STUNet: A Spatiotemporal Deep Learning Approach for 7T fMRI Denoising

Thomas Alan Loboy Ramos[1,2], Ana Beatriz Solana[2], Brice Fernandez[2], Florian Wiesinger[2], Benedikt Wiestler[3], and Marion Irene Menzel[1,2,3]

[1] Technische Hochschule Ingolstadt, Ingolstadt, Germany
[2] GE HealthCare, München, Germany
[3] Technische Universität München, München, Germany
thomasalan.loboyramos@thi.de

Abstract. Ultra-high-field fMRI at 7 T enables submillimeter spatial resolution but is fundamentally limited by thermal noise, which constrains temporal signal-to-noise ratio (tSNR) and statistical power in task activation studies. NORDIC, the current standard for thermal noise suppression, performs local, patch-wise decomposition using both spatial and temporal redundancy within each patch, but requires complex-valued (magnitude and phase) data. We present STUNet (SpatioTemporal U-Net), a supervised deep learning model that denoises 7 T task fMRI using magnitude data only, eliminating the need for complex-valued acquisition. STUNet combines a 3D U-Net encoder-decoder with a bidirectional convolutional GRU bottleneck and a lightweight temporal mixing block to jointly exploit spatial and temporal redundancies within a sliding window of consecutive volumes, applying a global, learned, nonlinear spatiotemporal model rather than NORDIC's local, linear, patch-based decomposition. The model was trained on noisy–clean pairs derived from repeated in-house 7 T acquisitions across seven healthy volunteers and evaluated against NORDIC and unprocessed data on a held-out test scan from an unseen volunteer, paradigm, resolution, and repetition time. STUNet improved temporal SNR relative to unprocessed data (7.89 to 10.23, NORDIC: 12.98) and volume SNR (10.86 to 11.52, comparable to NORDIC's 11.39), and yielded activation sensitivity in AFNI-derived GLM maps with a peak GLM t-statistic improvement (29.92 to 31.66), closely comparable to NORDIC (32.12), though with spatially over-extended activation clusters relative to NORDIC, attributable in part to the total variation regularization term used during training. Beyond its quantitative performance, STUNet's magnitude-only design halves the per-session data and storage requirements relative to NORDIC, offering a practical advantage for research and clinical workflows in which phase data are not retained.

Keywords: 7T fMRI · Deep Learning · Image Denoising · Spatiotemporal · NORDIC

1 Introduction

Ultra-high-field fMRI at 7 T enables submillimeter spatial resolution, opening a window into cortical laminar and columnar organization that is inaccessible at conventional field strengths. However, this resolution comes at a fundamental cost: at submillimeter voxel sizes, thermal noise dominates over physiological noise, severely limiting temporal signal-to-noise ratio (tSNR) and statistical power in task activation studies [8, 9]. Effective post-acquisition denoising is therefore essential to fully exploit the sensitivity gains that 7 T offers.

The current state of the art for thermal noise suppression in fMRI is NORDIC [5, 9], which applies locally low-rank PCA in patch space to remove signal components that are statistically indistinguishable from thermal noise. NORDIC has demonstrated substantial tSNR improvements and increased activation sensitivity across a range of resolutions and acquisition protocols [5, 9]. NORDIC operates on overlapping spatiotemporal patches, exploiting redundancy across both a local spatial neighborhood and the timeseries within each patch; however, it requires complex-valued data, requiring both magnitude and phase images — data that are not always retained in standard research workflows, and that roughly double per-session storage requirements. Moreover, NORDIC's decomposition is local and patch-wise, applied independently patch by patch, rather than modeling spatial and temporal structure jointly and globally across the full image volume and timeseries.

Deep learning approaches have begun to address the fMRI denoising problem from complementary angles. Early work proposed temporal convolutional and recurrent architectures operating on individual voxel timeseries to suppress physiological noise in task fMRI [10], while other methods framed denoising as automated ICA noise-component classification using spatiotemporal CNNs [2]. A more recent approach, architecturally closest to our own, combined 3D convolutions with a convolutional LSTM to generate denoised fMRI volumes in a generative adversarial framework [11]. However, existing deep learning methods share important limitations in the present context: they are predominantly designed for resting-state or standard-resolution acquisitions rather than task fMRI at 7 T [10, 2], they do not target thermal noise specifically, and the 3DConv-LSTM approach [11] requires the full volumetric timeseries for training and inference, limiting its applicability to sliding-window or real-time settings.

We present **STUNet** (**S**patio**T**emporal **U**-Net), a supervised deep learning model for 7 T task fMRI denoising that applies a global, learned, nonlinear spatiotemporal model across the full image volume and timeseries jointly at inference, in contrast to NORDIC's local, patch-wise, linear decomposition, while also operating exclusively on magnitude data and eliminating the need for complex acquisition and the associated storage overhead. A bidirectional convolutional GRU bottleneck and a lightweight temporal mixing block enable the model to jointly exploit spatial and temporal redundancies within a sliding window of consecutive fMRI volumes. Trained on synthetic noisy–clean pairs derived from repeated in-house 7 T acquisitions, STUNet is evaluated against NORDIC on tSNR, peak SNR, and activation maps across motor and visual task paradigms at 0.8 mm and 1 mm isotropic resolution.

2 Methods

2.1 NORDIC

NORDIC was applied using the standard publicly available implementation [4] with default parameters, providing both magnitude and phase images as input to enable optimal noise estimation.

2.2 STUNet Architecture

STUNet is a 3D U-Net encoder-decoder augmented with a bidirectional convolutional GRU (ConvGRU) bottleneck and a lightweight temporal mixing block, jointly exploiting spatial and temporal redundancies in magnitude-only fMRI data (Fig. 1). The network takes a sliding window of $W = 7$ consecutive magnitude volumes as input, permuted to (B, W, Z, X, Y) for sequential processing. The encoder applies three levels of Conv3D–BatchNorm–ReLU blocks with in-plane max-pooling, preserving the slice dimension Z throughout to maintain compatibility with the recurrent bottleneck. Spatial dropout (rate = 0.1) is applied at encoder levels 2–3. At the bottleneck, a bidirectional ConvGRU (hidden size = 48, temporal kernel = 3) propagates hidden states along Z in both directions, fusing forward and backward sequences via a 1×1×1 convolution to integrate context across the full slice stack. The decoder mirrors the encoder with transposed convolutions and skip connections [6]. A final 1×1×1 convolution projects to W output channels, followed by a depthwise temporal Conv1d across W with Group Normalization and ReLU, refining temporal consistency without additional spatial parameters. At inference, a spatial sliding window of size W is applied along Z, retaining only the central slice per position; this train/inference discrepancy may introduce mild boundary effects, discussed in Section 4.

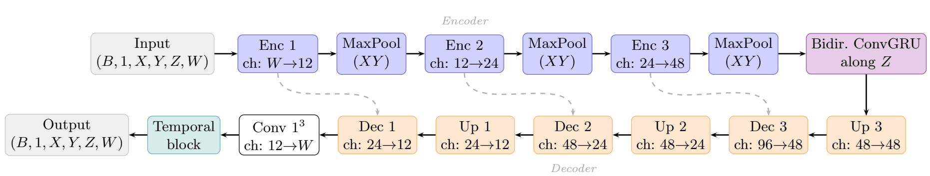

Fig. 1. STUNet architecture. Purple: encoder with in-plane max-pooling preserving Z. Pink: bidirectional ConvGRU bottleneck. Orange: decoder with transposed convolution upsampling and skip connections (dashed). Teal: temporal mixing block across W.

2.3 Data Acquisition

All data were acquired on a 7 T SIGNA scanner (GE HealthCare) using a 32-channel receive head coil. Seven healthy volunteers were scanned across multiple sessions, yielding a primary dataset of 35 scans spanning motor, visual, and sensory (brush tactile stimulation of left and right hand) task paradigms, at temporal resolutions of TR = 1 s or TR = 3 s, and spatial resolutions of 1.0 mm isotropic (whole-brain coverage) or 0.8 mm isotropic (reduced brain coverage), all acquired using a 2D gradient-echo EPI sequence. A subset of acquisitions included repeated runs of the same paradigm per volunteer, enabling the construction of high-quality cleaner reference volumes by averaging across repetitions, which served as training targets (see Section 2.4). Data acquisition is ongoing, and further volunteers will be incorporated into future versions of the dataset to improve statistical robustness.

All acquisitions were exported as magnitude, real, and imaginary components following DICOM-to-NIfTI conversion, making complex-valued data available for all scans. However, STUNet is designed to operate exclusively on magnitude images, eliminating the need for phase data at both training and inference time. Complex-valued data were retained solely to enable application of NORDIC as a reference method for comparison. Ethical approval was obtained and all participants provided written informed consent prior to scanning.

2.4 Training

Training pair generation. Two complementary strategies were used. Leave-one-out (LOO) pairs used each single repeat as the noisy input and the mean of the remaining $N-1$ repeats as the clean target, capturing real scanner noise statistics including the Rician distribution of magnitude MRI [1]. Synthetic pairs added simulated Rician noise (zero-mean Gaussian added independently to real and imaginary components, $\sigma \in [0.05, 0.20]$ of the local slice maximum) to the full N-repeat mean, yielding three augmented samples per group with a cleaner target than LOO alone. All volumes were globally min-max normalized across the training dataset to preserve relative intensity differences. The noise range $\sigma \in [0.05, 0.20]$ of the local slice maximum was randomly sampled per slice, chosen empirically to span from perceptible but mild degradation to severe noise corruption, ensuring the model is exposed to a range of noise levels during training rather than a single fixed level.

Loss function. The network was trained using a combined L1 and total variation (TV) loss [7]:

$$\mathcal{L} = \|f(x) - y\|_1 + \lambda \sum_{i,j,k} \sqrt{(\Delta_x f)^2 + (\Delta_y f)^2 + (\Delta_z f)^2} \tag{1}$$

where $f(x)$ is the network output, y is the clean target, $\Delta_x, \Delta_y, \Delta_z$ denote finite differences along each spatial dimension, and $\lambda = 3 \times 10^{-5}$ weights the TV regularizer relative to the L1 reconstruction term.

Hyperparameters. The model was trained for up to 95 epochs with a batch size of 4, using the Adam optimizer [3] with an initial learning rate of 1×10^{-4}, a 5-epoch linear warmup, followed by cosine annealing decay for the remaining epochs. Training was performed on a single NVIDIA A100 80GB PCIe GPU (3g.39gb partition), with 40 CPU cores and 300 GB RAM. Model selection across checkpoints was based on downstream functional metrics (tSNR, SNR and GLM-derived statistical measures of task activation) on a held-out external test volume rather than validation loss alone, as validation loss was observed to be a poor proxy for denoising quality at this dataset size. tSNR and SNR were tracked across checkpoints at 10-epoch intervals on the held-out test volume; epoch 60 was selected as it lies within the stable plateau of this trend rather than at a transient peak, balancing denoising performance against the risk of overfitting at later epochs. This checkpoint was used for all results reported in Section 3.

Dataset split. Groups with fewer than 2 repeats were excluded from training, as both the LOO and synthetic augmentation strategies require at least two acquisitions of the same paradigm. The held-out test scan belongs to a volunteer whose scans did not meet this criterion and were therefore absent from the training set entirely, ensuring the model has not been exposed to any data from this individual during training.

Of the 35 scans in the dataset, those belonging to volunteers with at least two repeats of the same paradigm were eligible for training; the remaining scans, belonging to volunteers with only single acquisitions, were excluded from both training strategies and constituted the held-out test set. Concretely, 26 scans from six volunteers entered training and validation (10% subject-level holdout for validation), while 9 scans from the remaining volunteer were excluded and used for evaluation only.

2.5 Evaluation

STUNet was evaluated on a held-out test set comprising scans not used at any stage of training or validation, a right-hand motor task fMRI scan of 1 mm isotropic resolution and $\mathrm{TR} = 3\,\mathrm{s}$. For all evaluations, three conditions were compared: unprocessed magnitude data, NORDIC-denoised data, and STUNet-denoised data. NORDIC [5, 9] was used as the primary baseline, applied to the held-out test scans using the standard publicly available implementation [4] with default parameters; NORDIC requires complex-valued (magnitude and phase) input, which was available for all test scans (Section 2.3).

Temporal SNR. Temporal signal-to-noise ratio (tSNR) was computed voxelwise as the ratio of the mean signal to its temporal standard deviation across the run. tSNR maps were averaged within a brain mask to yield a scalar summary per scan. As the primary measure of noise suppression, higher tSNR is expected following effective thermal denoising [9].

SNR. Volume signal-to-noise ratio (SNR) was computed as the ratio of mean signal intensity within a whole-brain mask to the standard deviation of signal in the background (non-brain) region of the same volume, providing a noise estimate independent of temporal variance and complementary to tSNR.

Activation maps. Task-based activation maps were generated using AFNI's `3dDeconvolve` general linear model (GLM) pipeline. For each condition, the initial four volumes were discarded to allow for signal stabilization, followed by motion correction (`3dvolreg`, rigid-body realignment to the first retained volume). The GLM design used a single stimulus regressor (right-hand motor response) convolved with a canonical SPM-style hemodynamic response function (`SPMG1`, 21 s duration), a polynomial baseline of automatically selected order (`-polort A`), and the six motion parameters from `3dvolreg` included as nuisance regressors of no interest. The same pipeline, design, and statistical threshold ($|t| > 3.39$, two-tailed $p < 0.001$, $df = 101$) were applied identically to all three conditions. Activation sensitivity was assessed by comparing peak GLM t-statistics and cluster extent across conditions, and spatial specificity was evaluated through visual inspection of activation overlays on anatomical reference images.

3 Experiments and Results

STUNet (base width 12, batch size 4, trained with in-plane flip and on-the-fly Rician noise augmentation) was selected as the configuration reported in this work. An earlier configuration trained without data augmentation and at higher model capacity (base width 16) achieved higher headline tSNR and SNR values, but exhibited a persistent train/validation L1 loss gap throughout training (a relative gap of approximately 15–20%, as observed in training logs, between final train and validation loss, versus under 5% for the configuration reported below) and spatial non-uniformity in tSNR maps, both consistent with overfitting on the limited training set; this configuration is therefore not reported further. The configuration below shows a markedly tighter train/validation gap and a checkpoint-stable improvement trend across epochs (Section 2.4), and is used for all results reported here.

Temporal SNR, SNR, and activation statistics. Table 1 summarizes tSNR, volume SNR, and peak activation t-statistics for the original, NORDIC-denoised and STUNet-denoised conditions on a single representative held-out test scan (1.0 mm isotropic, right-hand motor paradigm, TR = 3 s). This single-scan comparison is reported as an initial demonstration, and results should be interpreted with corresponding caution regarding generalizability; extending the evaluation to averaged metrics across the full held-out test set is planned as the dataset and evaluation pipeline continue to expand (Section 4).

STUNet improved tSNR relative to the unprocessed data ($7.89 \rightarrow 10.23$), indicating a reduction in temporal signal fluctuation consistent with effective thermal noise suppression, though more modest than the improvement achieved by NORDIC ($7.89 \rightarrow 12.98$). STUNet's volume SNR improvement ($10.86 \rightarrow 11.52$) was closely comparable to that of NORDIC ($10.86 \rightarrow 11.39$), as was its peak GLM t-statistic improvement ($29.92 \rightarrow 31.66$ for STUNet, versus $29.92 \rightarrow 32.12$

Table 1. Quantitative comparison of the original data and data denoised by STUNet v3 (Run B) and NORDIC. Performance was evaluated using whole-brain temporal signal-to-noise ratio (tSNR), volume SNR, and peak GLM activation t-statistics. Relative improvement over the original data is shown in parentheses.

Metric	Original	STUNet	NORDIC
Average tSNR	7.89	10.23 (+30%)	12.98 (+64%)
Volume SNR	10.86	11.52 (+6%)	11.39 (+5%)
Peak GLM t-stat	29.92	31.66 (+1.74)	32.12 (+2.20)

for NORDIC), indicating that STUNet achieves activation sensitivity within a comparable range to NORDIC, despite using only magnitude data.

Activation maps. Reported peak t-statistics and cluster metrics were computed after masking each condition's t-stat map with a brain mask derived from its own denoised functional volume, to exclude suprathreshold voxels falling outside brain tissue. On this representative scan, the brain-masked peak t-statistic for STUNet was 31.66, compared to 29.92 for the original data and 32.12 for NORDIC. Cluster analysis at the same statistical threshold ($|t| > 3.39$) showed that, as a percentage of each condition's own brain mask, suprathreshold voxels comprised 1.78% for the original data, 4.85% for NORDIC, and 7.35% for STUNet, with the largest cluster comprising 7,114 voxels (original), 21,276 voxels (NORDIC), and 26,540 voxels (STUNet). This indicates a real, though comparatively modest once restricted to brain tissue, increase in activation cluster extent for STUNet relative to both comparison conditions. This is discussed further in Section 4.

Visual inspection of the STUNet output (Fig. 2) revealed a mild oversmoothing effect relative to NORDIC and the original data, together with a faint horizontal banding pattern in the STUNet tSNR map. These observations are discussed further in Section 4.

4 Discussion

STUNet achieved meaningful tSNR and volume SNR improvements relative to unprocessed data, with activation sensitivity lower, but comparable to NORDIC despite requiring only magnitude input (Table 1). While NORDIC's tSNR gain was larger, STUNet's volume SNR and GLM sensitivity were closely matched, supporting its viability as a practical, lower-overhead alternative that approximately halves per-session storage and enables retrospective application to datasets where phase data were not retained.

The faint horizontal banding pattern visible in the STUNet tSNR map (Fig. 2) is consistent with the train/inference discrepancy along the slice dimension Z noted in Section 2.2: during training the network processes the full slice extent, whereas at inference a sliding window of size W is applied along Z with only the central slice retained per position, potentially introducing mild periodic boundary effects at window transitions. Addressing this discrepancy — for in-

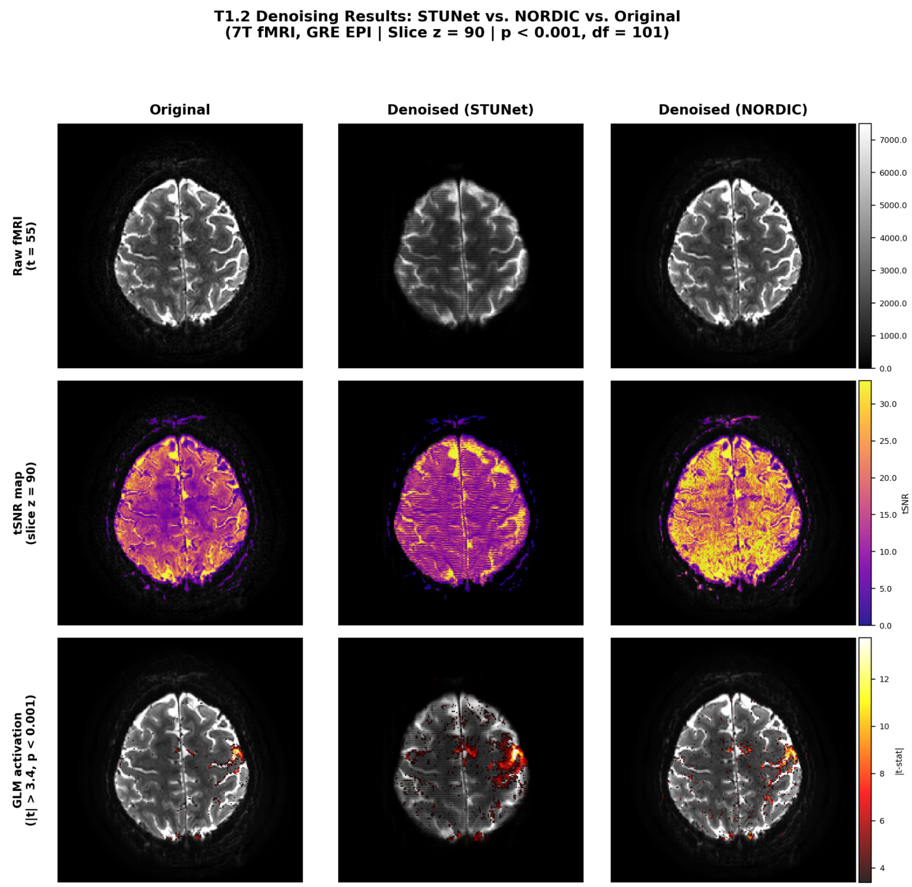

Fig. 2. Representative comparison of original, STUNet-denoised, and NORDIC-denoised data (left to right). Row 1: raw fMRI slice (t = 55, slice z = 90). Row 2: tSNR maps. Row 3: GLM activation maps overlaid on EPI ($|t| > 3.39$, $p < 0.001$, $df = 101$). Color scales are consistent across methods within each row.

stance through overlap-averaging of window outputs or a consistent Z-windowing strategy in both training and inference — is a planned next step.

The larger activation clusters observed for STUNet after brain masking (Section 3) could reflect a genuine sensitivity benefit, revealing task-related signal indistinguishable from noise in the unprocessed data. However, the spatial blurring already visible in the denoised images suggests that cluster inflation driven by smoothing plays at least some role. Three factors may likely contribute: the TV regularization term penalizes spatial gradients, encouraging piecewise-smooth outputs [7]; clean training targets were constructed by averaging repeated acquisitions without prior inter-run coregistration, so residual misalignment may have introduced blurring that the network learned to reproduce; and the dataset remains modest (seven volunteers, 35 scans), limiting the model's ability to learn fine spatial detail. Confirming whether additional suprathreshold voxels are anatomically plausible extensions of the motor cortex or scattered into unrelated regions would help distinguish genuine sensitivity gain from smoothing-driven inflation, and is an important direction for future evaluation.

This work is further limited by the use of a single representative scan for quantitative evaluation rather than averaged metrics across a fuller test set, and generalization to patient populations remains to be demonstrated. No statistical significance testing was performed, as the single-scan evaluation does not support meaningful inference about inter-subject variability. Planned next steps include: inter-run coregistration of training targets; a Fourier-domain loss term to better preserve high-frequency spatial detail [12]; a Gaussian-smoothed baseline to verify that tSNR gains reflect learned denoising rather than simple blurring; benchmarking against magnitude-only NORDIC; and expansion of the training and evaluation datasets.

STUNet is a global, learned, nonlinear spatiotemporal denoising model for 7 T task fMRI operating on magnitude data only. It achieves performance comparable to NORDIC while approximately halving storage requirements and removing the phase-acquisition constraint. Spatial smoothing is the primary current limitation and the focus of ongoing refinement.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Gudbjartsson, H., Patz, S.: The Rician distribution of noisy MRI data. Magnetic Resonance in Medicine **34**(6), 910–914 (1995). https://doi.org/10.1002/mrm.1910340618
2. Kam, T.E., Wen, X., Jin, B., Jiao, Z., Hsu, L.M., Zhou, Z., Liu, Y., Yamashita, K., Hung, S.C., Lin, W., Zhang, H., Shen, D.: A deep learning framework for noise component detection from resting-state functional MRI. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Lecture Notes in Computer Science, vol. 11766, pp. 754–762. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_84
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (2015), `https://arxiv.org/abs/1412.6980`
4. Moeller, S.: NORDIC_Raw: NORDIC for denoising in MRI. `https://github.com/SteenMoeller/NORDIC_Raw` (2021), accessed: 2025
5. Moeller, S., Pisharady, P.K., Ramanna, S., Lenglet, C., Wu, X., Dowdle, L., Yacoub, E., Uğurbil, K., Akçakaya, M.: NOise reduction with DIstribution Corrected (NORDIC) PCA in dMRI with complex-valued parameter-free locally low-rank processing. NeuroImage **226**, 117539 (2021). https://doi.org/10.1016/j.neuroimage.2020.117539
6. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

7. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena **60**(1–4), 259–268 (1992). https://doi.org/10.1016/0167-2789(92)90242-F
8. Triantafyllou, C., Hoge, R.D., Krueger, G., Wiggins, C.J., Potthast, A., Wiggins, G.C., Wald, L.L.: Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition parameters. NeuroImage **26**(1), 243–250 (2005). https://doi.org/10.1016/j.neuroimage.2005.01.004
9. Vizioli, L., Moeller, S., Dowdle, L., Akçakaya, M., De Martino, F., Yacoub, E., Uğurbil, K.: Lowering the thermal noise barrier in functional brain mapping with magnetic resonance imaging. Nature Communications **12**(1), 5181 (2021). https://doi.org/10.1038/s41467-021-25431-8
10. Yang, Z., Zhuang, X., Sreenivasan, K., Mishra, V., Curran, T., Cordes, D.: A robust deep neural network for denoising task-based fMRI data: an application to working memory and episodic memory. Medical Image Analysis **60**, 101622 (2020). https://doi.org/10.1016/j.media.2019.101622
11. Zhao, C., Li, H., Jiao, Z., Du, T., Fan, Y.: A 3D convolutional encapsulated long short-term memory (3DConv-LSTM) model for denoising fMRI data. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Lecture Notes in Computer Science, vol. 12267, pp. 479–488. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59728-3_47
12. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. In: IEEE Transactions on Computational Imaging. vol. 3, pp. 47–57 (2017). https://doi.org/10.1109/TCI.2016.2644865