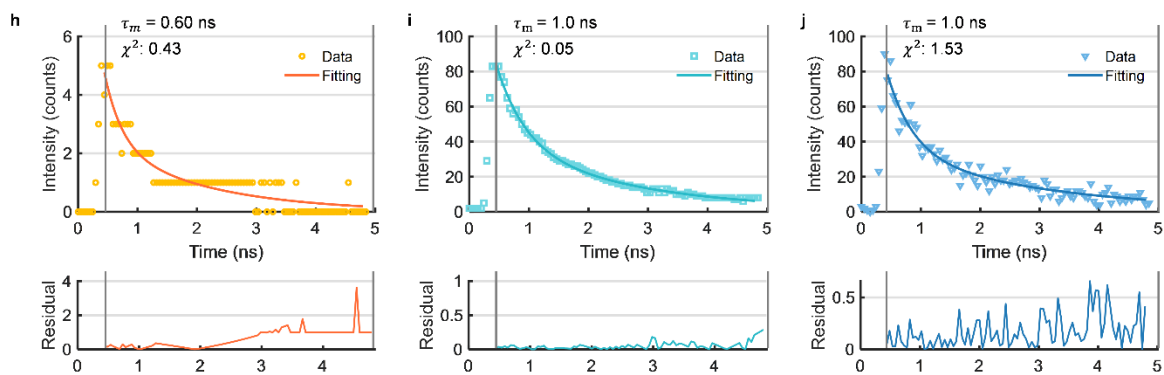
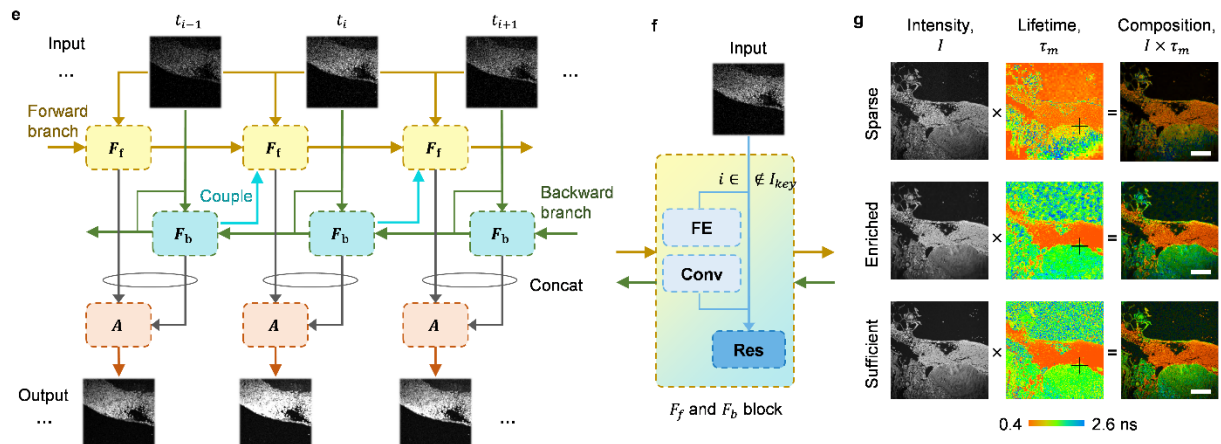
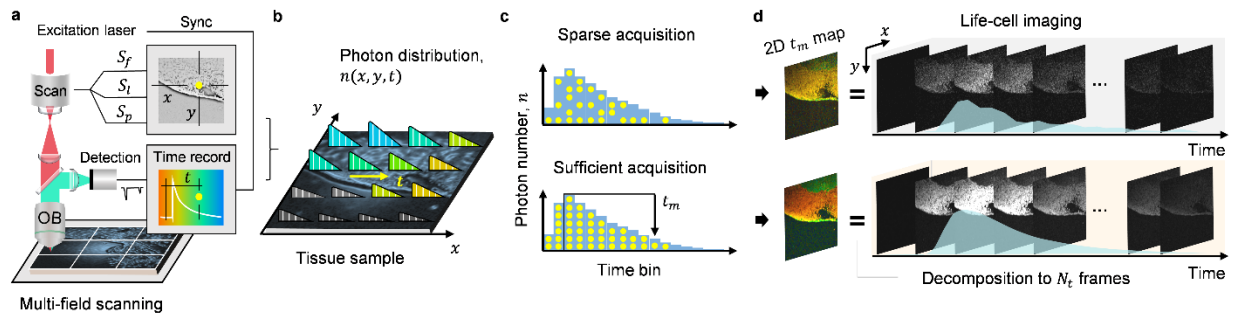


Fig. 1 General principle and validation of SparseFLIM. **a**, FLIM setup and data acquisition. **b**, Photon distribution in the x - y plane. **c**, Comparison of photon counts between sparse and sufficient acquisition modes. **d**, 2D mean lifetime (τ_m) images (left) and 3D lifetime stacks (right). **e**, Network architecture of SparseFLIM. t_i indicates the i th-frame time. F_f , forward propagation branch; F_b , backward propagation branch. A , aggregation blocks. F_f and F_b are coupled indicated by the cyan arrow line. **f**, Information refilling module. FE, feature extractor; Conv, convolution operations; Res, residual blocks **g**, Comparison of the sparse photon input, network output, and sufficient photon reference. **h**, Autofluorescence decay of the location indicated by the cross in **g** for the sparse photon acquisition. **i**, Data restored from the sparse photon recording using our network. **j**, Autofluorescence decay of the sufficient photon recording. The bottom panels correspond to the fitting residuals. Also, see Supplementary video 1. Scale bar, 100 μm .

What exactly happens with the sparseFLIM restoration process? The algorithm generates extra photons to fill the gaps? Or the discrete numbers of photons are replaced with averaged non-discrete numbers to approximate a more smooth curve, and subsequently rescaled to reflect an increased nr of photons?

[Reply] That’s a good question about the specifics of how SparseFLIM restores and enriches the sparse photon data. SparseFLIM might not literally generate or insert new photons into the data. Instead, it leveraged the spatial and temporal correlations learned from the forward and backward data to predict the underlying true fluorescence decay curve and photon distribution that would be observed with sufficient photon counts. More specifically, during the training process, the network learnt a mapping between the sparse, noisy input data and the corresponding high photon count reference data. It captured the relationships between the sparse spatial-temporal patterns and the underlying fluorescence decays they represent. At inference time, when given a new sparse FLIM input, the network used this learned mapping to reconstruct and predict the full, enriched spatial-temporal distribution and decay curve that aligned with the high photon count data distribution. So rather than directly generating photons, it predicted the continuous valued intensities (counts) across space and time that best reconstruct the true underlying signal, while suppressing noise.

The reconstructed intensities could be high precision continuous values approximating the true fluorescence decay curves and spatial distributions as if high photon counts were collected. These reconstructed continuous intensities could then be re-scaled or re-quantized to photon count levels matching high fluence acquisition, which is what we showed in the visualizations. But fundamentally, the network predicts the smooth, underlying distribution.

Modified Discussion:

“.... Notably, SparseFLIM might not literally generate or insert new photons into the data. Instead, it leveraged the spatial and temporal correlations learned from the forward and backward data to predict the underlying true fluorescence decay curve and photon distribution that would be observed with sufficient photon counts. More specifically, during the training process, the network learnt a mapping between the sparse, noisy input data and the corresponding high photon count reference data. It captured the relationships between the sparse spatial-temporal patterns and the underlying fluorescence decays they represent. At inference time, when given a new sparse FLIM input, the network used this learned mapping to reconstruct and predict the full, enriched spatial-temporal distribution and decay curve that aligned with the high photon count data distribution.”

Why, if we focus on fig. 1g, is the average lifetime in the large red part of the image (in the centre, above the cross) quite uniformly red (0.4ns) for Photon-rich and for Enriched cases, while in the sparse image it is orange (0.8 ns) without even a single pixel that is red? If it is just a matter of noise improvements, I would expect some very red pixels in the top image. I do not understand the L-shape of this data cloud.

How can it be that in suppl. fig 5a there appears to be NO correlation between the lifetimes calculated from sparse photon and from sufficient photon data. This just implies that the LMA algorithm does an exceptional bad job, so why using it as a standard against which to benchmark improvements? I do not understand the L-shape of this data cloud. Similarly, why is there NO correlation between reference data and sparse data in Fig. 2h?

[Reply] You raised good points about the discrepancy in the lifetime maps shown in Fig. 1g for the sparse input versus the photon-rich and network enriched cases. “Least Squares” is the conventional algorithm, based on a minimisation of the sum of the squared differences between the data points and the points of the model function, and works well for high photon numbers but has deficiencies when the photon number is low. We initially chose it because we wanted to use one consistent algorithm for the fitting problems of sparse photons, enriched photons, and full photons. And, for sparse photons, we utilized the LMA with Poisson noise modification which is more appropriate for low photon count scenarios compared to the typical Gaussian assumptions.

However, as you suggest, LMA is not very suitable for sparse photons. Now we have used the “Maximum Likelihood Estimation (MLE)” for the input data (except those have worse results than LMA). The fit accuracy is improved especially for low photon numbers, as shown in Fig. 1g.

The “L-shape” distribution likely arises because for most pixels, the noise in the sparse data is so severe that the fitted lifetimes become essentially random, decorrelating from the true lifetimes. Only a small subset may by chance produce lifetimes that weakly correlate. Regarding the lack of correlation between sparse and reference decays in Fig. 2h - this stems from the same issue. With insufficient photons (<100/pixel), the raw decay curves in the sparse data are extremely distorted by noise, decorrelating from the true underlying decays.

Nevertheless, after careful consideration, we determined that the lifetime correlation scatter plots (previous suppl. fig 5a) did not significantly aid the interpretation of our results. Since the mean lifetime values were obtained through algorithm fitting to the raw photon data, comparing the correlations of mean lifetime between different domains may inadvertently introduce artifacts or errors from the fitting process itself.

Given that we have already provided thorough 3D correlation and SNR comparisons between the raw photon data images before and after network reconstruction, the lifetime correlation scatter plots in the previous suppl. fig 5a felt redundant. To streamline the presentation and focus on the most direct, interpretable metrics, we removed this scatter plot figures from the revised Supplementary Information. Thank you again for the insightful feedback.

Modified Text:

“... In the sparse input, the lack of photons and excessive noise prevents resolving the correct lifetime. However, during training on the photon-rich data, the network learns the lifetime features of different lifetime populations/components. When reconstructing from the sparse input, it can leverage this learned knowledge to “unmix” the disordered lifetime distribution and recover the underlying bifurcated ~ 1 ns components seen in the reference. ...”

“... The low correlation likely arises because for most pixels, the noise in the sparse data is so severe that the fitted lifetimes become essentially random, decorrelating from the true lifetimes. Only a small subset may by chance produce lifetimes that weakly correlate. With insufficient photons ($<100/\text{pixel}$), the raw decay curves in the sparse data are extremely distorted by noise, decorrelating from the true underlying decays.”

Modified Methods:

“... We use the conventional Levenberg-Marquardt algorithm (LMA) fitting routine, based on a minimisation of the sum of the squared differences between the data points and the points of the model function, and works well for high photon count fitting. However, LMA is not very suitable for sparsely sampled data, which were thereby fitted using the Maximum Likelihood Estimation (MLE), based on calculating the probability that the values of the model function correctly represent the data points of the decay function.”

It is not intuitively clear why photon refill would result in significantly more signal restoration for ~ 1 ns lifetimes (over 12-fold) than for 3-4 ns lifetimes (just 4 to 5-fold; Fig. 2 i, j). Presented in this way, it looks like short lifetimes are under-represented in LMA fits of sparse data since the sparse FLIM algorithm injects more photons at ~ 1 ns bins than at ~ 3 ns bins. If so, then why are lifetimes longer in the sparse data (fig. 1g, top middle picture)? Yet, the close agreement between

sparseFLIM and photon-rich data shows the network is right, so this is just something that puzzles readers.

[Reply] You raised a good point regarding the seemingly disproportionate enhancement of the ~1 ns lifetime components compared to the 3-4 ns components in the photon distribution plots (Fig. 2i-j). We have provided some potential explanations for this behavior:

Tissue autofluorescence composition: The <1 ns lifetime likely corresponds to short lifetime species like bound FAD in cells/tissue. In contrast, the 3-4 ns components arise from longer lifetime fractions like free FAD. Shorter lifetimes mean faster fluorescence decay with sparser photons in fluorescence tail and more easily distorted when undersampled compared to longer lifetimes. The network learnt the short lifetime features and leveraged this learned knowledge to recover the underlying bifurcated <1 ns photon counts seen in the reference.

Lifetime fitting biases: You make an astute observation that the longer lifetimes seen in the sparse input of Fig. 1g appear contradictory. Our interpretation is that the lack of photons might cause the fitting to erroneously bias towards artificially longer lifetimes. The network is able to correct for this artifact and recover the true, prevalent shorter ~1 ns components seen in the photon-rich data.

You are correct that the close agreement between SparseFLIM and photon-rich data suggests our network is accurately reconstructing the underlying lifetimes and distributions. We likely need to clarify the sources and analysis of these apparent lifetime enrichment discrepancies in the manuscript.

Modified Text:

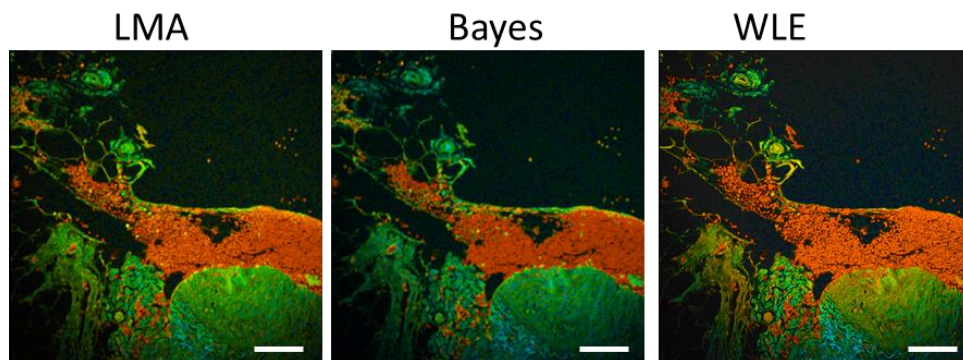
“... However, the <1 ns lifetime components were likely restored with more photon counts than the 3-4 ns components. Shorter lifetimes mean faster fluorescence decay with sparser photons in fluorescence tail and more easily distorted when undersampled compared to longer lifetimes. The lack of photons might cause the fitting to erroneously bias towards artificially longer lifetimes. The network is able to correct for this artifact and recover the true, prevalent shorter ~1 ns components seen in the photon-rich data. ...”

4) Main reservations I have concern the comparison against benchmarks. At the moment, the improvement in fitting is compared against the LMA fit of sparse data (100 photons in 100 bins). LMA is probably one of the worst choices for noisy exponential fitting, and according to literature it favors short lifetimes under these conditions. Everyone would agree that the fitted red line in Fig. S7g, right panel, is just an exceptional poor fit. LMA assumes that each point has symmetric (Gaussian) variance around the expectation value, and for low photon counts the Poisson distribution significantly deviates from a Gaussian.

Later in the paper additional benchmarks are introduced which do not differ so extremely from the sparseFLIM method. Why was the LMA selected rather than e.g. a bayesian method, e.g. maximum likelihood algorithm, which performs much better in sparse data? Or better even, against the best published result until this manuscript?

[Reply] Thank you for raising valid concerns regarding the choice of benchmark used for comparison in our work. As mentioned above, we have chosen the Maximum Likelihood Estimation (MLE) to fit the sparsephoton data instead. For the broader data analysis, we did attempt to use Bayesian estimation methods which can better handle Poisson noise compared to LMA. However, we encountered extremely long computation times required for the Bayesian fits to converge, especially across the not enough photon image. This proved impractical for comprehensive analysis.

We used the LMA fitting routine because works well at sufficient photon levels (see below). For data with sufficient photon levels, the results from LMA fitting approach the maximum likelihood algorithm results from SPCImage software that is widely used for FLIM analysis. This allowed us to benchmark our SparseFLIM reconstructions against a realistic baseline representing conventional sufficient photon analysis.



However, you make a fair point that comparing against cutting-edge sparse data fitting algorithms would provide more meaningful benchmarks. We will aim to incorporate comparisons against best-in-class methods from the latest literature in future works. Thank you again for raising this issue - it has prompted us to re-evaluate our benchmarking methods and more clearly describe the rationale behind the choices made given the computational constraints.

5) How easy and valid is it to transfer this learned network to data from completely unrelated preparations (non-tissue)? Does application to e.g. multicomponent analysis of fluorescent proteins in cell biology require a complete new training?

[Reply] This is a good question about the generalizability and transferability of the trained SparseFLIM network to new datasets from different imaging scenarios. The ability to generalize depends on how different the new data distribution is from the training data the network was originally trained on (photon counts, noise distributions, etc.). For substantially different sample types like fluorescent proteins in cells, which likely have distinct photophysical properties and noise characteristics compared to tissue autofluorescence, it may be necessary to retrain the network to accurately reconstruct lifetime decays. However, the manuscript does demonstrate that the pre-trained network could generalize reasonably well to other imaging modalities like multispectral FLIM and in vivo endoscopy (Fig. S6, 7), which involve different data characteristics from the training set. This suggests that if the new target data shares some fundamental similarities in terms of the decay statistics or imaging parameters, transfer learning by fine-tuning the pre-trained weights may be feasible without full retraining from scratch.

Modified Discussion:

“... Despite these advances, several limitations still need to be addressed. First, although the current results suggest a degree of transferability of the pre-trained weights for datasets sharing core characteristics, the generalization performance across broader datasets remains to be fully validated. In cases involving significantly different optical setups (e.g. confocal microscopy), parameters (e.g. much fewer time bins), and sample types (e.g. fluorescent proteins in cells), it may be necessary to retrain the network for accurate reconstruction of lifetime decays. ...”

6) In *supl.Fig 5*, why are the dots color-coded so that their color depicts the distance from the diagonal (δT)? this information is already in their position on the x and y axis. Rather, use the color to reflect the relative abundance of the number of times that x,y pair occurs in the image. 1 dot = 1 pixel, right? (cannot find that detail). That would be much more valuable to judge how important some clouds of dots are, compared to the whole image (similar to the representation of FACS plots).

[Reply] You’re correct that color-coding the dots based on their distance from the diagonal (δT) provides redundant information already conveyed by their x-y position. Visualizing which lifetime pairs are more prevalent across the full image could provide valuable context. After careful consideration, we determined that these lifetime correlation scatter plots did not significantly aid the interpretation of our results. Since the mean lifetime values are obtained through algorithm fitting to the raw photon data, comparing their correlations may inadvertently introduce artifacts or errors from the fitting process itself.

Given that we have already provided thorough 3D correlation and SNR comparisons between the raw photon data images before and after network reconstruction, the lifetime correlation scatter plots felt redundant. To streamline the presentation and focus on the most direct, interpretable

metrics, we removed those supplementary scatter plot figures from the revised Supplementary Information. Thank you for the insightful suggestion. While visualizing lifetime pair abundances could be valuable, we concluded that removing those figures entirely was preferable to avoid potential fitting artifacts obscuring the direct photon data comparisons.

7) on page 21, limitations of the studies are discussed, but it appears to be more on future directions. The paper also needs a real discussion on possible limitations, e.g.:

- How well does it perform on data from different setups? Is it only applicable on your custom two-photon FLIM setup? What if the system has fewer time bins (and thus, higher counts per bin).
- Does it always assume two single-exponential components? What happens if this is not the case? What is the rationale for sticking with two lifetimes?

[Reply] You make a fair point that the Limitations section reads more like future directions rather than critically discussing the actual limitations of the current study and SparseFLIM approach. We have added the mentioned potential limitations as given in the Discussion section:

Modified Discussion:

“ Despite these advances, several limitations still need to be addressed. First, although the current results suggest a degree of transferability of the pre-trained weights for datasets sharing core characteristics, the generalization performance across broader datasets remains to be fully validated. In cases involving significantly different optical setups (e.g. confocal microscopy), parameters (e.g. much fewer time bins), and sample types (e.g. fluorescent proteins in cells), it may be necessary to retrain the network for accurate reconstruction of lifetime decays. Second, the network restores the temporal and spatial distribution of fluorescence independently of single or double exponential fitting methods. While the bi-component nature of the spontaneous fluorescence of FAD has been extensively discussed in prior research¹⁻³, practical applications should tailor the choice of model based on the specific fluorescence distribution characteristics of the sample. ...”

8) page 7, the Rsquared value was used for estimation of goodness of fit. For non-linear models the Chi-squared value is more appropriate than the R-squared value. R-squared is actually not the goodness-of-fit, but the coefficient of determination.

A high R-squared does not mean that the fit is good, but instead means that the variation is explained by the fit model. Also, R-squared is not valid for non-linear models, to my understanding.

[Reply] Thank you for the suggestion. For non-linear curve fitting, like bi-exponential fluorescence lifetime analysis, reduced chi-squared or other statistical tests that evaluate the

residuals between the fitted model and data would likely be better indicators of fit quality. R-squared values can still provide a general sense of improvement from the deep learning reconstruction, but you make a valid point that more statistically rigorous metrics specific to non-linear fitting should be employed for comprehensive assessment.

The chi-squared statistic can indicate how much difference exists between the observed counts and the expected counts if there were no difference at all in the population. A low value for chi-square means there is little difference between what was observed and what would be expected. In theory, if the observed and expected values were equal (“no difference”) then chi-square would be zero. We have provide the Chi-squared values in Fig. 1h-j.

Minor:

page 10, "The Pearson correlation comparison the network reconstruction results and the sparse inputs also exhibit a substantial enhancement, with a...." Insert a comma after comparison, to reflect an enumeration of arguments.

[Reply] We appreciate the reviewer for the suggestion and have inserted a comma after comparison: “... The Pearson correlation comparison, the network reconstruction results ...”

page 16. " We assessed the feasibility of recovering time sparsity by removing multiple time frames (equivalent to reducing the 256 time channels)." We thought there were just 100 time channels?

[Reply] We appreciate the reviewer for the comment and have corrected the time channel number: “... (equivalent to reducing the 100 time channels) ...”

on page 24, " This facilitates handling misalignments and information..." I can't find out what 'misalignments' means and don't see it introduced in the paper.

[Reply] We appreciate the reviewer for the comment and have corrected the expression: “... This facilitates handling noise and information loss ...”

In summary, a very difficult paper to fully grasp, even after multiple times reading it. I think every effort should be made to make it more transparent.

[Reply] Thank you for your feedback. We understand that the previous technical details and concepts presented in our manuscript could be challenging to fully comprehend. In response to your comments, we have made efforts to improve the transparency and clarity of the paper. We have modified the manuscript to provide more detailed explanations and additional clarifications where needed. Our goal is to ensure that the methodology, results, and implications are accessible

and understandable to readers. We appreciate you taking the time to thoroughly review our work, and we hope that the revisions have enhanced the overall transparency of the paper.

Response 2 (Reviewer #2)

The manuscript introduces SparseFLIM, a deep learning model claiming to address issues related to low photon counts and spatio-temporal sparsity in fluorescence lifetime imaging (FLIM). While the potential of such deep learning approaches in overcoming challenges in FLIM lifetime estimation is acknowledged, however, the manuscript requires significant technical improvements.

[Reply] We appreciate you for the positive comments on our work and have made revision to address your concern:

1. *The manuscript asserts to handle slow photon counting routines in FLIM, yet fails to provide details regarding the speed gain achieved through this Deep Learning method. Additionally, the authors omitted crucial details regarding sparse photon and sufficient photon acquisition e.g., changes in system or camera parameters, illumination power, etc. Clarification is needed regarding whether the photon count mentioned in line 89 represents cumulative photons collected or photons at a specific frame. Furthermore, details regarding the laser repetition rate are necessary to ascertain if the 5ns data range captures truncated or whole decay.*

[Reply] You make a fair critique that the manuscript did not explicitly quantify the speed gain achieved by SparseFLIM compared to conventional FLIM techniques. We have now clarified in the manuscript that for a 512 x 512 image without exogenous fluorescent labeling, SparseFLIM captured a photon count of ~100 per pixel summed over all frames in nearly 10 seconds. In contrast, acquiring sufficient photons yielded an average of ~1,000 photons per pixel in nearly 150 seconds. This highlights a tenfold enhancement in photon collection efficiency enabled by our method. Concerning the laser repetition rate and the 5ns data range, we have added details specifying that the laser operated at a repetition rate of 80 MHz, and the 5ns data range is sufficient to capture the complete fluorescence decay without truncation for the autofluorescence lifetimes observed in our samples.

Modified Text:

“ FLIM measurements were performed using a synchronized system comprising a femtosecond laser (~100 fs, 80 MHz, Chameleon Discovery, Coherent), ...”

“... Without exogenous fluorescent labeling, SparseFLIM captured a photon count of ~100 per pixel summed over all frames for a 512 × 512 image (nearly 10 sec) as depicted in Fig. 1c. In contrast, the sufficient photon collection yielded an average of ~1,000 photons (nearly 150 sec), highlighting a tenfold enhancement. ...”

1. Line 89 states that a 5ns temporal range is sufficient for autofluorescence lifetime estimation, but the basis for this claim is not provided. Moreover, it is unclear how this model can be generalized for other exogenous fluorescence dye lifetimes. The assertion that working within a 5ns range reduces non-semantic information learning repetition requires elaboration.

[Reply] Thank you for pointing out the need for further clarification regarding the 5ns temporal range and its applicability. The basis for claiming that a 5ns temporal range is sufficient for autofluorescence lifetime estimation in our experiments lies in the typical lifetimes observed for the endogenous fluorophores present in the biological tissue samples we studied. Specifically, the autofluorescence from sources such as NADH and FAD, which are prominent endogenous fluorophores in tissues, generally exhibits fluorescence lifetimes in the range of a few nanoseconds (typically <5 ns). Therefore, a 5ns temporal range is adequate to capture the complete fluorescence decay profiles of these autofluorescent species without truncation.

While the current implementation focuses on a 5ns temporal range suitable for autofluorescence lifetimes, our SparseFLIM model is adaptable and can be applied to different imaging modalities and fluorophore lifetimes. Specifically, as mentioned in the manuscript, we have assessed the adaptability of our SparseFLIM model by applying it to three other distinct imaging modes, including *in vivo* endoscopic FLIM. This scenario is particularly relevant to your question, as it involves the use of exogenous fluorescent probes with potentially different lifetime characteristics compared to endogenous autofluorescence. As described in the manuscript, our pre-trained model was able to successfully reconstruct and enhance the low-SNR endoscopy imaging results, recovering both intensity and lifetime information, even with the challenges of fiber dispersion and low numerical aperture.

As for the assertion about reducing non-semantic information learning repetition by working within a 5ns range, we likely meant that by limiting the temporal range to the relevant timescale for the fluorescence decays under study, the network did not need to learn redundant or irrelevant information beyond that range. Notably, the full 256 channels require 2.5 times the GPU memory to run. Less time bins can improve the efficiency of the learning process and prevent the network from overfitting to noise or artifacts outside the region of interest.

We acknowledge that the manuscript could benefit from more detailed explanations and justifications for these points. In the revised version, we have provided a clearer rationale for the chosen temporal range

Modified Text:

“... This temporal range was sufficient to reconstruct the autofluorescence lifetime accurately^{7, 28, 29}, reducing non-semantic information learning repetitions and graphics memory requirements (Methods). ...”

Modified Methods:

“...We utilized an approximately 5 ns temporal range, corresponding to 100 time channels, for autofluorescence lifetime estimation. This range was chosen because the typical lifetimes observed for endogenous fluorophores present in the biological tissue samples under study, such as NADH and FAD, generally exhibit fluorescence lifetimes within a few nanoseconds (typically <5 ns)^{7, 28, 29}. Consequently, comparative analysis revealed negligible differences in fluorescence lifetime estimates between 5 ns and 10 ns time windows (Fig. S8). Moreover, by constraining the temporal range to the relevant timescale for the fluorescence decays of interest, the network circumvented the need to learn redundant or irrelevant information beyond that range. This approach enhances the learning efficiency and mitigates overfitting to noise or artifacts outside the region of interest. ...”

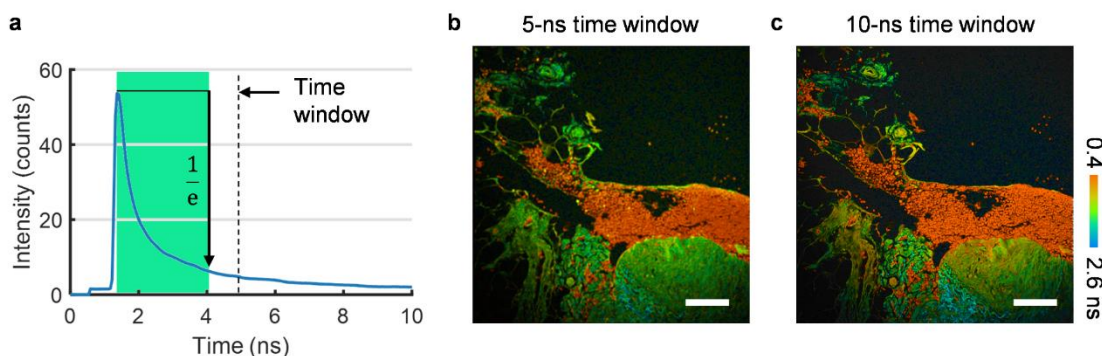


Fig. S2 Comparative analysis of lifetime estimates with different time windows.

a, Fluorescence intensity decay at a full time window of 10 ns for the sufficient photon reference data (cumulative from 262k pixels). The main fluorescence lifetime of endogenous FAD is located in the green region, and the dashed line indicates the time window for autofluorescence lifetime estimation in the experiments. Lifetime map reconstructed by fitting the fluorescence decays within a 5 ns temporal window (**b**) and 10-ns window (**c**). Scale bars, 100 μ m.

2. The manuscript mentions bidirectional information flow in the model, utilizing features from both the current and previous frames. However, it lacks clarity on how temporal decay information with long-term dependencies at each pixel is preserved. Typically, temporal dependencies are maintained through recurrent neural networks, yet evidence of long-term temporal dependency preservation in this model is absent.

[Reply] You are correct that typically, recurrent neural networks (RNNs) are used to maintain long-term temporal dependencies in sequential data. In fact, the model we employed is indeed based on a recurrent neural network architecture, specifically a bidirectional propagation framework adapted from a video super-resolution task.

RNNs are called recurrent neural networks because the current output of a sequence is also related to the previous output. The concrete manifestation is that the network will remember the previous information and apply it to the current output calculation, that is, the nodes between the hidden layers are connected, and the input of the hidden layer includes not only the output of the input layer but also the output of the hidden layer at the previous time. In theory, RNNs can process sequence data of any length. In practice, however, in order to reduce complexity, it is often assumed that the current state is only related to the previous states

The bidirectional propagation in our model consists of two branches: a forward propagation branch and a backward propagation branch. The forward branch processes the frames sequentially from start to end, while the backward branch processes the frames recursively from end to start. This bidirectional processing allows the model to accumulate and integrate long-term spatiotemporal information from both past and future frames at each time step. Additionally, the two branches are coupled, enabling them to exchange information and features during the propagation process. The forward branch receives features propagated from the backward branch, providing it with context from future frames. Similarly, the backward branch receives features from the forward branch, incorporating information from preceding frames.

This coupled bidirectional propagation mechanism, along with the recurrent nature of the architecture, allows our model to effectively capture and preserve long-term temporal dependencies at each pixel location. By leveraging information from both past and future frames, the model can better understand and reconstruct the underlying fluorescence decay dynamics, even in the presence of sparse or noisy data.

In the revised manuscript, we have provided a more detailed explanation of the recurrent neural network architecture employed in our model.

Modified Methods:

“The network employs a recurrent neural network architecture with two key components: (1) Forward and backward propagation branches to leverage temporal correlations and accumulate long-term spatiotemporal information from both past and future frames, and (2) Coupled propagation blocks that exchange information between the forward and backward branches, enabling each branch to incorporate context from the entire sequence. This coupled bidirectional propagation mechanism, along with the recurrent nature of the architecture, allows the model to effectively capture and preserve long-term temporal dependencies at each pixel location. Thereby, the extremely sparse photons (approaching zero) at the tail end of fluorescence lifetime decay can be precisely reconstructed, relying on the photon distribution pattern learned by the network at

high photon time points. Additional elements of the network include aggregation blocks to concatenate the complementary bidirectional information, and high-level feature fusion to reconstruct missing details and reduce artifacts by refining the propagation and aggregation processes. ...”

3. Terms such as "high-level features," "high-quality features," and "reduce cumulative errors" are used ambiguously in lines 124-126. It remains unclear which features the authors aim to capture and whether cumulative errors refer to spatial or temporal reconstruction and what metric is used to compare that.

[Reply] We appreciate you for the comment. The feature extractor module incorporating enhanced deformable convolutional networks extracts complementary spatial and temporal features from the key frames and their neighboring frames. These extracted features are then fused through convolution and integrated into the propagation branches. By refining the propagated features and their aggregation, it enables better reconstruction of fine spatial details and temporal dynamics that may have been lost or distorted in the sparse input data, leading to reduced artifacts and improved overall reconstruction accuracy. In this revised phrasing, we have clarified the following points:

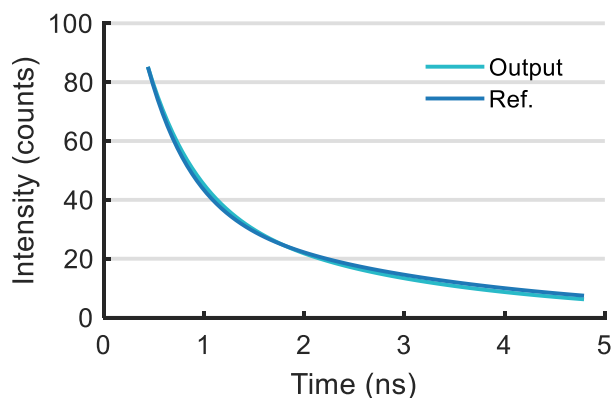
Modified Text:

“... Additionally, the network incorporates a feature extractor from video restoration with enhanced deformable convolutional networks (EDVR)⁴. This approach can extract spatial and temporal features from the key frames and their neighboring frames (Fig. 1f). The features extracted by EDVR include feature maps of different levels, which possess varying resolutions and distinct feature information. In the pyramid structure of cascading and deformable convolutions, the higher levels contain more structural information, but their positional information may undergo slight changes due to the blurring effect caused by repeated convolution and pooling operations. Conversely, the lower levels harbor richer details such as edge textures with more precise positional information. Therefore, performing deformable convolution based on different feature maps can generate more complex transformations, enabling the model to learn how to extract features from reference frames within complex spatiotemporal patterns. This mechanism is particularly beneficial for detailed regions where photon numbers are scarce. We also used temporal and spatial attention⁴ to fuse complementary features from the infrequent key frame extraction. This module aids the network in disregarding irrelevant feature information while emphasizing pertinent feature data for the reconstruction process.”

4. Figure 1i,j displays good non-linear least square fitting with higher photon counts, yet when both normalized graphs would be plotted together, they clearly exhibit different decay rates.

Contrary to the claim that both decays provide the same mean lifetime of 1 ns, both decay curves should be plotted in one graph, and metrics should be provided to compare the mismatch in temporal decay.

[Reply] You raised good points regarding the apparent discrepancy between the fitted decay curves shown in Fig. 1i and 1j. We carefully examined the fitting parameters and visualizations. We plotted the fitting curve of autofluorescence decay for the network output and reference:



Fitting curve of autofluorescence decay for the network output and reference (the reference curve was scaled to the maximum value of the network output curve)

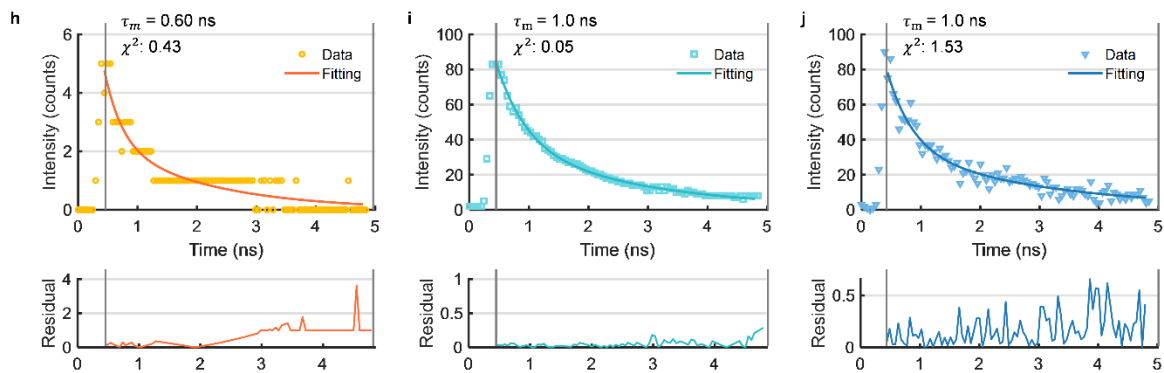
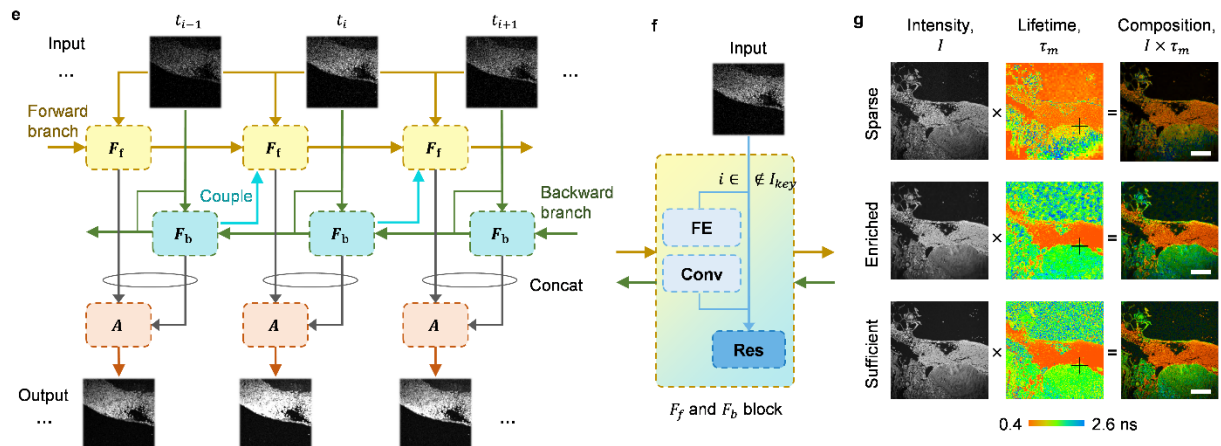
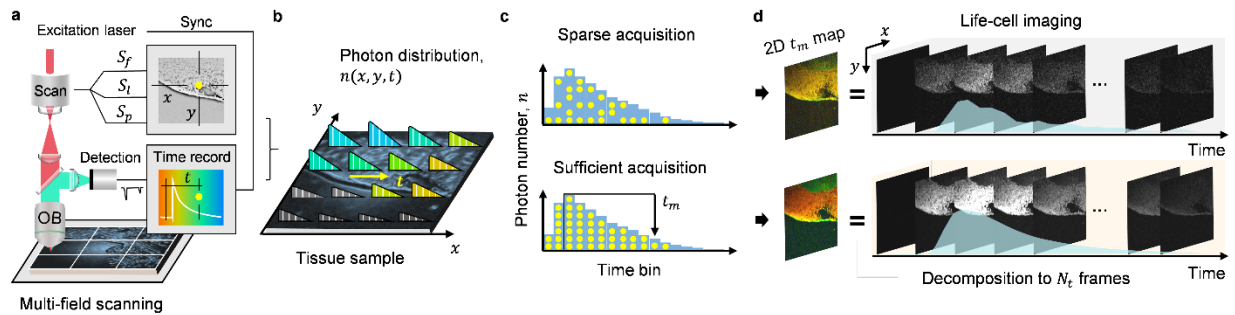


Fig. 2 General principle and validation of SparseFLIM. **a**, FLIM setup and data acquisition. **b**, Photon distribution in the x - y plane. **c**, Comparison of photon counts between sparse and sufficient acquisition modes. **d**, 2D mean lifetime (τ_m) images (left) and 3D lifetime stacks (right). **e**, Network architecture of SparseFLIM. t_i indicates the i th-frame time. F_f , forward propagation branch; F_b , backward propagation branch. A , aggregation blocks. F_f and F_b are coupled indicated by the cyan arrow line. **f**, Information refilling module. FE, feature extractor; Conv, convolution operations; Res, residual blocks **g**, Comparison of the sparse photon input, network output, and sufficient photon reference. **h**, Autofluorescence decay of the location indicated by the cross in **g** for the sparse photon acquisition. **i**, Data restored from the sparse photon recording using our network. **j**, Autofluorescence decay of the sufficient photon recording. The bottom panels correspond to the fitting residuals. Also, see Supplementary video 1. Scale bar, 100 μ m.

5. Line 167 suggests that SparseFLIM excels at restoring intrinsic correlation, but evidence supporting this claim is lacking.

[Reply] We appreciate you for the comment. The intrinsic correlation refers to the enhancement of correlation in Fig. S4b. However, such description was given in the following section. Thus, we have deleted this sentence to avoid redundant expressions.

6. In Figure 2, the network output fluorescence lifetime map appears significantly filtered (smooth) compared to the ground truth. Utilizing better statistical metrics (for all figures) such as SSIM and t-SNE could facilitate a more comprehensive comparison of feature extraction.

[Reply] We appreciate you for the comment. In fact, we included 3D Structural Similarity Index (SSIM) in Table S1 previously. Now we have presented the 3D SSIM in addition to the 3D SNR and Pearson correlation in Fig. 2g. Following network enhancement, the 3D SSIM demonstrates a significant improvement over the original sparse photon stacks, with an average increase of 137% (from ~ 0.35 to ~ 0.84). This substantial increase in SSIM indicates that the network output has enhanced structural and texture similarity to the ground truth, despite the perceived smoothing effect.

While we considered employing the t-SNE algorithm for visualizing and comparing feature representations, we encountered computational challenges and convergence issues when dealing with our large-scale dataset. We computed a stack of image ($512 \times 512 \times 100$) on a high performance computer but got Out of Memory after a couple of days. Additionally, t-SNE focuses primarily on maintaining local structure, but this can come at the cost of compromising the global

structure. Although the visualization results of t-SNE are intuitive, they may be difficult to interpret due to the nonlinear characteristics of the algorithm. For example, distance in a low-dimensional space is not necessarily proportional to distance in a higher-dimensional space.

Therefore, in the revised manuscript, we have focused on the commonly used and interpretable metrics of SNR, SSIM, and Pearson correlation to quantify and evaluate the model's output performance comprehensively. These metrics, along with the visual comparisons and analyses, provide a relatively comprehensive assessment of the network's ability to reconstruct high-fidelity fluorescence lifetime images from sparse photon data.

Modified Text:

“... Following network enhancement, the 3D SNR demonstrates a significant improvement over the original shot-limit stacks, with an average increase of 16.6 dB (~ 12 dB over ~ -4.6 dB), and the 3D SSIM also exhibits a substantial enhancement, with an average increase of 137% (from ~ 0.35 to ~ 0.84). ...”

7. The manuscript poorly explains the network architecture, with no mention of equations in the text. Certain terms remain unclarified, such as the significance of "h-bar," the distinction between "h-hat" (estimator) and "h" outputs, the necessity of LeakyReLU, and the functions of PETCP, SU-PETCP, TU-PETCP, U_s , and U_t in Equation (4).

[Reply] Thank you for pointing out the lack of clarity regarding certain terms and the network architecture details in the manuscript. We have made the following revisions and clarifications:

1. Regarding the term “h-bar”, this notation was used in the original EDVR architecture, where it represented the feature map after being warped for frame-to-frame alignment. However, in our SparseFLIM network, we do not require frame-to-frame alignment since the time frames are inherently aligned. Therefore, we have removed the unnecessary mention of “h-bar” from our manuscript to avoid confusion.

2. We have clarified the distinction between “h-hat” and “h” in the context of our network. “h-hat” represents the feature extraction results obtained from the key frames and their neighboring frames using the feature extractor module.

3. Regarding the use of LeakyReLU activation function, we have added an explanation for its necessity. LeakyReLU improves the training dynamics of neural networks by allowing gradients to flow more freely, reducing the risk of dead neurons (neurons that never activate due to the traditional ReLU function). This enhancement potentially enables the model to learn more complex representations and capture intricate spatial-temporal relationships within the FLIM data.

4. PETCP, SU-PETCP, and TU-PETCP were mistaken writings, and we have corrected them to SparseFLIM, SU SparseFLIM, and TU SparseFLIM, respectively. We have added information explaining the functions of the Spatial Upsampling (SU) and Temporal Upsampling (TU) modules, denoted as Us and Ut, respectively.

By incorporating these revisions and clarifications, we aim to improve the reader’s understanding of the network architecture, the terminology used, and the rationale behind certain design choices.

Modified Methods:

“... $h_i^f, h_i^b = 0$ for the first frame. ... Before these output features (h_i^f and h_i^b) were concatenated, we refilled the loss information due to undersampling.”

Information-refill

“... For example, sparse photon and boundary regions often suffer from noisy. By extracting contextual features from the key frames before and after such events, the lost information can be refilled. Concretely, if the current reference frame is in the key frame set, the input to the module consists of the key frame and its two adjacent supporting frames. The feature extraction module is realized with a relatively lightweight EDVR⁴, so the result of feature extraction is actually the fusion result of EDVR taking these three frames as input and finally output these three frames:

$$\begin{aligned} e_i &= E(x_{i-1}, x_i, x_{i+1}) \\ \hat{h}_i^{\{b, f\}} &= \begin{cases} C(e_i, h_i^{\{b, f\}}), & i \in I_{key} \\ h_i^{\{b, f\}}, & i \notin I_{key} \end{cases} \end{aligned} \quad (1)$$

where E is the feature extractor. C is the convolution. I_{key} represents the key frame number. $\hat{h}$ is the result after information-refill.

For feature extraction of key frames, the significance of the information contained within intra-frame regions varies. Support key frames may exhibit artifacts such as blur, noise, or loss of signal photons. To address this issue, we also included the temporal and spatial attention mechanism to disregard irrelevant feature information while focusing on pertinent data for accurate reconstruction, which was detailed in EDVR⁴.

Next, transfer the result of information-refill to the feature reconstruction:

$$h_i^{\{b, f\}} = R_{b, j}(x_i, \hat{h}_i^{\{b, f\}}) \quad (2)$$

where $R_{b, f}$ is the feature reconstruction module with eight residual blocks for backward and forward and branches. The information-refill can provide useful guidance to correct faulty lifetime decay reconstruction and enrich feature learning. ...”

“... Leaky ReLU improves the training dynamics of neural networks by allowing gradients to flow more freely, reducing the risk of dead networks neurons, and potentially enhancing the model’s ability to learn complex representations. ...”

“... The output frame y_i is obtained by

$$y_i = \begin{cases} F_f(A_i) & \text{for SparseFLIM} \\ U_s[F_f(A_i)] & \text{for SU SparseFLIM} \\ U_t[F_f(A_i)] & \text{for TU SparseFLIM} \end{cases} \quad (3)$$

where U_s denotes the upsampling module, which comprises convolutional layers for extracting features and a pixel shuffle layer for upsampling:

$$U_s = \text{Leaky ReLU}\left(\text{PixelShuffle}\left(\text{conv}(F_f(A_i))\right)\right) \quad (4)$$

The convolutional layer extracts features from the input, which is the result of the fusion of the feature maps. PixelShuffle performs the actual spatial upsampling by rearranging the feature channels into a higher-resolution output image. It effectively increases the spatial resolution by a spatial upsampling factor (e.g., $2\times$, $3\times$, or $4\times$). These ultimately yield a high-resolution image as output.

U_t represents a channel upsampling module that expands output channels of the last convolutional layer to more temporal channel bins and allocate adjacent ones:

$$U_t = \text{Leaky ReLU}\left(\text{conv}(F_f(A_i), s_t)\right) \quad (5)$$

where s_t is the number of output channels of the convolutional layer, which match the desired temporal scale factor (e.g., $2\times$, $3\times$, or $4\times$). Then, these adjacent output channels are allocated to the collected frames and the missing frames that need to be reconstructed. For instance, the four output channels of the convolutional layer ($s_t = 4$), C_1 , C_2 , C_3 , and C_4 can be allocated to $I(x, y, t_1)$, $I(x, y, t_2)$, $I(x, y, t_3)$, and $I(x, y, t_4)$, respectively. By expanding the output channels and assigning them appropriately, the TU module enables the network to reconstruct and synthesize the missing time frames, effectively recovering the temporal information lost due to temporal sparsity or downsampling during acquisition.

”

8. In the training options, the rationale behind selecting the 8th, 10th, 12th, and 16th frames as reference key frames is unclear. Additionally, the role of temporal offset and how it is addressed require explanation.

[Reply] Thank you for raising these points regarding the selection of reference key frames and the role of temporal offset. We have clarified these aspects: The 8th, 10th, 12th, and 16th frames were selected as reference key frames for the following reasons: 1. As shown in Table S1, these specific

frame selections achieved the highest attained SNR compared to other key frame selections during the training process. 2. The selected frames exhibit high contrast and low noise levels, which are desirable characteristics for effective feature extraction and network training.

The temporal offset refers to the delay before the fluorescence begins to decay after excitation. In our data, the temporal offset is approximately 10 frames (~ 0.5 ns) based on the following considerations: 1. A temporal offset of ~ 0.5 ns ensures that the rising edge of the fluorescence is preserved, which is important for capturing the complete fluorescence decay dynamics. 2. While preserving the rising edge, it is crucial to have a sufficient subsequent time range to capture the complete decay of the fluorescence signal. A longer temporal offset would result in an insufficient time range to observe the decay within the acquired data.

In the revised manuscript, we have provided a detailed explanation of the rationale behind the selection of reference key frames and the role of temporal offset, including the considerations mentioned above.

Modified Methods:

“... Reference key frames were selected from the [8, 10, 12, 16]th frames with the highest attained SNR compared to other key frame selections (Table S1). The network can effectively leverage the high SNR, contrast, and low noise levels in these frames for feature extraction and training. ...”

“... Additionally, the chosen temporal offset (i.e. the delay before the fluorescence begins to decay after excitation) of approximately 0.5 ns allows for preserving the fluorescent rising edge while ensuring a sufficient subsequent time range to capture the effective fluorescence decay. ...”

9. Information regarding Spatial Upsampling (SU) and Temporal Upsampling (TU) modules is absent, and Figures 3(a) and 4(a) fail to provide insight into these modules.

[Reply] Thank you for pointing out the lack of information regarding the Spatial Upsampling (SU) and Temporal Upsampling (TU) modules. As mentioned above, we have included more detailed descriptions of the SU and TU modules to provide better insights into their functionalities and how they contribute to the spatial and temporal upsampling capabilities of the SparseFLIM network.

The SU module, represented as U_s , is responsible for spatial upsampling in the network architecture. It consists of two key components:

1. Convolutional layer: This layer extracts features from the input, which is the result of the fusion of the feature maps from the forward and backward propagation branches.
2. PixelShuffle: This operation performs the actual spatial upsampling by rearranging the feature channels into a higher-resolution output image. It effectively increases the spatial resolution by a specified upsampling factor (e.g., $2\times$, $3\times$, or $4\times$).

The input to the SU module is the aggregated feature map from the bidirectional propagation branches, and the output is a high-resolution image with enhanced spatial details and enriched photon information.

The TU module, denoted as U_t , handles temporal upsampling by synthesizing missing time frames. It operates as a channel upsampling module, which works as follows:

1. Output channel expansion: The module increases the number of output channels of the convolutional layer to match the desired number of time frames, including both the collected frames and the missing frames that need to be reconstructed.

2. Channel assignment: adjacent output channels are assigned to represent the collected and reconstructed time frames, respectively.

By expanding the output channels and assigning them appropriately, the TU module enables the network to reconstruct and synthesize the missing time frames, effectively recovering the temporal information lost due to temporal sparsity or downsampling during acquisition.

Modified Text:

“... we employed a spatial upsampling (SU) module, realized by sub-pixel convolution (pixel shuffle)³⁷ following the feature aggregation within the network of SparseFLIM, as illustrated in Fig. 3a and described in detail in Methods. ...”

“... we employed a temporal upsampling (TU) module, realized by increasing the output channels of the network (Fig. 4a, see details in Methods). ...”

Modified Methods:

“... where U_s denotes the upsampling module, which comprises convolutional layers for extracting features and a pixel shuffle layer for upsampling:

$$U_s = \text{Leaky ReLU}\left(\text{PixelShuffle}\left(\text{conv}(F_f(A_i))\right)\right) \quad (6)$$

The convolutional layer extracts features from the input, which is the result of the fusion of the feature maps. PixelShuffle performs the actual spatial upsampling by rearranging the feature channels into a higher-resolution output image. It effectively increases the spatial resolution by a spatial upsampling factor (e.g., $2\times$, $3\times$, or $4\times$). These ultimately yield a high-resolution image as output.

U_t represents a channel upsampling module that expands output channels of the last convolutional layer to more temporal channel bins and allocate adjacent ones:

$$U_t = \text{Leaky ReLU}\left(\text{conv}(F_f(A_i), s_t)\right) \quad (7)$$

where s_t is the number of output channels of the convolutional layer, which match the desired temporal scale factor (e.g., $2\times$, $3\times$, or $4\times$). Then, these adjacent output channels are allocated to the collected frames and the missing frames that need to be reconstructed. For instance, the four output channels of the convolutional layer ($s_t = 4$), C_1 , C_2 , C_3 , and C_4 can be allocated to

$I(x, y, t_1)$, $I(x, y, t_2)$, $I(x, y, t_3)$, and $I(x, y, t_4)$, respectively. By expanding the output channels and assigning them appropriately, the TU module enables the network to reconstruct and synthesize the missing time frames, effectively recovering the temporal information lost due to temporal sparsity or downsampling during acquisition.”

Response 3 (Reviewer #3)

The manuscript presents an intelligent paradigm, SparseFLIM, which employs a coupled bidirectional propagation network to reconstruct high-fidelity fluorescence lifetime imaging microscopy (FLIM) data from sparse photon measurements. This addresses a longstanding limitation in FLIM, where there exists a trade-off between measurement duration and information content. The authors have shown over tenfold photon enrichment, dramatically improving SNR, lifetime accuracy, and correlation compared to the original sparse data, demonstrating that SparseFLIM can effectively address spatial and temporal sparsity through upsampling modules. The proposed SparseFLIM technique represents a significant advancement in fluorescence lifetime imaging, addressing a fundamental limitation and potentially enabling new applications in biomedical imaging and sensing.

[Reply] We appreciate you for the positive comments on our work and have made revision to address your concern:

Minor Revision Comments:

1. The fluorescence decay produced by applying SparseFLIM exhibits a remarkably higher level of smoothness (Fig 1i) when compared to the high SNR (SNR) reference recording (Fig 1j). This intriguing phenomenon raises questions about the underlying factors responsible for this stark contrast in smoothness. The authors should provide a concise clarification on the mechanisms that why the reference image-guided learning approach employed in SparseFLIM can yield lower-noise, higher R-squared restorations.

[Reply] We appreciate you for the comment. The key factor contributing to the smoother appearance of the SparseFLIM reconstruction is the network's ability to learn and reconstruct the underlying fluorescence decay patterns while effectively suppressing random noise present in the input data. During the training process, the network learns to map the sparse, noisy input data to the corresponding high SNR reference data. However, since independent noise components in the input data are expected to have a zero mean, the network cannot effectively learn or reconstruct these random noise patterns. Instead, the network focuses on learning the underlying fluorescence decay characteristics and the relationships between the sparse input and the clear, high SNR reference data.

As a result, when reconstructing from the sparse input during inference, the network can recover the high-SNR fluorescence decay patterns while inherently filtering out the random noise components that were present in the original sparse input. This noise suppression capability of the network leads to smoother and cleaner fluorescence decay traces in the SparseFLIM reconstruction compared to the high SNR reference, which may still contain some residual noise.

It is important to note that the smoothness in the SparseFLIM reconstruction does not necessarily imply a loss of important features or details. Instead, it reflects the network’s ability to separate and reconstruct the underlying fluorescence decay dynamics from the noise-contaminated sparse input, leveraging the guidance provided by the high SNR reference data during training.

In the revised manuscript, we have provided a concise explanation of this mechanism:

Modified Text:

“... Moreover, the residual error of the SparseFLIM result is largely lower than that of the sparse input and even smaller than that of the sufficient photons as observed in the bottom panel of Fig. 1h–j. This smoothness arises because the network cannot effectively learn or reconstruct independent noise components with a zero mean present in the input data. As a result, when reconstructing from the sparse input during inference, the network can recover the high-SNR fluorescence decay patterns while inherently filtering out the random noise components that were present in the original sparse input. This noise suppression capability of the network leads to smoother and cleaner fluorescence decay traces in the SparseFLIM reconstruction compared to the high SNR reference, which may still contain some residual noise. ...”

2. While the photon number histograms in Fig. 2j and Fig. 2k exhibit a high degree of similarity, the authors should offer an explanation for the slight discrepancies observed between these two distributions. Furthermore, it is crucial to address whether these subtle differences have any potential impact on the reliability and accuracy of the inference results obtained through the SparseFLIM approach.

[Reply] Thank you for raising this point about the slight discrepancies observed between the photon number histograms in Fig. 2j and Fig. 2k. While the photon number histograms in Fig. 2j (network-enriched photon) and Fig. 2k (photon-rich reference) exhibit a high degree of overall similarity, there are indeed subtle differences between these two distributions. These small discrepancies may arise due to 1) imperfect network reconstruction and 2) errors in fitting methods. Despite the network’s capabilities, there may be limitations in perfectly reconstructing the photon distribution to match the reference data, leading to minor deviations. The fitting methods used to extract the photon counts from the fluorescence decay traces may introduce some inaccuracies, contributing to the observed differences in the histograms.

However, it is important to note that these subtle differences have less potential impact on the reliability and accuracy of the inference results obtained through the SparseFLIM approach for the following reasons: 1) The histograms represent a statistical distribution across a large number of stacks (over 7 million fluorescence decay traces), and while the distributions are not completely consistent, the overall trends and shapes are closely aligned. 2) Some discrepancies may be beneficial for the model’s generalization capability, as it prevents overfitting to the reference data

and allows for better adaptation to different scenarios. 3) As presented in the manuscript, the network output exhibits significant improvements in terms of 3D SNR, SSIM, and correlation compared to the input data. These quantitative metrics, along with the visual comparisons, provide strong evidence for the reliability and accuracy of the SparseFLIM results.

In the revised manuscript, we have offered an explanation for the slight discrepancies observed between these two distributions and explained whether these subtle differences have any potential impact on the reliability and accuracy of the inference results

Modified Text:

“... While the photon number histograms in Fig. 2j and Fig. 2k exhibit a high degree of overall similarity, there are subtle discrepancies between these two distributions. This may arise due to imperfect network reconstruction and errors in fitting methods. Nevertheless, these subtle differences have less potential impact on the reliability and accuracy of the inference results of SparseFLIM because the overall lifetime trends and shapes are closely aligned.”

3. The fluorescence decay curve depicted in Fig. 2l (input trace) displays a low level of noise. Whether this curve still necessitates reconstruction through the SparseFLIM model, despite its seemingly low noise levels?

[Reply] Thank you for raising this observation The fluorescence decay curve shown in Fig. 2l is the cumulative decay curve of the lifetime components within the white dashed box region in the y - t view. As a result, it appears to exhibit a lower noise level due to the averaging effect of multiple pixels within that region. The noise level of a single pixel's fluorescence decay trace is significantly higher, as illustrated in Fig. 1h, that the sparse input data for a single pixel location displays a highly irregular and noisy decay pattern, deviating substantially from the expected exponential decay behavior. Therefore, the reconstruction through the SparseFLIM model is still necessary to recover the underlying fluorescence decay dynamics. We have added the explanation:

Modified Text:

“... The x - t and y - t views of the RBCs visualizing fluorescence decay within a 5-ns window, with cumulative decay trace plotted in the bottom right, which corresponds to the lifetime components within the white box region in the y - t view. ...”

4. The superiority of the proposed SparseFLIM model over alternative approaches, such as 3D UNet and 3D RCAN, is a critical aspect that warrants an explanation from the authors. A detailed elucidation of the specific reasons and mechanisms that render the SparseFLIM model more

effective and accurate in reconstructing high-fidelity FLIM data from sparse photon measurements is desirable.

[Reply] Thank you for highlighting the importance of explaining the superiority of the proposed SparseFLIM model over alternative approaches, such as 3D UNet and 3D RCAN. We have provided a detailed elucidation of the specific reasons and mechanisms that make the SparseFLIM model more effective and accurate in reconstructing high-fidelity FLIM data from sparse photon measurements.

The SparseFLIM model employs a novel bidirectional propagation mechanism, consisting of forward and backward propagation branches that are coupled and exchange information during the reconstruction process. This bidirectional approach allows the model to effectively capture and leverage long-term spatiotemporal correlations and dependencies within the FLIM data, both from past and future frames. In contrast, approaches like 3D UNet and 3D RCAN primarily rely on unidirectional processing, which may limit their ability to capture and utilize the rich spatiotemporal relationships present in the FLIM data. The coupled bidirectional propagation in SparseFLIM enables a more holistic sequence modeling and reconstruction, leading to enhanced photon enrichment and recovery of underlying fluorescence dynamics.

The SparseFLIM model incorporates a feature extractor module that leverages enhanced deformable convolutional networks to extract complementary spatial and temporal features from key frames and their neighboring frames. These extracted features are then fused and integrated into the propagation branches, refining the reconstruction process and reducing cumulative errors. This high-level feature fusion mechanism allows the SparseFLIM model to capture and utilize more comprehensive information from the input data, leading to improved reconstruction of missing details and suppression of artifacts. The alternative approaches, such as 3D UNet and 3D RCAN, may not explicitly incorporate such a feature extraction and fusion mechanism, potentially limiting their ability to recover fine-grained spatial and temporal information.

The SparseFLIM model is specifically designed to capture and reconstruct the temporal correlations and dynamics present in fluorescence decay data. The bidirectional propagation, coupled with the recurrent nature of the architecture, enables the model to effectively learn and preserve the long-term temporal dependencies at each pixel location, which is crucial for accurate reconstruction of fluorescence decay curves. Approaches like 3D UNet and 3D RCAN, while capable of spatial reconstruction, may not be optimized to effectively capture and reconstruct the intricate temporal dynamics and correlations inherent in FLIM data, potentially leading to less accurate recovery of fluorescence decay characteristics.

Modified Text:

“... The superior performance of SparseFLIM benefits from its unique strengths in capturing spatiotemporal correlations, leveraging feature extraction and fusion, and learning temporal dynamics. The bidirectional approach allows the model to effectively capture and leverage long-

term spatiotemporal correlations and dependencies within the FLIM data, both from past and future frames. In contrast, approaches like 3D UNet and 3D RCAN primarily rely on nonpropagating reconstruction, which may limit their ability to capture and utilize the rich spatiotemporal relationships present in the FLIM data. The feature fusion mechanism allows SparseFLIM to capture and utilize more comprehensive information from the input data, leading to improved reconstruction of missing details and suppression of artifacts. Other approaches may not explicitly incorporate such a feature extraction and fusion mechanism, potentially limiting their ability to recover fine-grained spatial and temporal information.”

5. The authors have employed channel expansion as a key strategy to address time sparsity in FLIM data. It would be beneficial for the authors to elucidate the advantages of their reconstruction methodology over traditional interpolation techniques. A discussion of the unique strengths and benefits offered by the reconstruction approach, particularly in terms of accurately recovering missing temporal information and preserving data integrity, would enhance the reader’s understanding and appreciation of the proposed method.

[Reply] Thank you for raising this point about the advantages of the proposed reconstruction methodology over traditional interpolation techniques in addressing time sparsity in FLIM data. We have elucidated the unique strengths offered by our approach:

Unlike traditional interpolation techniques that rely on mathematical assumptions or predefined rules, our approach leverages expansion of the output channels of the network to reconstruct the temporal sparsity in FLIM data. Through the bidirectional propagation architecture and the recurrent nature of the network, the network can accurately recover missing temporal information by leveraging the knowledge gained from the high-SNR reference data during training. Traditional interpolation methods, on the other hand, may not capture these complex temporal patterns, potentially leading to inaccuracies or artifacts in the recovered temporal information.

Our reconstruction methodology is designed to preserve the integrity of the FLIM data by leveraging the rich spatiotemporal context and avoiding overfitting to noise or artifacts. The coupled bidirectional propagation mechanism and the feature extraction and fusion components allow our model to effectively utilize the complementary information from both past and future frames, as well as high-level spatial and temporal features. This holistic approach helps mitigate potential distortions or artifacts that may arise from interpolation techniques, which often rely on local information or predefined assumptions that may not accurately capture the underlying fluorescence decay dynamics. By learning from the high-SNR reference data, our model can reconstruct missing temporal information while preserving the integrity of the fluorescence decay characteristics.

One key advantage of our deep learning-based reconstruction approach is its adaptability and ability to generalize to different scenarios and experimental conditions. Through the training

process, our model can learn to reconstruct a wide range of fluorescence decay patterns and dynamics, enabling its application to diverse FLIM modalities and sample types. Traditional interpolation techniques, on the other hand, may be limited by their predefined assumptions or rules, which may not be universally applicable or easily adaptable to varying experimental conditions or sample characteristics. Our approach's ability to learn from diverse data during training allows for better generalization and applicability across different FLIM scenarios.

Modified Text:

“... Unlike traditional interpolation techniques that rely on mathematical assumptions or predefined rules, our approach leverages expansion of the output channels of the network to reconstruct the temporal sparsity in FLIM data. Through the bidirectional propagation architecture and the recurrent nature of the network, the network can accurately recover missing temporal information by leveraging the knowledge gained from the high-SNR reference data during training. Traditional interpolation methods, on the other hand, may not capture these complex temporal patterns, potentially leading to inaccuracies or artifacts in the recovered temporal information. The coupled bidirectional propagation mechanism and the feature extraction and fusion components helps mitigate potential distortions or artifacts that may arise from interpolation techniques.”

In summary, this manuscript presents a new approach to address a longstanding challenge in fluorescence lifetime imaging, with a strong potential for enabling new applications in biomedical imaging and sensing. I recommend its publication in the journal after addressing the minor revision comments outlined above.

Thank you again for the highly positive comments on our work.

References

1. Shen, B.L. et al. Label-free whole-colony imaging and metabolic analysis of metastatic pancreatic cancer by an autoregulating flexible optical system. *Theranostics* **10**, 1849–1860 (2020).
2. Skala, M.C. et al. In vivo multiphoton microscopy of NADH and FAD redox states, fluorescence lifetimes, and cellular morphology in precancerous epithelia. *P Natl Acad Sci USA* **104**, 19494–19499 (2007).
3. Ranjit, S. et al. Measuring the effect of a Western diet on liver tissue architecture by FLIM autofluorescence and harmonic generation microscopy. *Biomed Opt Express* **8**, 3143–3154 (2017).
4. Wang, X.T. et al. EDVR: Video Restoration with Enhanced Deformable Convolutional Networks in *2019 Ieee/Cvf Conference on Computer Vision and Pattern Recognition Workshops* 1954–1963 (IEEE, Long Beach; 2019).

Dear Reviewers,

Thank you very much for your careful review of our manuscript and additional feedback. The comments are insightful and highly helpful for revising our manuscript and our next work. We have studied the comments and revised our manuscript carefully. Our reply to the reviewers' comments is shown in blue. Our revised text is shown in red here and highlighted in the manuscript. Please find below our point-by-point responses to the reviewers' comments.

Response (Reviewer #2)

The explanation and terminology regarding the speed improvement by SparseFLIM are difficult to understand, and I couldn't grasp the rebuttal. Is the author suggesting that the SparseFLIM method can achieve similar efficiency with data captured for 10 seconds as it would with data captured in 150 seconds?

[Reply] Yes, SparseFLIM can achieve similar efficiency with data captured in a much shorter time compared to conventional methods. For a 512×512 image with sufficient photon counts (around 1,000 photons per pixel), the acquisition time is nearly 150 seconds. For the same 512×512 image size, but with sparse photon counts (around 100 photons per pixel), the acquisition time is nearly 10 seconds. The SparseFLIM method is designed to take this sparse photon data and reconstruct it to match the quality of the sufficient photon data (which would normally take 150 seconds to acquire). This means that SparseFLIM can potentially achieve a 15 $\times$ speed improvement (150 seconds reduced to 10 seconds) while maintaining similar image quality and information content. The method accomplishes this by using deep learning to reconstruct high-quality FLIM data from the sparse photon measurements, effectively "filling in" the missing information that would normally require longer acquisition times.

Modified Text:

".... For a 512×512 image with sufficient photon counts (around 1,000 photons per pixel), the acquisition time is nearly 150 seconds without exogenous fluorescent labeling. For the same image size, but with sparse photon counts (around 100 photons per pixel), the acquisition time is nearly 10 sec (Fig. 1c). The SparseFLIM method is designed to process this sparse photon data and reconstruct it to match the quality of the sufficient photon data, which traditionally demands a much longer acquisition time. This suggests that SparseFLIM can achieve a remarkable 15-fold improvement in imaging speed while enhancing image quality and preserving critical information content."

Based on Figure S8, there is a noticeable difference between the lifetime estimated using the full 10 ns decay curve and the 5 ns decay. Specifically, the pixels initially in green (approximately 1.1

ns) shifted towards 0.4 ns (red), while the orange pixels moved to a darker range (red). Hence, the authors' claim that for lifetime estimation only decay is located between the (max, max/e) regions is not correct.

[Reply] We appreciate the your careful observation of Figure S8. You are correct that there are noticeable differences between the lifetime estimates using the full 10 ns decay curve and the 5 ns decay curve. We acknowledge that our initial claim about the negligible differences was overstated, and we thank you for bringing this to our attention.

The decay curve of fluorescence lifetime typically follows an exponential decay pattern, which is why it is generally not necessary to capture the entire decay curve to estimate the lifetime accurately. This principle is also the basis for gated lifetime measurements. In theory, the most informative portion of the decay curve for lifetime estimation is indeed between (max, max/e), as we initially suggested. However, we recognize that in practice, the actual fitting process can be influenced by various factors, particularly noise. This is especially true at the tail end of the decay curve where photon counts are low. These noise effects can introduce errors in the fitting process, which may explain the discrepancies observed in Figure S8.

Given these considerations, we propose to modify our expression and clarify our approach:

Modified Text:

“... Nevertheless, comparative analysis revealed a little differences in fluorescence lifetime estimates between 5 ns and 10 ns time windows (Fig. S8). Our choice of a 5 ns window represents a careful compromise between capturing sufficient decay information and minimizing the influence of noise in the low-photon tail region. In sparse photon conditions, the tail end of the decay curve often suffers from increased noise due to low photon counts. We aim to capture the most informative portion of the decay curve while reducing the impact of noise-induced artifacts. By constraining the temporal range to the relevant timescale for the fluorescence decays of interest, the network circumvented the need to learn redundant or irrelevant information beyond that range. This approach enhances the learning efficiency and mitigates overfitting to noise or artifacts outside the region of interest despite impact for longer lifetime reconstructions. ... Moving forward, we will explore adaptive windowing techniques to potentially extend the analyzable decay range without compromising the robustness of our approach in low-photon conditions. These refinements will help to broaden the applicability of SparseFLIM while maintaining its advantages in processing speed and photon efficiency.”

The authors' argument that "in practice, however, to reduce complexity, it is often assumed that the current state is only related to the previous states" is incorrect. Especially in the case of

fluorescence lifetime decay, each pixel has long-term dependencies (entire decay), not just on previous and next states. Especially when dealing with a bi-exponential decay profile, the maximum signal part of the temporal point spread function (TPSF) is essential for a short lifetime, and the tail part of the decay is crucial for a long lifetime. That's why the difference in mean lifetime estimation is visible in Fig. S8. However, the model and architecture explanation show dependencies only on previous and next elements, ignoring long-term dependencies such as LSTM or GRU units do. I believe this model architecture is only suitable for processing and improving (up-sampling) video signals, which require only the previous and next frame to enhance resolution, but it is not suitable for lifetime TPSF data. I believe the author could have achieved better results by using a simple bi-directional LSTM model or transformer model.

[Reply] Thank you for your insightful comments. It is a mistake that “in practice, however, to reduce complexity, it is often assumed that the current state is only related to the previous states” in the previous response letter (not in the manuscript), was claimed for traditional RNN networks, and not for our proposed model. Our model does extract information from adjacent frames, but it also propagates this information over a long distance, as we mentioned in the article:

“The network employs a recurrent neural network architecture with two key components: (1) Forward and backward propagation branches to leverage temporal correlations and accumulate long-term spatiotemporal information from both past and future frames, and (2) Coupled propagation blocks that exchange information between the forward and backward branches, enabling each branch to incorporate context from the entire sequence. This coupled bidirectional propagation mechanism, along with the recurrent nature of the architecture, allows the model to effectively capture and preserve long-term temporal dependencies at each pixel location. Thereby, the extremely sparse photons (approaching zero) at the tail end of fluorescence lifetime decay can be precisely reconstructed, relying on the photon distribution pattern learned by the network at high photon time points. Additional elements of the network include aggregation blocks to concatenate the complementary bidirectional information, and high-level feature fusion to reconstruct missing details and reduce artifacts by refining the propagation and aggregation processes. ...”

“... The forward branch propagates frame features sequentially from the start to the end of the sequence. Each frame's features are computed by fusing information from the current frame and the features propagated forward from the prior frame. This allows accumulating contextual guidance from preceding frames. Conversely, the backward branch propagates frame features recursively from the end to the start of the sequence. Each frame's features are computed by fusing information from the current frame and the features propagated backward from the next frame. This maximizes the spatiotemporal information available to each frame for reconstruction, surpassing unidirectional or localized propagation schemes. ...”

We understand that in the case of fluorescence lifetime decay, each pixel has long-term dependencies, not just on previous and next states. We agree that when dealing with a bi-exponential decay profile, the maximum signal part of the temporal point spread function (TPSF) is essential for a short lifetime, and the tail part of the decay is crucial for a long lifetime. However, we would like to emphasize that our model architecture is designed to handle such long-term dependencies by propagating information over a long distance, effectively capturing the entire decay process. We understand your perspective that our model architecture might be more suitable for processing and improving video signals, which require only the previous and next frame to enhance resolution. However, we believe that our model is capable of handling fluorescence lifetime TPSF data as well. The propagating nature of our network branch allows it to capture the long-term dependencies present in the data, making it suitable for this application.

Regarding your suggestion to use a simple bi-directional LSTM model or transformer model, we appreciate your input. However, we would like to point out that our proposed model has been specifically designed and optimized for the task of fluorescence lifetime imaging, taking into account the unique characteristics and requirements of this application. We hope this clarification addresses your concerns and provides a better understanding of our model architecture and its applicability to fluorescence lifetime decay data.

As mentioned in point 4 of the earlier review, both raw data in 1i and 1j must be normalised, and their fitting (NLSF) must be plotted together (raw and fit for 1i and 1j). The reason for repeating this point is simple: in 1j, multiple points (at the peak of decay) are not part of the fitting, resulting in a poor chi-squared value of 1.53. Additionally, these decay curves should be plotted on semilogarithmic scales for better clarity. This exercise will highlight the error in lifetime estimation between fitting and the SparseFLIM method.

[Reply] We appreciate your continued guidance on improving the clarity and accuracy of our research. Based on the suggestion of another reviewer, we adopted the actual photon count instead of normalization to reflect the increase in photon count. Although in the previous response letter, we combined the fitting curves of Figures 1i and 1j together to verify their consistency. But for the unfitted discrete data points, they are easily confused when placed together, which is why we did not put them together in the manuscript before. We modified Figure 1 by putting them together and using semilogarithmic scales.

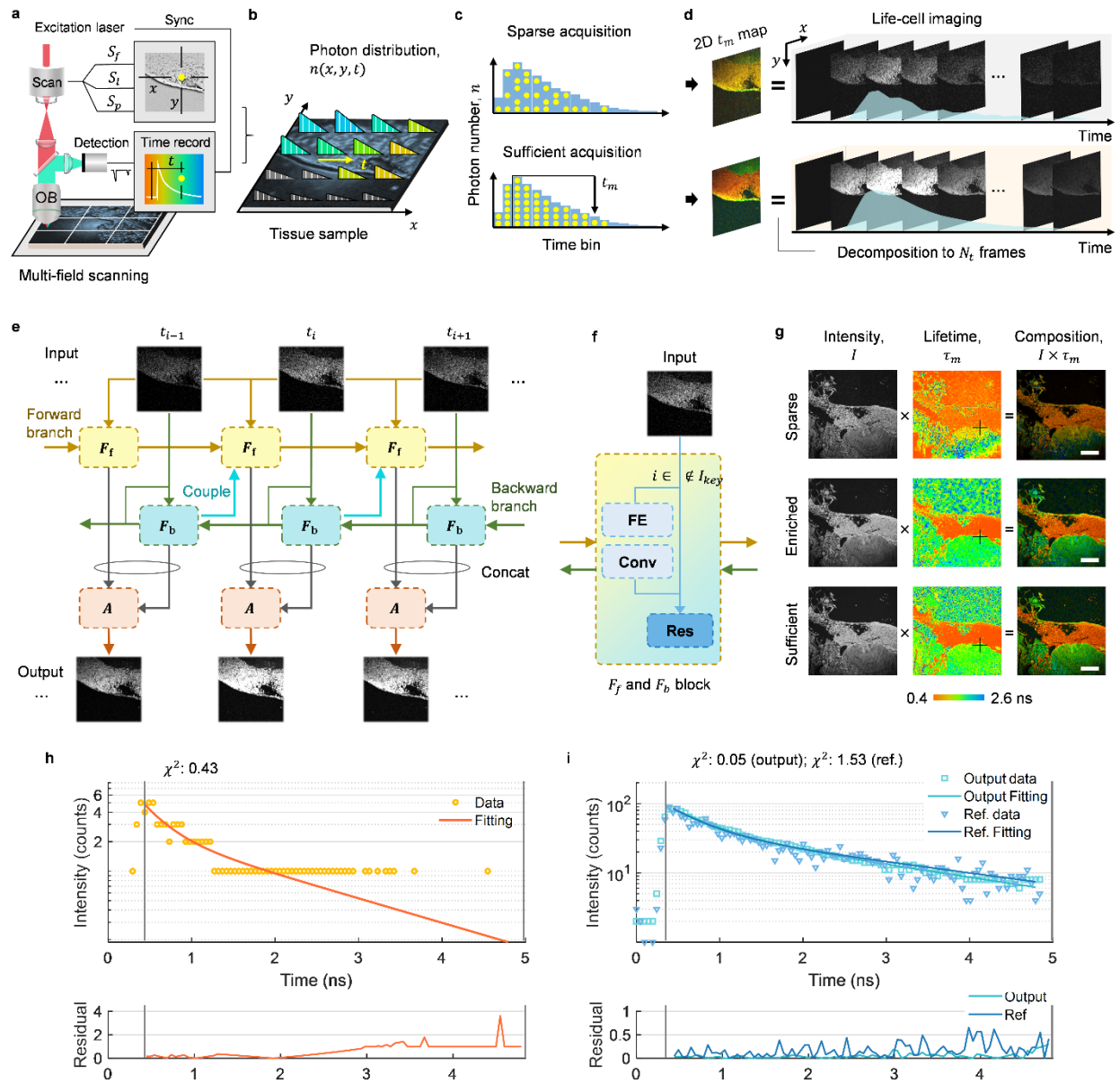


Fig. 1 General principle and validation of SparseFLIM. **a**, FLIM setup and data acquisition. **b**, Photon distribution in the x - y plane. **c**, Comparison of photon counts between sparse and sufficient acquisition modes. **d**, 2D mean lifetime (τ_m) images (left) and 3D lifetime stacks (right). **e**, Network architecture of SparseFLIM. t_i indicates the i th-frame time. F_f , forward propagation branch; F_b , backward propagation branch. A , aggregation blocks. F_f and F_b are coupled indicated by the cyan arrow line. **f**, Information refilling module. FE, feature extractor; Conv, convolution operations; Res, residual blocks **g**, Comparison of the sparse photon input, network output, and sufficient photon reference. **h**, Autofluorescence decay of the location indicated by the cross in **g** for the sparse photon acquisition. **i**, Data restored from the sparse photon recording using our network, which is consistent with the autofluorescence decay of the sufficient photon recording. The bottom panels correspond to the fitting residuals. Also, see Supplementary video 1. Scale bar, 100 μm .

In the explanation of the model, the authors mentioned a module called feature extractor (E), but its function has not been explained in the manuscript or supplementary materials. Simply referencing it without explanation is insufficient. The authors also mentioned temporal and spatial attention mechanisms. Where are the equations or parts of the model architecture where the attention mechanism is applied? Which query (Q), key (K), and value (V) were used for attention? Again, merely referencing it without sufficient explanation is inadequate.

[Reply] Thank you for your valuable feedback. The manuscript could benefit from clearer labeling of components and more details for feature extractor and attention. The EDVR paper described the application of feature extraction, temporal and spatial attention mechanisms. We have added that into the manuscript:

Modified Text:

“where E is the feature extractor. C is the convolution. I_{key} represents the key frame number. $\hat{h}$ is the result after information-refill. **Briefly, the feature extractor E uses strided convolution filters to downsample the input frames and generate multi-scale pyramid representations, transforming the raw input frames into a set of feature maps that are suitable for subsequent fusion operations.** For feature extraction of key frames, the significance of the information contained within intra-frame regions varies. Support key frames may exhibit artifacts such as blur, noise, or loss of signal photons. To address this issue, we also included the temporal and spatial attention mechanism to disregard irrelevant feature information while focusing on pertinent data for accurate reconstruction, which was detailed in EDVR³¹. **Briefly, the temporal attention map is computed between the extracted features e_i of a neighboring frame and the aligned features e'_i of the reference frame:**

$$h(e'_i, e_i) = \text{sigmoid}(\theta(e_i)^T \varphi(e'_i)) \quad (1)$$

where θ and φ are embedding functions implemented by simple convolution filters. In this context, the key and value are the embedded features $\varphi(e'_i)$ of the reference frame, while the query are the embedded features $\theta(e_i)$ of a neighboring frame. The temporal attention map h serves as the weights for the value, indicating how informative each neighboring frame's features are for reconstructing the reference frame.

After temporal attention is applied and features are fused, spatial attention masks are computed from the fused features using a pyramid design to increase the attention receptive field. These spatial attention masks are then used to modulate the fused features through element-wise multiplication and addition. For spatial attention, the query, key, and value are all derived from the fused features that have already gone through temporal attention."

Lastly, the reference key frames selected for this study [8, 10, 12, 16] might not be the same for all samples or other implementations. Does the user need to train the model for each experiment with different endo and exogenous dyes?

[Reply] It is insightful to raise this important point about the reference key frames and the potential need for retraining the model. The selection of reference key frames [8, 10, 12, 16] was specific to this study and may not be optimal for all samples or implementations. We chose these frames based on their highest attained SNR compared to other key frame selections, as mentioned in the "Training options" section of the Methods. Regarding the need to retrain the model for each experiment with different endogenous and exogenous dyes, we suggests some level of generalization but also acknowledges potential limitations:

1. Generalization: we demonstrates that the model showed "strong generalization across experimental modalities including multispectral FLIM and in vivo endoscopic FLIM." This suggests some degree of adaptability to different experimental conditions.

2. Limitations: however, in the Discussion section, we note: "Although the current results suggest a degree of transferability of the pre-trained weights for datasets sharing core characteristics, the generalization performance across broader datasets remains to be fully validated", "In cases involving significantly different optical setups (e.g. confocal microscopy), parameters (e.g. much fewer time bins), and sample types (e.g. fluorescent proteins in cells), it may be necessary to retrain the network for accurate reconstruction of lifetime decays."

"SparseFLIM generalized well across experimental modalities including multispectral FLIM and in vivo endoscopic FLIM with the pre-trained weights. ... While this work focuses on FLIM, the

core technique of learning from sparse data may generalize. ...Future work should focus on advancing the network, assessing more varied experimental scenarios, and clinical translation. For experiments with similar setups and fluorophores to those used in the study, the pre-trained model may work without retraining. However, for significantly different experimental conditions, optical setups, or fluorophores, retraining the model would likely be necessary to ensure accurate reconstruction of lifetime decays.”

In Figures S2 and S3, there are mismatches in the tau_m values of SparseFLIM and Photon-rich; what is the explanation for that?

[Reply] We appreciate your keen observation of the discrepancies in tau_m values between the SparseFLIM results and the Photon-rich reference in Figures S2 and S3. It’s important to acknowledge that achieving perfect consistency between deep learning network outputs and reference images is inherently challenging due to several factors: 1) Information are significantly influenced by the noise present in the raw input data. This noise can lead to variations in the reconstructed lifetime values. 2) It’s crucial to note that the Photon-rich reference images themselves are products of fitting procedures, which can introduce their own set of errors and uncertainties. Their tau_m values should not be considered as absolute ground truth values, but rather as approximate estimates. 3) While our SparseFLIM network is designed to reconstruct high-fidelity FLIM data, it cannot completely overcome the fundamental limitations imposed by sparse photon inputs and the stochastic nature of fluorescence emission.

The mismatches observed in tau_m values likely stem from a combination of these factors. The network is attempting to balance noise reduction, feature preservation, and adherence to learned patterns from the training data.

It’s also worth noting that these discrepancies don’t necessarily indicate poor performance of the SparseFLIM method. In fact, in some instances, the network’s reconstruction might be correcting for noise-induced artifacts present in the Photon-rich reference, potentially leading to more accurate lifetime estimates in certain regions.

Modified Text:

“... Despite small discrepancies in τ_m values between SparseFLIM results and Photon-rich reference in Fig. S2 and S3, achieving perfect consistency is challenging due to noise in raw input data, errors in fitting procedures used to create the Photon-rich reference images, and fundamental limitations of sparse photon inputs and fluorescence emission. The network aims to balance noise reduction, feature preservation, and adherence to learned patterns from the training data.”

Figure S5e represents background; what is this background? Which is present in the photon-rich section but missing in SparseFLIM. The authors must highlight the background region using pseudocolor to better understand the processing of the model. How is the model getting rid of the background information?

[Reply] We appreciate your careful examination of Figure S5e and the questions regarding the background. In this context, the background refers to background noise, reflected in the random fluctuation in Fig. 1i, where SparseFLIM achieved a smoother decay curve. To avoid confusion, we have changed it to noise. This noise is an inherent part of FLIM imaging and is captured in the photon-sparse and photon-rich data. The selected area in Figure S5e corresponds to the region on the sample containing biological tissue signal. As our primary focus is on the signal-containing regions, we did not specifically highlight or display the noise in pseudocolor. When processing sparse photon data, the model focuses on reconstructing and enhancing the signal-rich areas. In doing so, it effectively suppresses noise in low-signal regions. This noise suppression is a byproduct of the model's training on paired sparse and photon-rich data, where it learns to differentiate between meaningful signal and noise.

Modified Text:

“...Additionally, we noted the noise-only component at long lifetime fractions was effectively suppressed in the network outputs (Fig. S5c,e). This observation indicates additional denoising capabilities where the network learns to remove non-informative signal unrelated to fluorescent species of interest.”

Chi-squared values are missing in all other fits except Fig. S1, for example, Fig. S6.

[Reply] We have added missing Chi-squared values.

In conclusion, even after several reads, the article still lacks clarity in a number of sections, particularly the explanations of the model and architecture. I believe authors need to put greater effort into making them sound.

[Reply] We sincerely appreciate your thorough review and the time you have taken to read our article multiple times. Your feedback is invaluable in helping us improve the quality and clarity of our work. We have revised the model and architecture explanations according to the above comments and hope you are satisfied with the modifications.

Response (Reviewer #3)

The authors have addressed my concerns.

[Reply] Thank you for your feedback. We are glad to hear that we have successfully addressed your concerns. Your thorough review has undoubtedly contributed to improving this work on SparseFLIM. Thank you again for the highly positive comments on our work.