Enhanced Localization in Ultrafast Ultrasound Imaging through Spatio-Temporal Deep Learning

Vassili Pustovalov, Duong-Hung Pham, Denis Kouamé

Enhanced Localization in Ultrafast Ultrasound Imaging through Spatio-Temporal Deep Learning

Vassili Pustovalov
Université Toulouse III Paul Sabatier
IRIT, CNRS UMR 5505
Toulouse, France

Duong Hung Pham
Université Toulouse III Paul Sabatier
IRIT, CNRS UMR 5505
Toulouse, France

Denis Kouamé
Université Toulouse III Paul Sabatier
IRIT, CNRS UMR 5505
Toulouse, France

***Abstract*—The integration of ultrasound localization microscopy (ULM) in ultrasound imaging has enabled an unprecedented enhancement in resolution, offering insights into blood flow direction and velocity measurement. Despite its potential, ULM remains a complex and time-intensive procedure, continually improved through advancements in deep learning (DL) methods. Current DL techniques for micro-bubble (MB) super-localization encounter complexity stemming from the use of high-resolution images within their networks. Consequently, these convolutional neural networks (CNNs) incur longer execution times compared to traditional ULM, necessitating arbitrary filtering of their outcomes prior to integration with a subsequent tracking algorithm. To address these challenges, our study introduces a novel DL approach inspired by the success of single-molecule localization microscopy DL techniques. Our proposed 3D CNN, named 3DML-ResNet, enables fast and scalable super-localization while providing explicit estimations of the number of MBs in each frame.**

***Index Terms*—Super Resolution, Ultrasound Imaging, 3D CNN, Ultrasound Localization Microscopy, Temporal Context**

I. Introduction

Ultrasound localization microscopy (ULM) has achieved remarkable success in imaging microvascular structures down to a few micrometers. Standard ULM algorithms surpass the fundamental diffraction limit by first super-localizing micro-bubbles (MBs) individually and subsequently tracking them over thousands of ultrasound (US) temporal frames [1]. Nevertheless, these algorithms introduce a new trade-off between MB localization precision and acquisition time [2]. Employing low-concentration MB perfusions to achieve spatially sparse signals enhances MB localization accuracy and precision. However, the acquisition duration needed to ensure complete perfusion of vascular structures is increased. Conversely, a higher concentration of MBs enables the perfusion of more vasculature within a shorter acquisition window, albeit at the expense of increased overlapping MB signals, resulting in inaccurately reconstructed features. Addressing this new compromise in ULM is an active area of research in super-resolution US. Some CNN-based algorithms [3], [4] have been explored to enhance the localization of individual MBs at higher perfusion concentrations, aiming to shorten the acquisition time in ULM. While these algorithms demonstrate improved detection and super-localization of MBs in simulation data and promising results on *in vivo* data, their integration into the classic ULM pipeline remains problematic due to their poor alignment with the tracking step.

Furthermore, recent studies utilizing CNNs for US super-localization reveal two serious limitations. Firstly, these networks explicitly utilize images sized to match the final super-resolved image for MB super-localization on a frame-by-frame basis. Consequently, the time required for detection and super-localization, as well as the volume processed, increases exponentially with the resolution factor. Secondly, MB detection on a super-resolved pixel grid results in the detection of more potential MB than actually present. While this approach suffices when the microvascular is rendered by accumulating the estimated MB positions, it complicates the tracking step and may lead to artifacts that degrade flow direction and velocity estimations. To manage tracking complexity, post-processing filtering of detected bubbles is necessary to eliminate outliers. However, existing automatic methods proposed in the literature [5], [6] rely on local maximum detection algorithms, limiting the detection of overlapping MBs.

In this study, drawing inspiration from DECODE [7], a single-molecule localization microscopy DL technique, and the context-aware DL concept [8], we introduce an efficient 3D CNN for accurate single MB localization. This approach integrates the aforementioned post-processing filtering for the tracking step by explicitly estimating the number of MBs. Referred to as 3DML-ResNet hereafter, for 3D micro-bubble localization based on ResNet, our method distinguishes itself from existing CNN-based methods by processing image sequences inputs and incorporating temporal information from neighboring US frames for MB localization. Furthermore, we reduce the overall complexity of the CNN, facilitating faster execution speed and a reduction in processed data volume. The remainder of the paper is structured as follows. Section II introduces the main ingredients to build the proposed method, 3DML-ResNet. Subsequently, Section III presents rendering results on the PALA datasets [9], [10], demonstrating the improvement achieved by our approach over the most efficient algorithms for ULM to date. Finally, Section IV concludes and suggests directions for further research.

With financial support from ITMO Cancer of Aviesan within the framework of the 2021-2030 Cancer Control Strategy, on funds administered by Inserm.

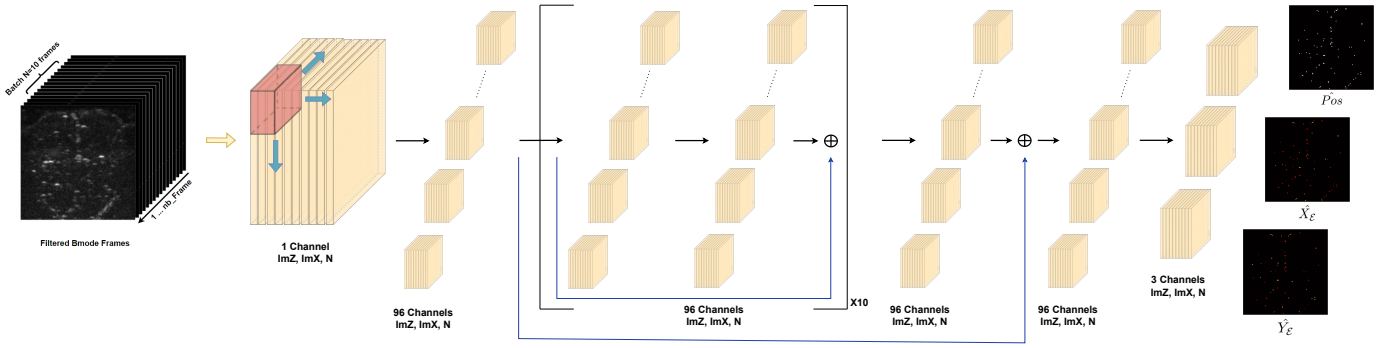


Fig. 1. **3DML-ResNet architecture.** The network processes a batch of normalized B-mode frames as input. Initially, a first 3D convolution layer is employed to generate 96 layers for feature extraction. Subsequently, high-frequency information is extracted using 10 consecutive residual blocks, each incorporating skip connections (illustrated by blue arrows). Within each residual block, two 3D convolution layers are applied, separated by a PReLU activation layer, and followed by a weighted element-wise sum with the input. The extracted feature information then passes through another 3D convolution layer before being combined with the layered network input via an element-wise sum. Finally, a reconstruction 3D convolution layer is utilized to reduce the number of channels to 3, thus matching the desired output representation. For each frame in the input batch, the output consists of three channels. One channel is the probability map $\hat{Pos}$ indicating the likelihood of an MB presence near each pixel. The super-localized coordinates of detected MBs are encoded in the subsequent channels as offsets from the pixel center ($\hat{X_\mathcal{E}}$ and $\hat{Y_\mathcal{E}}$).

II. Methods

A. Network Architecture

To develop an efficient CNN for MBs super-localization across multiple frames, we drew inspiration from the output representation and loss function of DECODE [7]. It introduced an efficient method for computing the super-localized coordinates of emitters in single-molecule localization microscopy. Rather than using a U-net structure as in [4], [7], which imposes constraints on input image sizes, we opted for a 3D ResNet architecture tailored for use with US data. By performing temporal context analysis using 3D convolution, we avoided replicating the time-consuming two-stage architecture of DECODE. Our 3DML-ResNet architecture, illustrated in Fig. 1. It is worth mentioning that 3D convolution kernels to analyze sequences of N images grouped in sets of 3 were employed.

Training a network to accurately detect MBs across various datasets poses a challenge due to the inherent biases in the training set. To tackle this issue, we opted to train a separate classification-style network instead of integrating the number of detected bubbles as a regularization term in the loss function during training. This classification-style network, composed of convolutional layers followed by fully connected layers, predicts the number of bubbles present in each ultrasound frame independently.

In optimizing the performance of our network, we found that rather than increasing the number of feature channels in the hidden layers, it was more effective to infer super-localization for each input US frame several time by shifting the batch of N images a single frame at a time rather than as a whole batch. This approach also enhances the accuracy of bubble position estimation for frames at the edge of the original batch.

B. Integration into the ULM pipeline

The conventional protocol of the ULM pipeline utilized in this study unfolds as follows: upon the injection of contrast agents (MBs), several thousand US frames are acquired. The MB signals are extracted by filtering the beamformed frames. Subsequently, individual MBs are detected and super-localized at a sub-pixel resolution precision. Following this, a tracking algorithm is applied to pair detected MBs across consecutive frames, eliminating inconsistent MB trajectories and establishing velocity profiles. The MB tracks are then interpolated to smooth the trajectories. Finally, the interpolated tracks are accumulated onto a high-resolution grid to render a super-resolved image of the microvasculature.

Our network aims to integrate directly into the ULM pipeline, replacing the detection and super-localization steps. Similar to most CNN-based methods, we combine these two steps within the network. The 3DML-ResNet takes a batch of N normalized B-mode images as input, processing them in groups of 3 frames to generate N 3-channels images. Simultaneously, the same batch of input images is fed into the second network, which outputs the estimated number of bubbles for each frame within the batch. Subsequently, the N_i MBs with the highest probability of existence in each US frame are selected based on the number of MBs determined by the secondary network. This approach effectively resolves the problem of detecting an excessive or insufficient number of bubbles when working on test datasets. Finally, the three filtered channels are converted into a list of coordinates in super-resolved pixels, directly usable by the tracking algorithms.

C. Loss function

We developed a new loss function as in Eq. (1) to train our network in a supervised manner to detect and super-localize MBs in spacetime. Unlike networks using super-resolved pixel grids, this loss function emphasizes accurate detection results at the expense of sub-pixel localization precision. It comprises a detection term L_{pos} and a localization term L_{loc}:

$$Loss = L_{pos} + L_{loc}, \tag{1}$$

The detection (or position) loss L_{pos} is computed on the first channel of the network output, indicating the probability of an MB existence on the low-resolution pixel. In [5], the ground-truth was dilated to encompass slight deviations in detection, enhancing convergence and learning stability. To achieve this, both the ground truth and the probability map of estimated MBs underwent convolution with a small Gaussian blur H:

$$L_{pos} = \left\| H \circledast Pos - H \circledast \hat{Pos} \right\|_F^2, \tag{2}$$

where Pos represents the ground-truth position of the MBs (with a probability of one), $\hat{Pos}$ denotes the inferred probability map, $\circledast$ denotes the 2D convolution operator, and $\|.\|_F^2$ represents the Frobenius norm.

The super-localization loss L_{loc} minimizes the difference between the sub-pixel position of the ground-truth and the estimated position for MBs detected in the first channel of the network. The ground truth is a sparse matrix where all coefficients not attached to an existing MB are zero.

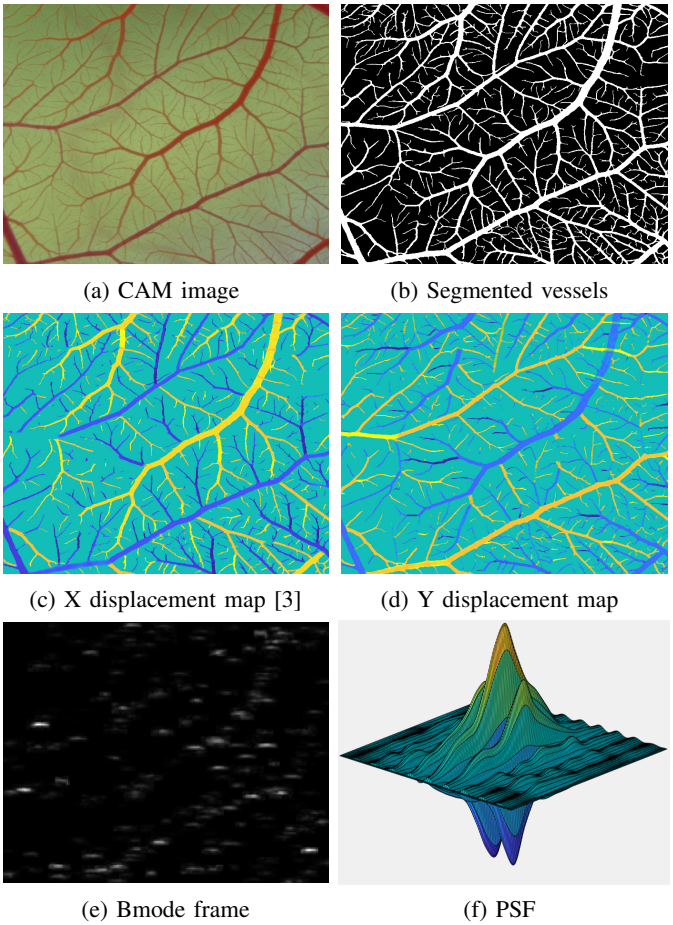
$$L_{loc} = \frac{1}{2} \left\| X_{\mathcal{E}} - \hat{X_{\mathcal{E}}} \right\|_F^2 + \frac{1}{2} \left\| Y_{\mathcal{E}} - \hat{Y_{\mathcal{E}}} \right\|_F^2, \tag{3}$$

where $X_{\mathcal{E}}$ and $Y_{\mathcal{E}}$ respectively denote the ground-truth sub-pixel coefficients for the vertical and horizontal axes, and $\hat{X_{\mathcal{E}}}$ and $\hat{Y_{\mathcal{E}}}$ represent the inferred values.

D. Simulation Data

Due to the unavailability of ground-truth data for supervised learning in *in vivo* ULM, we trained the network using simulated data generated from the optical images of chorioallantoic membrane (CAM) of chicken embryos from [2], with some adjustments made to the simulation process. While numerical simulation tools for US imaging, such as SIMUS and Field II, are available, we opted for the experimental Point Spread Function (PSF) simulation setup outlined in [2] to generate realistic images of MBs. This choice helps avoiding discrepancies between *in vivo* and *in vitro* PSFs.

Initially, a microvasculature mask was derived by segmenting CAM images, generating a sequence of MBs located within the mask for each instance. Subsequently, we extracted 200 MB image patches from the B-mode of the initial block of the rat brain beamformed data subset from [10] to generate a set of PSF. Since extracting the PSF solely from the B-mode image yields its envelope, we introduced an oscillation to the PSFs to achieve a more realistic representation. RF signals were generated through direct convolution (without grid approximation) between MBs and randomly selected PSFs from the set. Each individual MB maintained the use of the same PSF throughout its existence. A constant background was incorporated into the MBs RF signals, achieved through convolution between a simulated PSF (derived from L22-14v linear probe parameters) and scatterers randomly positioned from a normal distribution. The amplitude ratio between the MBs and the background was estimated from *in vivo* data. The interference between background scatterers and MBs results in slight variations in the PSF of each MB across different frames in a realistic manner.



(a) CAM image (b) Segmented vessels

(c) X displacement map [3] (d) Y displacement map

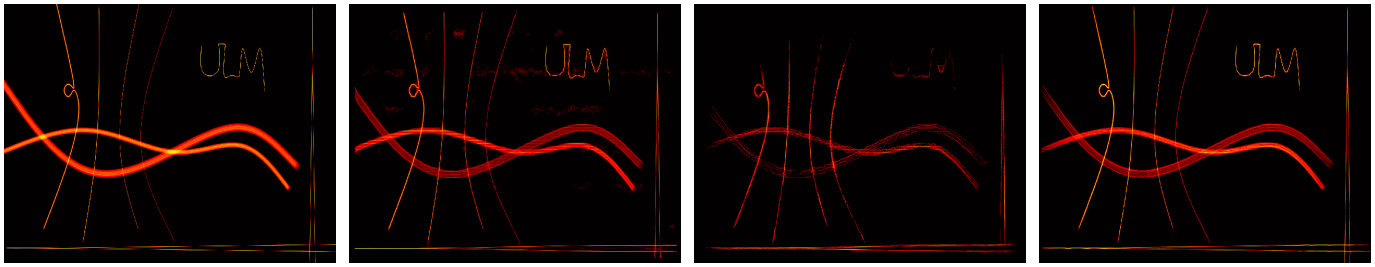
(e) Bmode frame (f) PSF

Fig. 2. Adapted numerical simulation procedure: (a) Cropped CAM optical image. (b) Segmentation result. (c) Vertical displacement map. (d) Horizontal displacement map. (e) Example B-mode frame. (f) An extracted PSF.

Rather than utilizing a directed vessel graph as in the original method, we opted to create displacement maps to simulate MBs motion (see Fig. 2). The motion was applied to the MBs locations before convolution, with a maximum flow velocity of 5 mm/s utilized in the simulation. Out-of-plane MB motion is not addressed in this study.

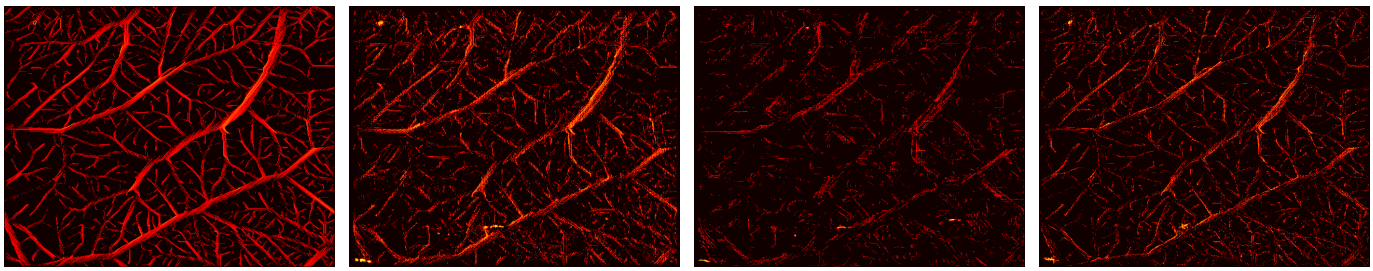
E. Training

A total of 20,000 US frames depicting MBs flowing through the extracted vasculature, each with dimensions of 250x250 pixels, were simulated for training purposes. The simulation utilized four distinct CAM images. Our network is trained to detect and super-localize MBs in ULM in a supervised manner. The ground-truth data is constructed from the simulated MBs list, comprising three channels: one for detection probability and two for sub-pixel localization in the axial and lateral directions of each MB. The detection probability in the ground-truth is binary, with values of either zero or one. Moreover, sub-pixel localization values are normalized to the $[0, 1]$ interval, ensuring that all CNN outputs remain positive.



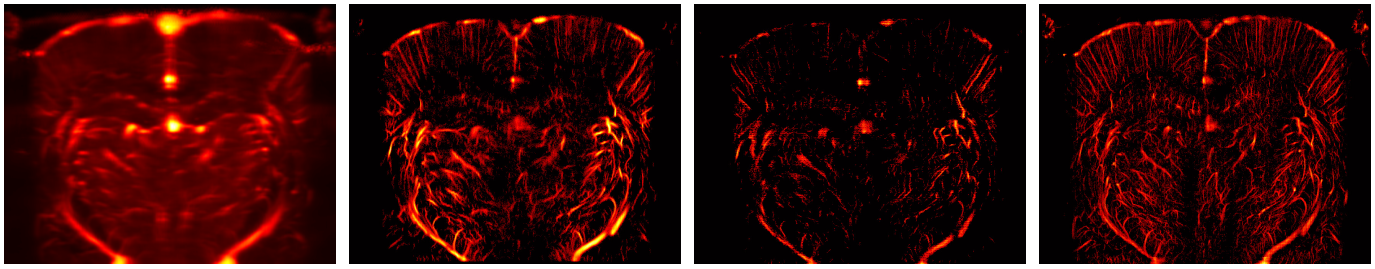
(a) Ground Truth (b) RS ULM [1] (c) mSPCN [3] (d) 3DML-ResNet

Fig. 3. Comparison of the rendering results obtained using Radial Symmetry ULM, mSPCN and 3DResNet super-localization methods on the *in silico* validation subset from [9]: (a) Ground Truth; (b) Radial Symmetry ULM [1]; (c) mSPCN [3]; (d) 3DML-ResNet (ours).



(a) Ground Truth (b) RS ULM [1] (c) mSPCN [3] (d) 3DML-ResNet

Fig. 4. Comparison of the rendering results obtained using Radial Symmetry ULM, mSPCN and 3DResNet super-localization methods on numerical simulations: (a) Ground Truth; (b) Radial Symmetry [1]; (c) mSPCN [3]; (d) 3DML-ResNet (ours).



(a) Power Doppler (b) RS ULM [1] (c) mSPCN [3] (d) 3DML-ResNet

Fig. 5. Comparison of the rendering results obtained using Radial Symmetry, mSPCN and 3DResNet super-localization methods on the rat brain *in vivo* set [10]: (a) Power Doppler; (b) Radial Symmetry ULM [1]; (c) mSPCN [3]; (d) 3DML-ResNet (ours).

During training, we employed the Adam optimizer over 200 epochs, with a batch size of 10 (each batch containing 10 US B-mode frames), and a learning rate of 0.001. Input normalization for our CNN was executed on a per-batch basis by dividing all values by the maximum value within the batch.

For our comparative study, we utilized the *in silico* subset of the PALA dataset [9], generated by the Verasonics Research Ultrasound Simulator, for both training and inference. This subset was partitioned into a test set comprising the initial 15 sequences and a training/validation set with a split ratio of 90% for training and 10% for validation, encompassing the last 5 sequences.

III. Results

Our 3DML-ResNet approach underwent validation using distinct test sets, including 20,000 frames of CAM simulation data, the *in silico* subset [9], and the complete rat brain *in vivo* dataset from [10]. We conducted a comparative analysis between our super-localization method, the Radial Symmetry (RS) ULM [1], and the mSPCN CNN [3]. Notably, results from the weighted average super-localization method are excluded as it consistently exhibited slightly inferior performance compared to RS in our evaluations.

Before applying localization methods to the rat brain *in vivo* data, clutter signals underwent filtering using a singular value decomposition and a bandpass Butterworth filter as proposed in [1]. However, no filtering was applied to the *in vivo* and *in silico* data as they were clutter-free. B-mode frames from different data subsets served as the input for the super-localization methods. Subsequent to the detection and super-localization of MBs, the PALA [1] tracking method with interpolation was employed to compute the MB tracks and accumulate them to produce the rendering results.

Unfortunately, we were unable to compare our method with the RF-ULM due to its ongoing development and the unavailability of the complete code. Nonetheless, we adopted its splitting ratio on the *in silico* data subset for training and testing to offer, at least, a visual basis for comparison.

A. Numerical simulation comparison

A comparative analysis was conducted, involving our 3DML-ResNet and other techniques on both the *in silico* and CAM test sets, as previously described. The rendering results, demonstrating a tenfold increase in resolution, are respectively showcased in Fig. 3 and Fig. 4. Quantitative comparison of the outcomes, presented in Table I and II, utilizes the Structural Similarity Index (SSIM) and Root Mean Square Error (RMSE) metrics.

Our approach exhibits superior performance in terms of these metrics on both the *in silico* data and simulated CAM data. Notably, our method produces results with the least noise and is the only one free of tracking artifacts, which manifest as rectilinear trajectories in the rendering, on the *in silico* data. This success is attributed to the accurate estimation of the number of MBs per frame by our secondary network, coupled with our loss function prioritizing MB detection accuracy over super-localization precision. Consequently, this reduces instances of undetected bubbles and prevents mismatches during the tracking phase. While the RS ULM method demonstrates the capability to reconstruct the complete simulated structure using the initial 15 sequences of the dataset, it exhibits some noise. Moreover, the mSPCN method encounters difficulties in accurately reconstructing the entire structure.

Results on CAM data exhibit more variability. While our method achieves superior metrics, it struggles to reconstruct all vessels accurately. Conversely, the RS ULM method successfully reconstructs most of the underlying structure, albeit with a noisy rendering. mSPCN faces considerable challenges in accurately reconstructing the structure in this scenario.

TABLE I
SSIM AND RMSE COMPARISON FOR THE ESTIMATIONS OF FIG. 3. BEST RESULTS ARE MARKED IN BOLD.

	ULM RS	mSPCN	3DML-ResNet
SSIM [%]	84.37	84.14	**91.40**
RMSE	7.043	9.440	**6.616**

TABLE II
SSIM AND RMSE COMPARISON FOR THE ESTIMATIONS OF FIG. 4. BEST RESULTS ARE MARKED IN BOLD.

	ULM RS	mSPCN	3DML-ResNet
SSIM [%]	40.34	42.02	**54.43**
RMSE	2.009	2.892	**1.764**

B. Comparison on Rat Brain subset

We utilize the complete set of 250 sequences from the *in vivo* rat brain dataset for this comparison. The rendering results, with a tenfold increase in resolution, are illustrated in Fig. 5. In the absence of available ground truth, we present the average Power Doppler calculated from each of the 250 sequences. We conduct a visual comparison of our method against RS ULM and mSPCN based on these rendered results. Qualitatively, only RS ULM and our 3DML-ResNet approach effectively capture the entire vascular network. RS ULM output exhibits the most prominent contrast, with minimal gridding artifacts. Conversely, our method detects numerous smaller vessels but is accompanied by a higher occurrence of gridding artifacts. Similar to the *in silico* data, we attribute the capability to visualize small vessels to the detection accuracy of our network. However, on *in vivo* data, this comes at the expense of less precise super-localization, leading to an increase in artifacts.

IV. CONCLUSION

This paper presented a novel approach leveraging advancements in CNN for ULM, combining and adapting various techniques to develop a new 3DML-ResNet method. Validation of these techniques on both simulation and *in vivo* data demonstrated at least comparable results to state-of-the-art methods. We illustrated the feasibility of utilizing output representations from optical microscopy to construct an efficient 3D neural network for MB super-localization. Moreover, we highlighted the importance of accurate MB detection in producing clean rendering images from the tracking step and imaging the microvasculature. We anticipate the widespread adoption of temporal context-aware network architectures, an emerging trend in CNN for ULM, as this provides an advantage not featured in traditional ULM. Our forthcoming studies will prioritize precise localization without compromising detection accuracy, aiming to mitigate gridding artifacts.

REFERENCES

[1] Baptiste Heiles, Arthur Chavignon, Vincent Hingot, Pauline Lopez, Eliott Teston, and Olivier Couture, "Performance benchmarking of microbubble-localization algorithms for ultrasound localization microscopy," *Nat. Biomed. Eng.*, vol. 6, pp. 605–616, 2022.

[2] Xi Chen et al., "Localization free super-resolution microbubble velocimetry using a long short-term memory neural network," *IEEE TMI*, vol. 42, no. 8, pp. 2374–2385, 2023.

[3] Xin Liu et al., "Deep learning for ultrasound localization microscopy," *IEEE TMI*, vol. 39, no. 10, pp. 3064–3078, 2020.

[4] Ruud J. G. van Sloun et al., "Super-resolution ultrasound localization microscopy through deep learning," *IEEE TMI*, vol. 40, no. 3, pp. 829–839, 2021.

[5] Léo Milecki et al., "A deep learning framework for spatiotemporal ultrasound localization microscopy," *IEEE TMI*, vol. 40, no. 5, pp. 1428–1437, 2021.

[6] Christopher Hahne, Georges Chabouh, Olivier Couture, and Raphael Sznitman, "Learning super-resolution ultrasound localization microscopy from radio-frequency data," in *2023 IEEE IUS*, 2023, pp. 1–4.

[7] Artur Speiser et al., "Deep learning enables fast and dense single-molecule localization with high accuracy," *Nat Methods*, vol. 18, no. 9, pp. 1082–1090, 2021.

[8] YiRang Shin et al., "Context-aware deep learning enables high-efficacy localization of high concentration microbubbles for super-resolution ultrasound localization microscopy," unpublished, 2023.

[9] Chavignon Arthur, Baptiste Heiles, Hingot Vincent, Lopez Pauline, Eliott Teston, and Couture Olivier, "OPULM PALA," 2020.

[10] Chavignon Arthur, Baptiste Heiles, Hingot Vincent, Lopez Pauline, Eliott Teston, and Couture Olivier, "In vivo rat brain for ultrasound localization microscopy: raw and beamformed data," 2023.