Abstract #01941 ISMRM ABSTRACT

Hybrid 3D CNN-Transformer for Accelerated and Robust Cerebral Blood Volume (CBV) Reconstruction from PWI-MRI

Primary: Acquisition & Reconstruction - Image Reconstruction: AI 〉 **Secondary:** Analysis Methods - Classification and Prediction 〉 **Presentation:** Oral, PowerPitch Oral, Traditional Poster, Digital Poster 〉 **Keywords:** DEEP LEARNING RECONSTRUCTION AI-ACCELERATED CEREBRAL BLOOD VOLUME PERFUSION-WEIGHTED IMAGING

Akito Yamauchi [1], **Zongyu Li** [2,3], **Zhonghui Qie** [4], **Jia Guo** [2,5]

[1]Columbia University, New York, United States of America
[2]Biomedical Engineering, Columbia University, New York, United States of America
[3]Department of Biomedical Engineering, Columbia University, New York, United States of America
[4]Zuckerman Institute, Columbia University, New York, NY, USA, New York, NY, United States of America
[5]Psychiatry, Columbia University, New York, United States of America

Presenting Author: *Akito Yamauchi (ay2668@columbia.edu)*

Impact

Our hybrid 3D CNN-Transformer (3D ConvFormer) reconstructs motion-robust CBV maps from PWI using >2.5-fold fewer time-points (15 vs. 40). It provides superior artifact correction and quantitative accuracy compared to conventional methods and a baseline spatiotemporal CNN (ST-CNN).

Synopsis

Motivation: Conventional CBV reconstruction is susceptible to noise and motion, as it underutilizes the PWI time-series and relies on problematic uniform time-to-peak (TTP) assumptions.

Goals: We aimed to develop an AI architecture that leverages spatiotemporal dynamics to enable robust CBV reconstruction from significantly accelerated acquisitions.

Approach: We developed a Hybrid 3D CNN-Transformer (3D ConvFormer) that combines a 3D-CNN (local features) with a Transformer encoder (global spatiotemporal features) and compared it to a baseline ST-CNN.

Results: The 3D ConvFormer robustly reconstructed CBV maps using only 15 time-points (>2.5-fold acceleration), significantly outperforming the ST-CNN and actively mitigating motion-induced artifacts.

Introduction

Cerebral blood volume (CBV) is a key physiological marker for brain disorders such as stroke [1, 2], Alzheimer's disease [3, 4] and schizophrenia [5, 6]. Conventional CBV reconstruction from perfusion-weighted imaging (PWI) often requires a relatively long scan time (one hour for 40 time points) and is highly sensitive to motion artifacts. This stems from underutilizing signal dynamics, typically using only a few time points and incorrectly assuming a uniform time-to-peak (TTP) for contrast concentration [7]. We propose an AI-driven approach that learns from the entire spatiotemporal PWI series to predict the full contrast dynamics (defined as ΔR2*max and ΔR2*min). This strategy enables robust CBV generation from significantly accelerated acquisitions. We developed and compared two deep learning architectures: a baseline spatiotemporal CNN (ST-CNN) and our proposed Hybrid 3D CNN-Transformer (3D ConvFormer).

Methods

We used 175 PWI-MRI datasets (>=32 time points) from healthy mice (4-21 months) acquired on a 9.4T Bruker scanner (Fig. 1a). Raw DICOMs were converted to NIFTI, within-subject robust registration, and brain-extracted using an in-house AI based tool with manual verification.

To create a ground truth that captures the full signal dynamics without TTP assumptions, we calculated ΔR2* time series for each voxel. We extracted the maximum (ΔR2*max) and minimum (ΔR2*min) values from this curve to serve as the prediction targets for our models (Fig. 1b), representing the peak contrast effect and baseline. We developed and compared two models: 1) The ST-CNN (Fig. 1c), which processes 2D spatial patches (7x7) over time (T), using a 2D-CNN for spatial features followed by 1D-CNNs (dilated and regular) for local and global temporal information. 2) The 3D ConvFormer (Fig. 1d) processes 3D spatial patches (7x7x7) over time (T), first using a 3D-CNN for local feature extraction, followed by a Transformer encoder for global spatiotemporal integration. We replaced standard Layer Normalization with a dynamic tanh function to improve training efficiency [8]. The CLS token was fed into an MLP to predict both ΔR2*max and ΔR2*min. Models were trained and tested using 20 and 15 time points.

Models were trained in PyTorch (Adam optimizer, LR: 1e-3, batch: 256) on an NVIDIA Quadro 6000 GPU using Mean Absolute Error loss with a 100:1 weighting for ΔR2*max vs. ΔR2*min to account for magnitude disparities and our focus on ΔR2*max. Performance was evaluated for the whole brain using Pearson correlation coefficient (PCC), Spearman's correlation coefficient (SCC), Concordance correlation coefficient (CCC), normalized root mean squared error (NRMSE), Kullback–Leibler (KL) divergence, structural similarity index measure (SSIM), and peak signal-to-noise ratio (PSNR), with statistical comparisons via paired t-test.

Results

Figure 2 confirms significant variability in GBCA accumulation speed across subjects and brain regions. For some regions, GBCA even started to wash off towards the end of the scan, validating our ground truth approach that avoids a uniform TTP assumption. Qualitatively (Fig. 3a) all AI models accurately reconstruct ΔR2*max and ΔR2*min in motion-free data. However, in motion-corrupted data (Fig. 3b), the conventional ΔR2*max derived from 40 time points without motion correction showed severe laterality artifacts. In contrast, the 3D ConvFormer model trained on only 15 time points produced a significantly improved map, effectively correcting the motion artifacts. Quantitative analysis (Fig. 4 & 5) confirms that 3D ConvFormer significantly ($p<0.0001$) outperformed the ST-CNN across nearly all metrics (PCC, SCC, CCC, NRMSE, PSNR, SSIM) at both 15 and 20 time points.

Discussion

Our results demonstrate that AI-driven approaches, particularly 3D ConvFormer, provide high-quality CBV reconstruction while reducing scan time by over 2.5-fold (15 vs. 40 time-points). This acceleration provides a dual benefit: it inherently reduces the acquisition window for motion to occur, and its

spatiotemporal architecture actively mitigates motion artifacts (Fig. 3b). The 3D ConvFormer's superior performance to the ST-CNN highlights the importance of utilizing 4D spatiotemporal input and its hybrid design which combines local 3D convolutional features with a Transformer' global attention to model complex perfusion dynamics.

Conclusion

We proposed 3D ConvFormer, a hybrid 3D CNN-Transformer architecture, for accelerated and robust CBV reconstruction. It outperformed a baseline ST-CNN and provided more accurate, motion-robust maps than conventional methods on 9.4T mouse PWI, n=175 datasets, all while reducing required scan time by over 2.5-fold. This result demonstrates high potential for faster, more reliable perfusion imaging in clinical and preclinical settings.

References

1. Derdeyn, Colin P., et al. "Variability of cerebral blood volume and oxygen extraction: stages of cerebral haemodynamic impairment revisited." Brain 125.3 (2002): 595-607. DOI: https://doi.org/10.1093/brain/awf047

2. Röther, Joachim, et al. "Assessment of regional cerebral blood volume in acute human stroke by use of single-slice dynamic susceptibility contrast-enhanced magnetic resonance imaging." Stroke 27.6 (1996): 1088-1093. DOI: https://doi.org/10.1161/01.STR.27.6.1088

3. Alsop, David C., John A. Detre, and Murray Grossman. "Assessment of cerebral blood flow in Alzheimer's disease by spin-labeled magnetic resonance imaging." Annals of neurology 47.1 (2000): 93-100. DOI: https://doi.org/10.1002/1531-8249(200001)47:1%3C93::AID-ANA15%3E3.0.CO;2-8

4. Bozzao, Alessandro, et al. "Diffusion and perfusion MR imaging in cases of Alzheimer's disease: correlations with cortical atrophy and lesion load." American Journal of Neuroradiology 22.6 (2001): 1030-1036.

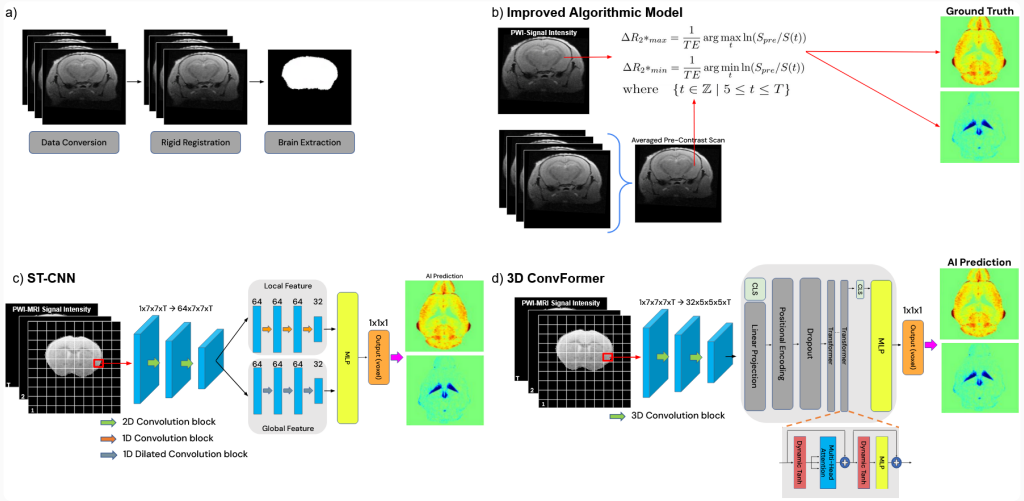
5. Lieberman, J. A., et al. "Hippocampal dysfunction in the pathophysiology of schizophrenia: a selective review and hypothesis for early detection and intervention." Molecular psychiatry 23.8 (2018): 1764-1772. DOI: https://doi.org/10.1038/mp.2017.249

6. Talati, Pratik, et al. "Increased hippocampal blood volume and normal blood flow in schizophrenia." Psychiatry Research: Neuroimaging 232.3 (2015): 219-225. DOI: https://doi.org/10.1016/j.pscychresns.2015.03.007

7. Kim, Seong-Gi, et al. "Cerebral blood volume MRI with intravascular superparamagnetic iron oxide nanoparticles." NMR in Biomedicine 26.8 (2013): 949-962. DOI: https://doi.org/10.1002/nbm.2885

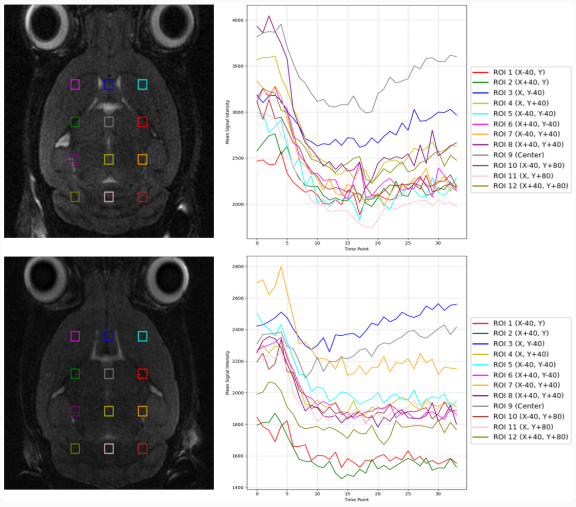
8. Zhu, Jiachen, et al. "Transformers without normalization." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025. DOI: https://doi.org/10.48550/arXiv.2503.10622

Figures and Tables



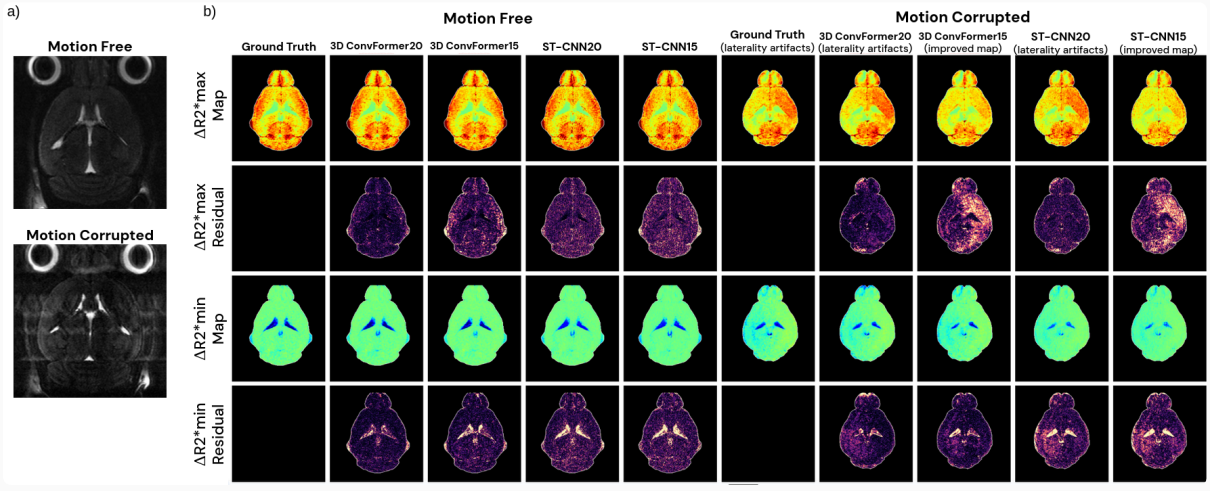
Scan for high-resolution version

Figure 1: *The experiment pipeline. (a) PWI data went through NIFTI conversion, registration, and brain extraction. (b) Ground truth (GT) generation, where ΔR2*max and ΔR2*min are extracted from the full voxel-wise ΔR2* time-series, avoiding uniform time-to-peak assumptions. (c) Baseline ST-CNN, which uses 2D-CNN for spatial and 1D-CNNs for temporal features. (d) Proposed 3D ConvFormer, which combines a 3D-CNN for local feature extraction with a Transformer encoder for global spatiotemporal integration.*



Scan for high-resolution version

Figure 2: *Mean PWI signal intensity curves from 12 distinct ROIs in two different subjects (top/bottom). The time-to-peak (signal nadir) varies significantly between regions within a subject and between subjects, and in some regions, contrast agents start to wash off towards the end of the scan. This inter- and intra-subject variability justifies the need for a ground truth method (detailed in Fig. 1b) that does not assume a uniform peak time.*



Scan for high-resolution version

Figure 3: *(a) Example motion-free (top) and motion-corrupted (bottom) PWI scans. (b) Comparison of ΔR2*max/min and residual maps (jet/ magma colormap). In motion-free data, all models accurately reconstruct the ground truth. In motion-corrupted data, the full-length (40-point) ground truth map suffers from severe laterality artifacts. In contrast, the 3D ConvFormer model trained on only 15 time-points successfully mitigates these artifacts, producing more robust and accurate maps than the ST-CNN models.*

a)

Model	PCC ↑	SCC ↑	CCC ↑	NRMSE ↓	KL ↓	PSNR ↑	SSIM ↑
Delta R2 Max							
3D ConvFormer20	**0.921**	**0.933**	**0.906**	**0.007**	**0.017**	**43.838**	**0.977**
ST-CNN20	0.887	0.887	0.874	0.008	0.020	41.909	0.959
3D ConvFormer15	**0.875**	**0.880**	**0.841**	**0.009**	0.057	**41.408**	**0.955**
ST-CNN15	0.836	0.830	0.806	0.010	**0.054**	40.378	0.940
Delta R2 Min							
3D ConvFormer20	**0.942**	**0.901**	**0.937**	**0.016**	**0.024**	**36.911**	**0.917**
ST-CNN20	0.875	0.769	0.844	0.024	0.097	33.287	0.812
3D ConvFormer15	**0.923**	**0.896**	**0.911**	**0.018**	**0.021**	**35.595**	**0.895**
ST-CNN15	0.848	0.734	0.805	0.027	0.214	32.447	0.772

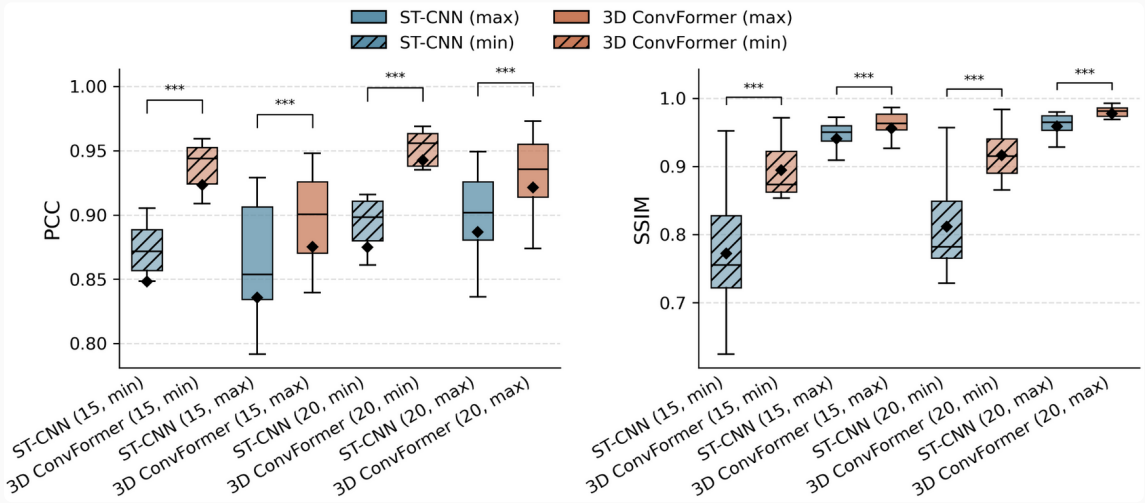
b)

Metric	15 time points	20 time points
PCC	$p < 10^{-5}$ ***	$p < 10^{-7}$ ***
SCC	$p < 10^{-7}$ ***	$p < 10^{-8}$ ***
CCC	$p < 10^{-4}$ ***	$p < 10^{-7}$ ***
NRMSE	$p < 10^{-5}$ ***	$p < 10^{-4}$ ***
KL	$p = 0.729$ ns	$p = 0.313$ ns
PSNR	$p < 10^{-5}$ ***	$p < 10^{-5}$ ***
SSIM	$p < 10^{-5}$ ***	$p < 10^{-4}$ ***

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Scan for high-resolution version

Figure 4: *(a) Table of performance metrics for 3D ConvFormer and ST-CNN at 15 and 20 time-points for predicting ΔR2*max and ΔR2*min calculated for the whole brain. 3D ConvFormer outperforms the ST-CNN in nearly all cases. (b) Statistical analysis (paired t-test) confirms the 3D ConvFormer is significantly (p<0.001) superior to the ST-CNN across most key metrics, including PCC, SCC, CCC, NRMSE, PSNR, and SSIM.*



Scan for high-resolution version

Figure 5: *Box plots for PCC (left) and SSIM (right) metrics evaluated for the whole brain confirms that 3D ConvFormer consistently outperforms the ST-CNN across all scenarios (15/20 time-points, max/min prediction). This suggests the superiority of using 4D spatiotemporal input and Transformer in global spatiotemporal feature extraction.*