OFDVDnet: A Sensor Fusion Approach for Video Denoising in Fluorescence-Guided Surgery

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Many applications in machine vision and medical imaging require the capture of images from a scene with very low radiance, which may result in very noisy images and videos. An important example of such an application is the imaging of fluorescently-labeled tissue in fluorescence-guided surgery. Medical imaging systems, especially when intended to be used in surgery, are designed to operate in well-lit environments and use optical filters, time division, or other strategies that allow the simultaneous capture of low radiance fluorescence video and a well-lit visible light video of the scene. This work demonstrates video denoising can be dramatically improved by utilizing deep learning together with motion and textural cues from the noise-free video.

Keywords: Fluorescence guided surgery, video denoising, neural network denoising, deep learning, fluorescence video dataset, guided denoising

1. Introduction

In medical imaging modalities such as fluorescence image-guided surgery (FGS), thermal imaging, imaging of Cerenkov radiation in radiotherapy, or Raman scattering, there are many photon-starved or low contrast signals that are important to image at video frame rates. For example in FGS, a patient is injected with a fluorescent compound that binds to a specific tissue type. Then the tissue is illuminated with an excitation light source during surgery and the fluorescent marker emits a weak fluorescent signal that can be picked up by a fluorescent camera. Existing procedures use fluorescent markers to visualize tumors (Ishizawa et al., 2009; Zhang et al., 2017), blood vessels (Li et al., 2010), lymph nodes (Kitai et al., 2005; Frumovitz et al., 2018), necrotic tissue (Xie et al., 2015; Zajac et al., 2022), or nerves (Gibbs-Strauss et al., 2011). A surgeon will use the fluorescent camera feed to make real-time intraoperative decisions, where having high fluorescent signal is key. Fluorescence is much weaker than standard light sources and this low signal is compounded by the need for short exposure times for real-time video. Additionally, some useful fluorescent dyes and intrinsic fluorescence compounds often have low photon yields or low specificity leading to noisy videos that may lack the required sensitivity for clinical decision-making. For example, the auto-fluorescence from the inferior parathyroid can be weak during thyroidectomy (Kim et al., 2016) making it hard to detect the target. The low signal levels in FGS can be improved computationally through video denoising methods that will be able to take into account past captured frames to denoise a current frame as well as temporal and spatial information. Denoising methods will increase signal-to-noise levels in FGS to allow surgeons to lower the injection dose, allow for longer imaging time periods, and allow use of lower contrast and low sensitivity fluorescent agents even auto-florescence.

© 2023 Anonymous.

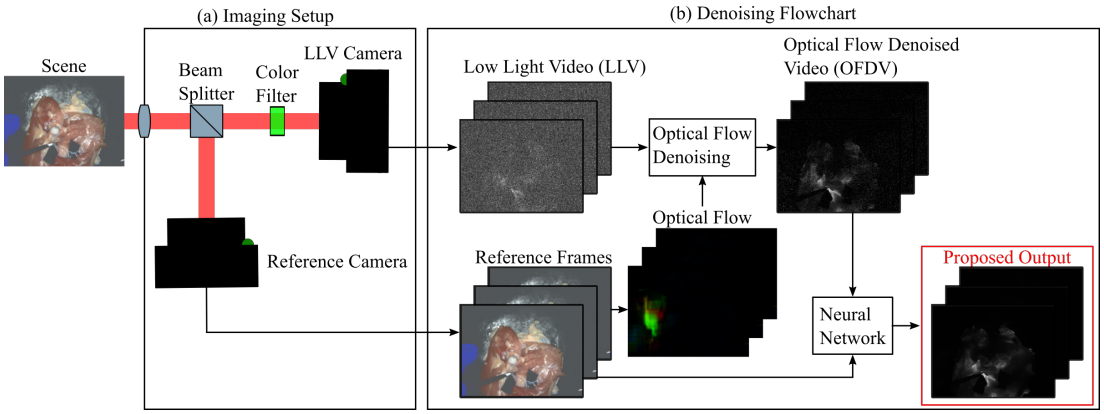

Figure 1: Overview: (a) Our imaging setup consists of two co-registered cameras: the reference camera measures a conventional intensity image of the scene while the low light video (LLV) camera measures a noisy fluorescence video. (b) To denoise the LLV, we first find optical flow throughout the video using the reference frames. We then apply the optical flow to the LLV to align frames in time, we merge the aligned frames to create the OFDV. The OFDV and reference frames are then fed into a video denoising neural network that will produce our final denoised video.

Due to the low brightness of fluorescent markers, fluorescence signals are flooded by background and ambient illumination. Hardware solutions have been developed to increase signal by filtering out ambient light. One method, called transient lighting (Velten et al., 2020), switches on and off the room light in short intervals between the exposure of fluorescence camera frames. The blinking light is unnoticeable to the human eye due to the flicker-fusion threshold making the surgical room appear well-lit while providing a dark environment for fluorescent capture. Another method is wavelength filtering which blocks photons from the illumination light source and photons outside of the fluorescence emission band. Both wavelength filtering and transient lighting lend themselves to a two-camera approach where one camera captures the low light video (LLV) of fluorescence, and another co-located camera captures a good quality RGB video called the reference video (RV).

In this work, we aim to recover a denoised video from a noisy low light video (LLV) of a faint fluorescent compound with noise standard deviations one magnitude higher than what current methods consider. To address this extreme noise, we propose a sensor fusion approach where the RV provides motion and structural cues that act as a guide when denoising the LLV. Due to the low signal level of the fluorescent compound we consider using a shot noise-limited camera such as a single photon avalanche diode (SPAD) to capture the LLV.

We identify 3 useful properties for FGS denoising, spatial, temporal, and RV correlation properties, many existing algorithms were only designed to use 2 of these factors. Video denoising methods, such as the non-learning based V-BM4D (Maggioni et al., 2012) and learning based FastDVDnet (Tassano et al., 2020), DVDNet (Tassano et al., 2019), and VNLNet (Davy et al., 2019) are meant to denoise videos without a RV so use only temporal and spatial properties. These algorithms rely on explicit or implicit feature mapping to properly denoise frames, which is difficult with the extreme noise considered in this paper. Additionally, convolutional neural network (CNN) approaches do not scale well to take as input a large number of video frames due to GPU memory limitations. A second class of denoisers are guided image denoisers such as guided filtering and joint bilateral filtering

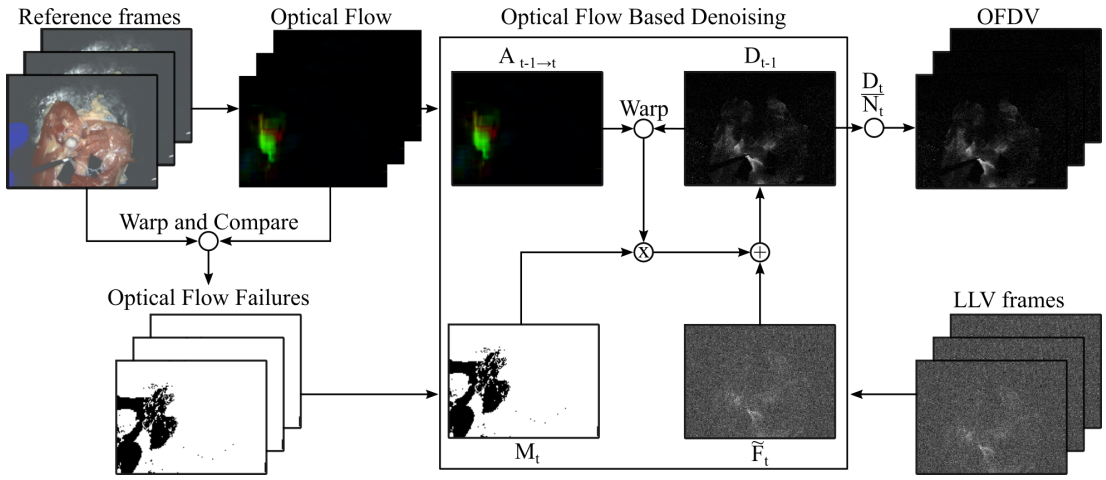

Figure 2: Optical Flow Denoising: First, we calculate optical flow from the RV frames. We then calculate an optical flow failure mask by warping the RV frames forward in time and comparing the warped frames to the true RV frame at that time. We then iterate over every frame and create a running sum of the LLV video. In one iteration we warp the previous running sum result forward in time 1 frame with the optical flow. We then reset pixels in the running sum according to the optical flow failure mask. The current LLV frame is then added to the result which becomes the next running sum frame and finally the OFDV.

that were designed for images with a guide image so make use of spatial properties and RV spatial correlations but no temporal information. A third category of denoisers are align and merge techniques (Hasinoff et al., 2016) which are often used to denoise videos generated by single photon cameras (Seets et al., 2021; Ma et al., 2020; Istvan et al., 2015; Gyöngy et al., 2017; Gyongy et al., 2018) and exploit temporal correlation; however, at extreme noise levels, alignment is difficult. To utilize all three properties, we propose OFDVDnet which draws inspiration from both align and merge techniques and deep learning methods in a two-stage approach. First, we align the LLV with motion extracted from the clean RV to compress the long temporal correlations in the LLV followed by a video denoising CNN based on FastDVDnet.

An overview of OFDVDnet is shown in Fig 1, first, our LLV and RV data is collected using two co-registered cameras. We use the RV to compute optical flow and an occlusion mask between successive frames. The optical flow and mask are then applied to the LLV to create a denoised motion-compensated estimate of the LLV, called the optical flow denoised video (OFDV). The OFDV incorporates information from many distant frames. One main advantage of the OFDV is it is a compressed representation of many different frames that can be easily loaded into GPU memory. Finally, we feed the ODFV and the RV frames into a video denoising neural network inspired by FastDVDnet to create a final denoised output.

We capture a RV and LLV aligned dataset using a commercial FGS imaging system (OnLume Surgical, Madison, WI) in a simulated surgery based on the blue blood chicken model used in micro-surgical training (Albano et al., 2021) where we inject a fluorescent agent to highlight blood vasculature in chicken thighs. To collect training data we use indocyanine green (ICG) as our fluorescent agent to capture low-noise video and simulate much weaker fluorescence giving us noisy and ground truth pairs for training.

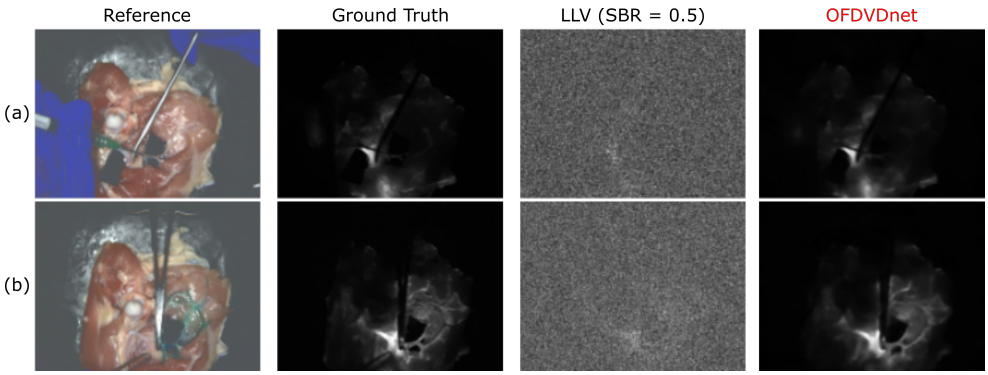

Figure 3: Injection of ICG: In this figure, we show the RV frame, ground truth fluorescence frame, noisy LLV frame, and our proposed denoised result of a scene before and after ICG injection. (a) At the start of ICG injection, there is a small amount of innate fluorescence near the injection site. (b) After injection, the vascular structure becomes visible in the fluorescence channel. OFDVDnet is able to reconstruct both scenes well, maintaining most of the important vascular structure in (b). Note that at high concentrations the dye appears green, but this color does not necessarily indicate the near infrared fluorescence from the ICG for example see the tip of the syringe in (a).

2. Method

2.1. Dataset

We capture data with a transient lighting-enabled (Velten et al., 2020), clinical wide-field FGS imaging system (OnLume Surgical, Madison, WI). OnLume's system uses two camera sensors for both the reference and fluorescent cameras. In a surgical training model (Albano et al., 2021), we inject the near-infrared fluorescent agent, ICG, via syringe into the femoral artery of chicken thighs (n=4) to simulate vascular surgery. We prepared varying doses of ICG up to the clinical guidelines of 2.5 mg/mL to generate fluorescent videos with visual contrast and low noise that can be treated as ground truth and simulate much lower fluorescence. In future work, we would like to detect markers that have much fewer photons than our captured fluorescent videos. We capture about 100 minutes of simulated surgical footage with a variety of motion such as cutting, pulling, squeezing, injecting, and working with surgical tools. The 100 minutes of footage is broken up into 590 100-frame long videos.

To simulate our noisy LLV, we scale the fluorescent frames between $[\phi_{\text{back}}, \phi_{\text{sig}} + \phi_{\text{back}}]$ photons per frame, where ϕ_{back} is the number of background photons and ϕ_{sig} is the maximum number of signal photons in a pixel, and then we apply Poisson noise to the scaled frames. We use $\phi_{\text{back}} = 10$ photons accounting for ambient light and sensor dark counts. We use 3 different signal photon levels, $\phi_{\text{sig}} = [1, 5, 20]$, to represent a range of fluorescent strengths. We measure the noise level with the signal-to-background (SBR) ratio, $\text{SBR} = \frac{\phi_{\text{sig}}}{\phi_{\text{back}}}$. Giving us three SBR levels, SBR$= [0.1, 0.5, 2]$ or standard deviations of $\sigma = [826, 180, 57]$ on an 8-bit image, see Appendix E.

2.2. Denoising Algorithm

Our goal is to denoise the LLV with the help of a co-registered noise-free RV. For our noise model, we consider the case where the LLV was captured with a photon-limited camera such as a SPAD and is dominated by Poisson noise (Hasinoff, 2014). While a pure Poisson noise model is a good approximation for single photon cameras, commercial CMOS cameras may introduce additional noise terms and next-generation CMOS cameras (e.g. Hamamatsu qCMOS) have low enough noise to measure single photons. While the noise model in our

method is interchangeable we choose to focus on the unavoidable Poisson noise coming from light quantization. At time t, let $\tilde{F}_t$ and W_t denote the captured LLV frame and RV frame, respectively. Let, $\tilde{F}_t = Poiss(F_t)$ where F_t represents the ground truth LLV frame, and $Poiss$ represents Poisson sampling. Because we consider very noisy cases it is difficult to denoise $\tilde{F}_t$ alone, so we use the RV as a guide.

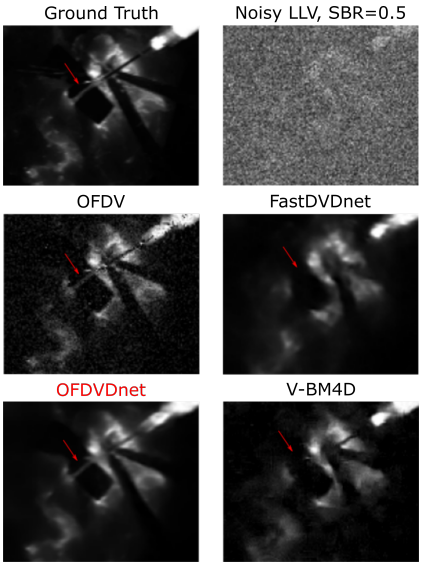

Figure 4: Vessel Detail: OFDVDnet correctly reconstructs a small vessel (red arrow) while comparison methods remove it entirely.

We use a two-step denoising method, first our optical flow based denoising exploits the alignment of the LLV and RV to find motion within the scene from the RV and apply it to the LLV. Once the LLV is aligned, frames can be merged to reduce noise to create the optical flow denoised video (OFDV). We warp the LLV both forward and backward in time aggregating information from all frames. However, if a video is needed to be displayed with little latency, then the OFDV could be warped only forward in time resulting in lower latency at the cost of quality. In our second step, a CNN takes the OFDV and RV frames as input to further denoise. This two-step approach uses frames from distant points in time to denoise a single frame without increasing GPU memory use.

Optical Flow Based Denoising: One technique to denoise $\tilde{F}_t$ is to average $2T$ frames around time t, $\tilde{F}_{t-T}...\tilde{F}_{t+T}$. Higher T will reduce noise but also increase motion blur. Therefore, we would like to spatially align the frames together before averaging. We use the RV to find optical flow and flow failure masks, which are then applied to the LLV to create the OFDV. The method is shown in Fig 2.

First, we use the RV frames to find optical flow warp maps, $A_{t\to t+1}$ that warp a frame at time t to time $t+1$. $A_{t\to t+1}$ can be found using any optical flow method. We found that Gunnar-Farneback's (Farnebäck, 2003) optical flow algorithm was sufficient for our problem and, on our hardware, ran faster than the CNN based MaskFlowNet (Zhao et al., 2020). We use $A_{t\to t+1}$ to average the LLV frames along motion trajectories with a running sum strategy. Let D_t^+ be the running sum at time t, then,

$$D_t^+ = A_{t-1\to t}(D_{t-1}^+) + \tilde{F}_t \tag{1}$$

Ideally, we could estimate a frame at time t as $\frac{D_t^+}{t}$. However, optical flow can fail in a variety of circumstances such as occlusions. To deal with flow failures, for each frame we detect pixels with flow errors and reset the summation for those pixels. Then due to the resets each pixel in D_t^+ may represent a sum over a different period of time, so we also record the length of time since the last reset in each pixel given by N_t^+. Therefore, the new LLV estimate at time t is given by $\frac{D_t^+}{N_t^+}$, which represents an average in each pixel since the last optical flow failure. Let M_t be a binary mask which has a pixel value of 0 if $A_{t-1\to t}$

fails to warp correctly, then Eq. 1 becomes,

$$D_t^+ = M_t \odot A_{t-1\to t}(D_{t-1}^+) + \tilde{F}_t \quad (2)$$

$$N_t^+ = M_t \odot A_{t-1\to t}(N_{t-1}^+) + \mathbb{1} \quad (3)$$

where $\odot$ represents pixel-wise multiplication and $\mathbb{1}$ represents an image of all ones. To find M_t we warp W_{t-1} and compare it to W_t by,

$$M_t = \left|1 - \frac{A_{t-1\to t}(W_{t-1})}{W_t}\right| > \tau \quad (4)$$

where τ is a threshold value, we use $\tau = 0.07$. We then compute, $\widehat{F}_t^{flow+}$,

$$\widehat{F}_t^{flow+} = \frac{D_t^+}{N_t^+} \quad (5)$$

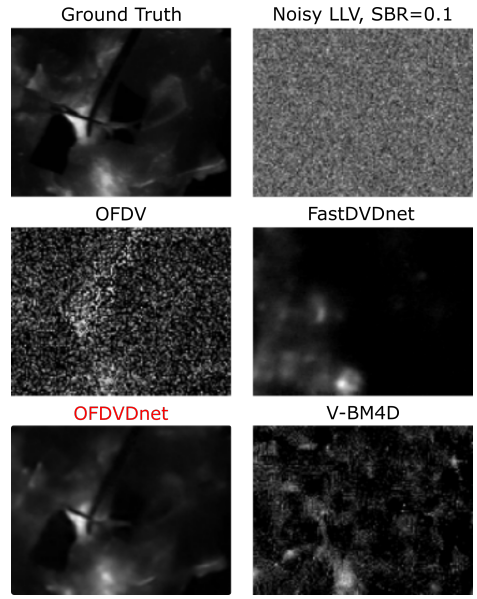

Figure 5: High Noise: This scene has no occlusions so OFDV averages over all frames giving good results where comparisons fail.

$\widehat{F}_t^{flow+}$ is the forward OFDV, we will combine $\widehat{F}_t^{flow+}$ with the backward OFDV $\widehat{F}_t^{flow-}$ to create out final OFDV. The backward OFDV is calculated by running the same process reversed in time. Similar to the forward OFDV, we calculate D_t^-, and N_t^- that represent the running sums on a time-reversed video. We can then create the final OFDV, $\widehat{F}_t^{flow}$, by combining the forward and backward estimations as follows,

$$D_t = D_t^+ + D_t^- - \tilde{F}_t \quad (6) \qquad N_t = N_t^+ + N_t^- - \mathbb{1} \quad (7) \qquad \widehat{F}_t^{flow} = \frac{D_t}{N_t} \quad (8)$$

where each pixel in D_t and N_t represents the sum and number of aligned pixels between two optical flow failures, respectively. Note that to find D_t and N_t we need to subtract out what is contained in both forward and backward OFDVs. A pixel value in the OFDV is the average value of an aligned LLV between optical flow failures. For example, if optical flow is correctly found for all frames, the OFDV will average along motion trajectories over the entire video, lowering noise and avoiding motion blur. Whereas if there is an occlusion the OFDV will avoid motion blur and only average pixel values between successive occlusion events. This process leads to the OFDV having high temporal consistency from frames being correlated and spatially varying noise levels from different averaging lengths per pixel.

Neural Network: In order to further remove the remaining noise and any warping artifacts in $\widehat{F}_t^{flow}$, we use a CNN denoiser based on FastDVDnet (Tassano et al., 2020) which takes five noisy frames to denoise the middle frame. We provide the CNN with five consecutive OFDV frames $\widehat{F}_{[t-2:t+2]}^{flow}$, averaging time maps $N_{[t-2:t+2]}$, and RV frames $W_{[t-2:t+2]}$ as input. We train the CNN to reconstruct the middle ground truth frame F_t using 1000 training pairs over 100 videos with mean square error (MSE), see Appendix C,D for details.

The four neighboring OFDV frames of $\widehat{F}_t^{flow}$, $\widehat{F}_{\{t-2,t-1,t+1,t+2\}}^{flow}$, provide the CNN with additional information on the center frame and reduce flickering. The averaging time maps

Table 1: Comparison of PSNR (higher is better) for LLV frames denoised using OFDVDnet, OFDV, and the comparison methods at 3 noise levels.

SBR	OFDVDnet	OFDV	FastDVDnet	V-BM4D	Guided Filtering	Joint Bilateral
0.1	**29.30**	10.80	24.25	19.73	16.40	15.81
0.5	**34.04**	21.48	30.76	29.86	28.13	26.33
2.0	**36.91**	30.77	35.66	36.73	33.72	31.49

$N_{[t-2:t+2]}$ act as noise maps to indicate per-pixel noise levels because OFDV pixels have averaged signal over a varying number of frames and thus different noise characteristics. RV frames $W_{[t-2:t+2]}$ let the CNN exploit structural similarities between the RV and OFDV.

3. Results

Our testing set consists of the middle 96 frames from 100 videos. We compare our results to video denoisers only given the LLV frames, the CNN FastDVDnet (Tassano et al., 2020) re-trained on our data, and V-BM4D (Maggioni et al., 2012), a popular block matching and filtering technique. We also compare our technique to two image denoising techniques that use a RV frame to assist in denoising a LLV frame, guided filtering (He et al., 2012), and joint bilateral filtering (Gastal and Oliveira, 2011). In Fig 3, we show an example scene before and after ICG injection into the vessels in a chicken thigh. Before injecting ICG, Fig 3(a), fluorescence is only seen near the injection site and in areas of the chicken thigh that are either auto-fluorescent or fluorescent due to the chicken treatment process. As the dye injects into the femoral artery, it perfuses and smaller vascular branches begin fluorescing. OFDVDnet is able to reconstruct most of the details of the vascular system with only slight blurring.

OFDVDnet performs well in high noise when the OFDV can average over a large number of frames allowing use of information in the entire video to create the denoised output. Fig. 5 shows an example scene (SBR= 0.1) with no occlusion events and only small motion that can be easily taken care of by the OFDV. OFDVDnet reconstructs most of the fluorescent structure correctly, but struggles to reconstruct the moving syringe fully due to large movement. Both FastDVDnet and V-BM4D fail in this high-noise example.

We found that OFDVDnet could better reconstruct small vessels, one example is shown in Fig. 4. In this example, the OFDV contains the blood vessel that has been injected with the marker (red arrow) so the denoising network has the information needed to properly reconstruct this important feature. Whereas the vessel is lost in the noisy LLV frames so the comparison methods, while showing a good looking image, completely remove this vessel. This failure mode is arguably worse than the un-processed noisy image since it suggests to the surgeon that the image is accurate but in fact conceals the vein.

By leveraging the correlation between successive OFDV frames, OFDVDnet produced videos with very little flickering. When our network denoises successive frames in a video the inputs are very similar, so the output will have little room to flicker due to noise fluctuations. Video results for OFDVDnet and the comparison methods can be found in Appendix B.

We use the standard Peak Signal-to-Noise Ratio (PSNR) as a quantitative quality metric. The PSNR results are summarized in Table 1. We calculate PSNR on each frame and average the results. OFDVDnet outperforms at all noise levels tested and OFDVDnet's PSNR drops slower with increasing noise when compared to other methods.

Network Ablation Study: We study the effects of the reference frames, averaging time maps, and OFDV on the performance of our network. We retain the network for each case at SBR= 0.5. The results are summarized in Table 2. First, we tested removing the averaging time maps or the RV frame inputs which both decrease the PSNR of the result. Next, we test using only the RV frames as input which obtains a PSNR of 24.13 indicating strong structural priors, such as visible veins, in the RV.

Table 2: Ablation study PSNRs.

SBR=0.5	PSNR
OFDVDnet	34.04
No RV Frames	33.82
No Averaging Time Maps	33.30
Only RV Frames	24.13
Switch OFDV with 5 LLV	31.05
Switch OFDV with 17 LLV	32.74

Finally, we replaced the OFDV input with the corresponding LLV frames which resulted in worse PSNR and significant flickering artifacts. We further studied the case of increasing the number of input LLV frames to 17, which was the maximum allowable due to GPU memory constraints. 17 LLV frames produced better quality results compared to five LLV. However, 17 LLV frames resulted in flickering artifacts and a decrease in PSNR compared to using five OFDV frames while also requiring a substantial increase in the required GPU memory (3x), training time (18x), and evaluation time (4x).

4. Conclusions and Discussion

In this work, we demonstrated a guided video denoising method meant for applications in FGS that is able to leverage deep learning in a memory-efficient manner with an explicit align and merge step. We captured and evaluated our method on a new dual-camera FGS dataset. OFDVDnet makes use of three properties spatial, temporal, and RV correlations to the LLV while comparison methods only use two of the three. FastDVDnet (Tassano et al., 2020) and V-BM4D (Maggioni et al., 2012) use spatial and temporal properties to provide decent results without the need for the RV, but fall off quickly as the noise level increases showing the importance of the RV at high noise. Guided filtering (He et al., 2012), and joint bilateral filtering (Gastal and Oliveira, 2011) make use of the RV, but only spatially, and use no temporal information. These methods have the worst PSNR and result in flickering in the final video at high noise levels but are computationally the least expensive. Finally, our intermediate OFDV only makes use of temporal information from the RV and its results show the importance of using spatial information to denoise.

While OFDVDnet is able to produce strong results, we identify three key areas that future work can improve upon. First, OFDVDnet relies on the time-consuming computation of optical flow between the RV frames which disallows real-time use. An efficient patch-based approach (Hasinoff et al., 2016) may be faster using the RV frames as a guide. Second, OFDVDnet struggles when strong motion or occlusion disrupts the averaging of the OFDV leading to higher noise results. Better motion tracking that deals with these cases will be useful in denoising more challenging scenarios. Finally, our dataset is limited to simulated chicken thigh data, so it is unclear how learned priors will translate to other applications such as oncology. While our dataset is useful for evaluation of new algorithms, new application-specific datasets will be required for learning-based approaches. We hope our dataset and method can be used in further algorithm development for medical imaging applications that require higher signal under scene motion.

References

Nicholas J. Albano, Weifeng Zeng, Christie Lin, Adam J. Uselmann, Kevin W. Eliceiri, and Samuel O. Poore. Augmentation of chicken thigh model with fluorescence imaging allows for real-time, high fidelity assessment in supermicrosurgery training. *Journal of reconstructive microsurgery*, 37(6), 2021. ISSN 0730-0301. doi: 10.1055/s-0040-1722184. URL https://doi.org/10.1055/s-0040-1722184.

Axel Davy, Thibaud Ehret, Jean-Michel Morel, Pablo Arias, and Gabriele Facciolo. A non-local cnn for video denoising. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2409–2413, 2019. doi: 10.1109/ICIP.2019.8803314.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *SCIA*, 2003.

Michael Frumovitz, Marie Plante, Paula S Lee, Samith Sandadi, J F Lilja, Pedro F. Escobar, Lilian T. Gien, Diana L. Urbauer, and Nadeem R Abu-Rustum. Near-infrared fluorescence for detection of sentinel lymph nodes in women with cervical and uterine cancers (film): a randomised, phase 3, multicentre, non-inferiority trial. *The Lancet. Oncology*, 19 10: 1394–1403, 2018.

Eduardo S. L. Gastal and Manuel M. Oliveira. Domain transform for edge-aware image and video processing. *ACM TOG*, 30(4):69:1–69:12, 2011. Proceedings of SIGGRAPH 2011.

Summer L. Gibbs-Strauss, Khaled A. Nasr, Kenneth M. Fish, Onkar Khullar, Yoshitomo Ashitate, Tiberiu M. Siclovan, Bruce F. Johnson, Nicole E. Barnhardt, Cristina A. Tan Hehir, and John V. Frangioni. Nerve-highlighting fluorescent contrast agents for image-guided surgery. *Molecular imaging*, 10(2):91–101, Apr 2011. ISSN 1536-0121. URL https://pubmed.ncbi.nlm.nih.gov/21439254.

I. Gyöngy, T. A. Abbas, N. Dutton, and R. Henderson. Object tracking and reconstruction with a quanta image sensor. In *Proceedings of the International Image Sensor Workshop*, 2017.

Istvan Gyongy, Neale AW Dutton, and Robert K Henderson. Single-photon tracking for high-speed vision. *Sensors*, 18(2):323, 2018.

Samuel W. Hasinoff. *Photon, Poisson Noise*, pages 608–610. Springer US, Boston, MA, 2014.

Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (TOG)*, 35(6): 1–12, 2016.

Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2012.

Takeaki Ishizawa, Noriyoshi Fukushima, Junji Shibahara, Koichi Masuda, Sumihito Tamura, Taku Aoki, Kiyoshi Hasegawa, Yoshifumi Beck, Masashi Fukayama, and Norihiro Kokudo. Real-time identification of liver cancers by using indocyanine green fluorescent imaging. *Cancer*, 115(11):2491–2504, 2009. doi: https://doi.org/10.1002/cncr.24291. URL https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.1002/cncr.24291.

Gyongy Istvan, Dutton Neale, Luca Parmesan, Davies Amy, Saleeb Rebecca, Duncan Rory, Rickman Colin, Dalgarno Paul, and Robert K Henderson. Bit-plane processing techniques for low-light, high speed imaging with a spad-based qis. In *International Image Sensor Workshop*, pages 1–4, 2015.

Sung Won Kim, Seo Hyun Song, Hyoung Shin Lee, Woong Jae Noh, Chulho Oak, Yeh-Chan Ahn, and Kang Dae Lee. Intraoperative Real-Time Localization of Normal Parathyroid Glands With Autofluorescence Imaging. *The Journal of Clinical Endocrinology and Metabolism*, 101(12):4646–4652, 09 2016. ISSN 0021-972X. doi: 10.1210/jc.2016-2558. URL https://doi.org/10.1210/jc.2016-2558.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Toshiyuki Kitai, Takuya Inomoto, Mitsuharu Miwa, and Takahiro Shikayama. Fluorescence navigation with indocyanine green for detecting sentinel lymph nodes in breast cancer. *Breast Cancer*, 12(3):211–215, 2005. doi: 10.2325/jbcs.12.211.

Yiwen Li, Ying Song, Lian Zhao, Gabriel Gaidosh, Alan M Laties, and Rong Wen. Direct labeling and visualization of blood vessels with lipophilic carbocyanine dye dii. *Nature protocols*, 2010. doi: https://doi.org/10.1038/nprot.2008.172.

Sizhuo Ma, Shantanu Gupta, Arin C. Ulku, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Quanta burst photography. *ACM Trans. Graph.*, 39(4), July 2020. ISSN 0730-0301. doi: 10.1145/3386569.3392470. URL https://doi.org/10.1145/3386569.3392470.

Matteo Maggioni, Giacomo Boracchi, Alessandro Foi, and Karen Egiazarian. Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 21:3952–66, 05 2012. doi: 10.1109/TIP.2012.2199324.

Trevor Seets, Atul Ingle, Martin Laurenzis, and Andreas Velten. Motion adaptive deblurring with single-photon cameras. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1945–1954, January 2021.

Matias Tassano, Julie Delon, and Thomas Veit. Dvdnet: A fast network for deep video denoising. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1805–1809, 2019. doi: 10.1109/ICIP.2019.8803136.

Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

A. Velten, A.J. Uselmann, S. Prajapati, J.S. Bredfeldt, T.R. Mackie, and K.W. Eliceiri. Transient room lighting for ambient light multiphoton microscopy. 2020. doi: 10.1101/2020.08.04.236364.

Bangwen Xie, Marieke A. Stammes, Pieter B. A. A. van Driel, Luis J. Cruz, Vicky T. Knol-Blankevoort, Martijn A. M. Löwik, Laura Mezzanotte, Ivo Que, Alan Chan, Jeroen P. H. M. van den Wijngaard, Maria Siebes, Sven Gottschalk, Daniel Razansky, Vasilis Ntziachristos, Stijn Keereweer, Richard W. Horobin, Mathias Hoehn, Eric L. Kaijzel, Ermond R. van Beek, Thomas J. A. Snoeks, and Clemens W. G. M. Löwik. Necrosis avid near infrared fluorescent cyanines for imaging cell death and their use to monitor therapeutic efficacy in mouse tumor models. *Oncotarget*, 6(36):39036–39049, Nov 2015. ISSN 1949-2553. doi: 10.18632/oncotarget.5498.

Jocelyn C Zajac, Aiping Liu, Adam J Uselmann, Christie Lin, Sameeha E Hassan, Lee D Faucher, and Angela Lf Gibson. Lighting the way for necrosis excision through indocyanine green fluorescence-guided surgery. *Journal of the American College of Surgeons*, 235(5):743—755, November 2022. ISSN 1072-7515. doi: 10.1097/xcs.0000000000000329. URL https://europepmc.org/articles/PMC9753148.

Ray R. Zhang, Alexandra B. Schroeder, Joseph J. Grudzinski, Eben L. Rosenthal, Jason M. Warram, Anatoly N. Pinchuk, Kevin W. Eliceiri, John S. Kuo, and Jamey P. Weichert. Beyond the margins: real-time detection of cancer using targeted fluorophore. *Nature reviews. Clinical oncology*, 2017. doi: https://doi.org/10.1038/nrclinonc.2016.212.

Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I-Chao Chang, and Yan Xu. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Appendix A. Additional Ablation Studies

Warp input frames: Explicit motion compensation could benefit video denoising quality (Tassano et al., 2019). We implemented explicit motion compensation on the input data to our CNN by warping four of the five frames in $\widehat{F}^{flow}_{[t-2:t+2]}$, $N_{[t-2:t+2]}$, $W_{[t-2:t+2]}$ to align with their respective middle frames: $\widehat{F}^{flow}_{t}$, N_t, and W_t using optical flow calculated from the reference frames. Warping did not improve video denoising quality for our case. This results in a PSNR of 33.92.

OFDV forward vs. forward backward OFDV forward(OFDV:FW) warps video frames forward in time, utilizing only information before the current frame in time space. OFDV:FWBW warps video frames both forward and backward in time, utilizing information from both before and after the current frame in time space.We compare the OFDV:FWBW to OFDV:FW by changing the inputs into the neural network at a variety of SBR levels. The results are

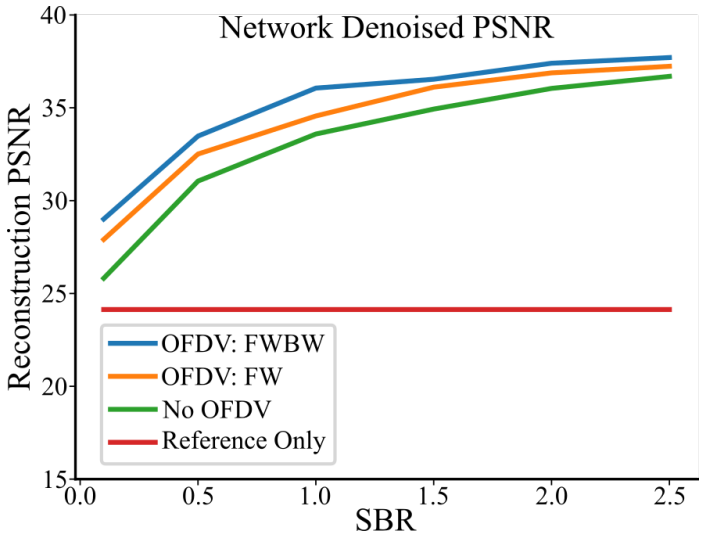

Figure 6: Reconstruction PSNR at different SBR: This plot shows our neural network reconstruction PSNR for a variety of SBR levels and different inputs into the network.

summarized in Fig. 6 along with PSNR curves for the only reference frames and replacing the OFDV with the LLV. We find the OFDV:FWBW obtains the best results due to its ability to use many frames from both the future and the past. Using the OFDV:FW achieves strong PSNR; however, more flickering is observed in the denoised results when compared to the OFDV:FWBW. This is likely due to stronger input correlations and lower noise in the OFDV:FWBW.

Appendix B. Video Results

We have included 3 videos from the test set at each noise level. In each video we include the reference video, ground truth LLV, noisy LLV, the OFDV, our result, FastDVDNet's(Tassano et al., 2020) result, and V-BM4D's(Maggioni et al., 2012) result. A brief description of each included video:

- Video 323: There is slight movement in this video from tool use.
- Video 383: In this video there is moderate movement due to pulling tissue.
- Video 499: The camera is bumped in this video leading to large motion in the scene.

We also include a video of denoised results from our ablation study. The result includes 25 test videos from the neural network denoised output with different input configuration. We use the following different inputs: OFDV:FWBW, OFDV:FW, 5 noisy LLV frames, 17 noisy LLV frames, and only reference frames. These results are at $SBR = 0.5$.

We share our video results on an anonymous YouTube channel: `https://www.youtube.com/@fgs_denoising/playlists`. For best viewing quality, we recommend enabling "fullscreen" by clicking the bottom-right button in the YouTube video player when viewing the videos. See the "About" tab in the channel for a short description on navigating the channel. The

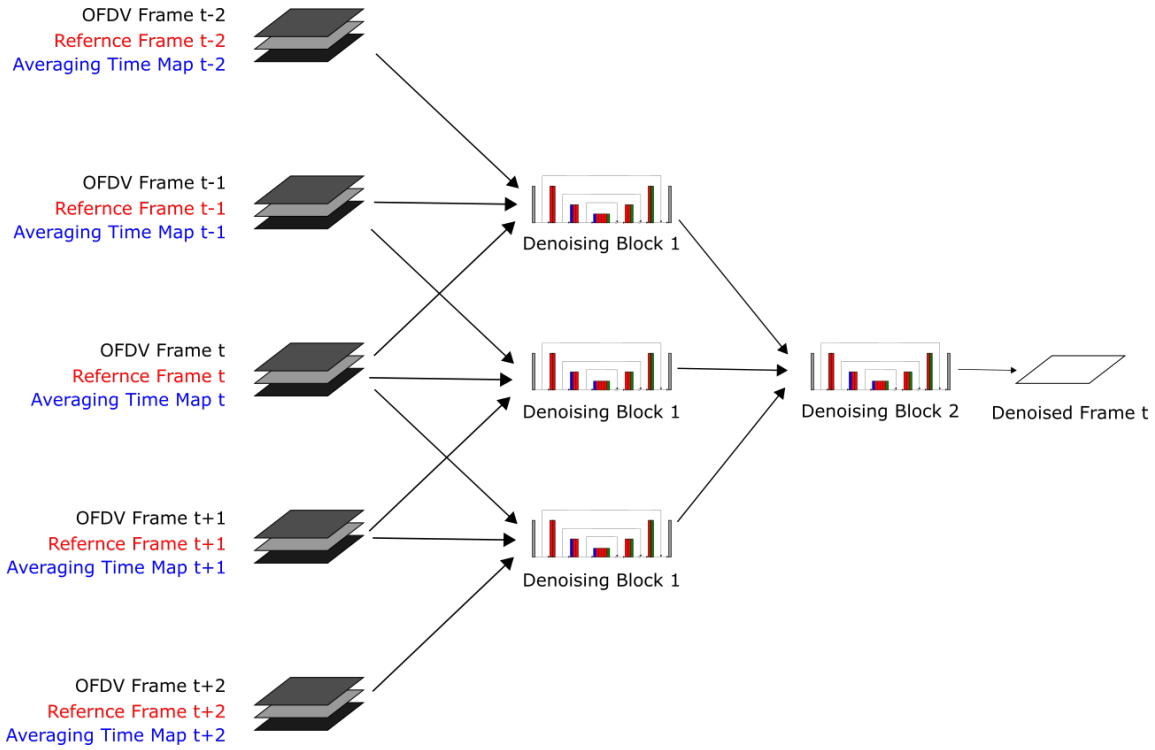

Figure 7: Diagram of the video denoising network

original videos can also be downloaded from an anonymous Google drive: https://drive.google.com/drive/folders/1QarTJx4h7TiVlR2zekRwcmqoV9fRL8Je?usp=sharing. The included .txt file provides a short description on how to navigate the folders.

Appendix C. Neural Network Architecture

We use the same general network architecture as FastDVDNet (Tassano et al., 2020) for our video denoising network. Suppl. Fig. 7 shows a diagram of the architecture. When denoising an OFDV frames $\widehat{F}_t^{flow}$, the network takes four of its neighboring OFDV frames $\widehat{F}^{flow}_{\{t-2,t-1,t+1,t+2\}}$ along with their corresponding reference frames $W_{[t-2:t+2]}$ and averaging time maps $N_{[t-2:t+2]}$ as inputs. During the forward pass, adjacent OFDV frames along with their corresponding reference frames and averaging time maps are passed into denoising blocks in groups of three. A diagram of the denoising block is shown in Fig. 8.

Appendix D. Training Details

Our CNN is trained on the output frames from the optical flow denoising algorithm or OFDV frames. The training dataset consists of input/ground-truth frame pairs P_t^k:

$$P_t^k = \left((\widehat{F}^{flow,k}_{[t-2:t+2]}, W^k_{[t-2:t+2]}, N^k_{[t-2:t+2]}), F_t^k \right)$$

where $k \in \{1, 2, \cdots, n_{\text{tot}}\}$.

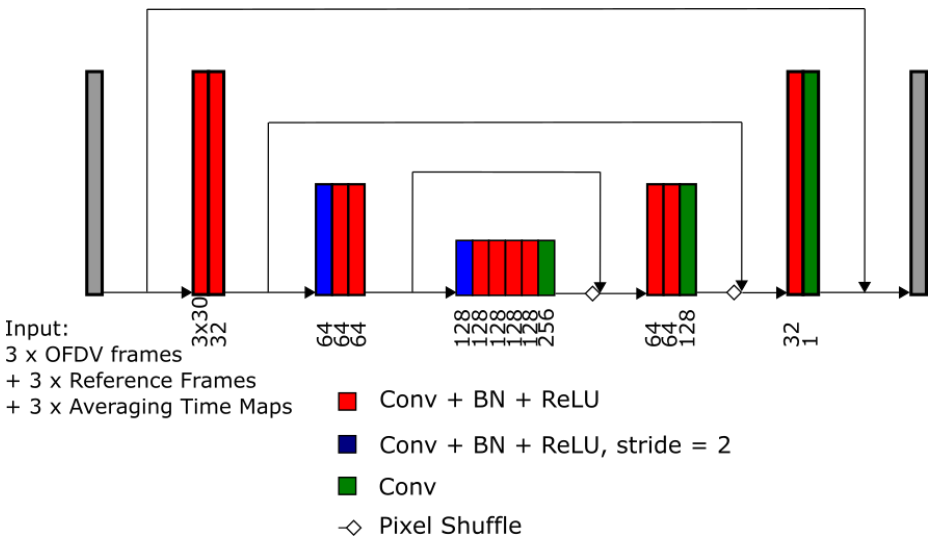

Figure 8: Architecture of the denoising block

For a given noise level, the training dataset contains a total number of $n_{tot} = 1000$ input/ground-truth pairs. We sample 1000 ground-truth fluorescent frames from the first 250 out of 590 ground-truth LLVs, and fill in their corresponding OFDV frames $\widehat{F}^{flow,k}_{[t-2:t+2]}$, reference frames $W^k_{[t-2:t+2]}$, and averaging time maps $N^k_{[t-2:t+2]}$.

We use Mean Squared Error (MSE) as our loss function:

$$L(\theta) = \frac{1}{n_{\text{tot}}} \sum_{k=1}^{n_{\text{tot}}} \|\widehat{F}^{net,k}_t - G^k_t\|^2 \tag{9}$$

where θ is the set of learnable parameters; $\widehat{F}^{net,k}_t$ is the output of the CNN.

ADAM optimizer(Kingma and Ba, 2015) is used to minimize the loss function with all its parameters set to default values and the initial leaning rate set to 10^{-3}. The network is trained for 100 epochs with a batch size of 8. Our CNN is trained separately for three different noise levels, SBR $= [0.1, 0.5, 2]$.

Retraining FastDVDNet: When comparing our method to FastDVDNet (Tassano et al., 2020) we found it was necessary to retrain FastDVDNet on our dataset and noise levels in order for FastDVDNet to produce reasonable results. We feed FastDVDNet the raw LLV frames and an estimated noise map. We estimate a constant noise map for each frame by using the average number of photon counts captured at each frame to estimate the noise standard deviation (see E) . We train FastDVDNet using the same training procedure as our method.

Appendix E. Finding Equivalent Gaussian Standard Deviation

Before we generate Poisson noise we scale our images between $[\phi_{\text{back}}, \phi_{\text{sig}}]$ then we add Poisson noise and re-scale the noisy image between [0,255]. For comparison to methods

using the standard deviation of additive white Gaussian noise, the comparable standard deviation of our Poisson corrupted images scaled between [0,255] is given by,

$$\sigma = \sqrt{\phi}\frac{255}{\phi_{\text{sig}}} \tag{10}$$

where $\phi \in [\phi_{\text{back}}, \phi_{\text{sig}}]$ is the expected numbers of photons for a given pixel. Although, Poisson noise is signal dependent we can get an estimate of the noise in a scene by using $\phi = \frac{1}{2}(\phi_{\text{sig}} + \phi_{\text{back}})$, which gives us a standard deviation of $\sigma = [826, 180, 57]$ for our three noise levels of SBR= $[0.1, 0.5, 2]$.