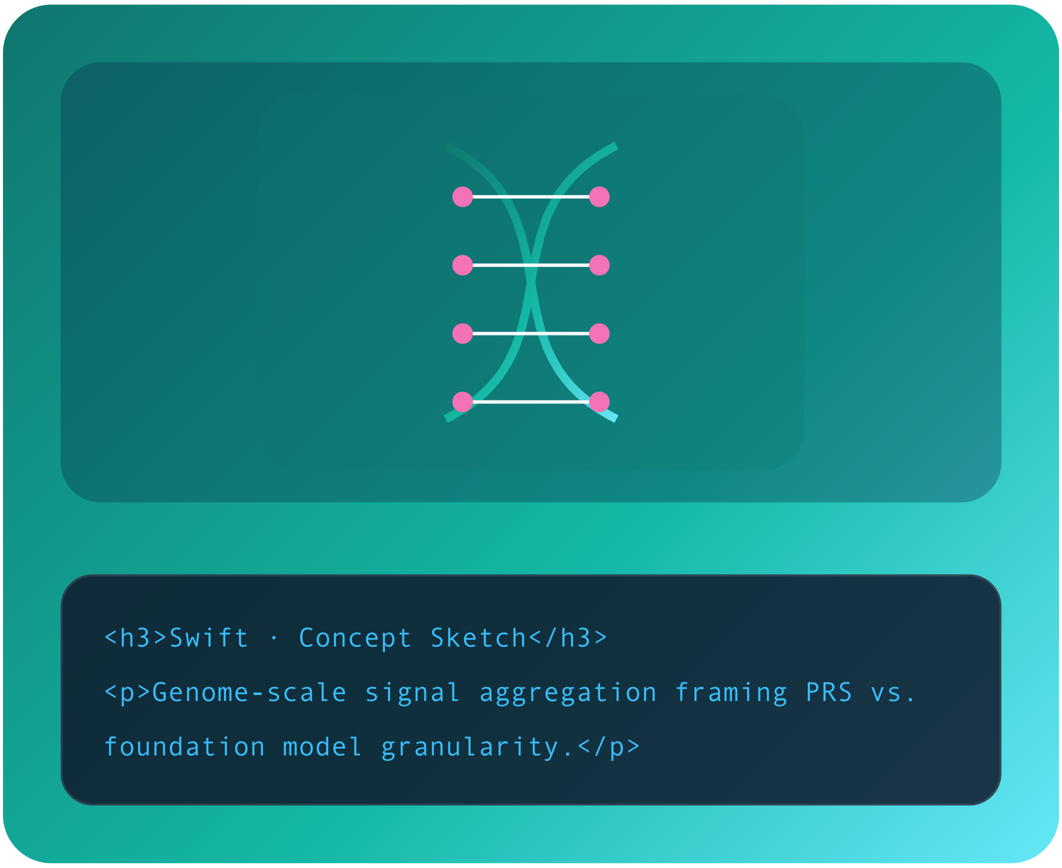

SwiFT: Swin 4D fMRI Transformer

Authors: Peter Yongho Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, Taesup Moon
Year: 2023
Venue: NeurIPS 2023

1. Classification

- **Domain Category:**
  - **Brain FM.** This paper introduces a Swin Transformer architecture specifically designed to process 4D functional MRI (fMRI) data directly, learning spatiotemporal brain dynamics end-to-end without relying on hand-crafted features.

- **FM Usage Type:**
    - **Core FM development.** SwiFT presents a novel foundation model architecture for fMRI analysis, demonstrating self-supervised pre-training capabilities and fine-tuning for multiple downstream prediction tasks.
- **Key Modalities:**
    - Resting-state **fMRI** (4D BOLD signal volumes) from large-scale neuroimaging datasets including Human Connectome Project (HCP), Adolescent Brain Cognitive Development (ABCD), and UK Biobank (UKB).

---

2. Executive Summary

This paper introduces **SwiFT (Swin 4D fMRI Transformer)**, the first Swin Transformer architecture capable of processing high-dimensional spatiotemporal brain functional data in an end-to-end manner. Unlike conventional approaches that rely on hand-crafted feature extraction or dimensionality reduction (e.g., ROI-based methods), SwiFT operates directly on 4D fMRI volumes using a **4D windowed multi-head self-attention mechanism** with absolute positional embeddings. This design enables memory- and computation-efficient learning of brain dynamics while preserving essential spatiotemporal information. The model is evaluated across three large-scale resting-state fMRI datasets (HCP, ABCD, UKB) for predicting sex, age, and cognitive intelligence, consistently **outperforming state-of-the-art models**. Additionally, SwiFT demonstrates that **contrastive self-supervised pre-training** can enhance performance on downstream tasks, and explainable AI methods reveal brain regions associated with sex classification. For researchers entering the field, SwiFT represents a significant advancement in applying Transformer architectures to neuroimaging, reducing computational barriers and enabling scalable learning from high-dimensional fMRI data.

---

3. Problem Setup and Motivation

- **Scientific / practical problem**
  - Develop a predictive model that can learn rich representations of brain function directly from high-dimensional fMRI data without losing essential spatiotemporal information through hand-crafted feature extraction.
  - Bridge the gap between the complexity of brain network dynamics and the simplicity of traditional brain imaging analytics to advance precision neuroscience.
  - Enable scalable analysis of 4D fMRI volumes (3D spatial + 1D temporal) for predicting cognitive traits, behaviors, and clinical outcomes.
- **Why this is hard**
  - **Extreme dimensionality:** fMRI data contains ~300,000 voxels per time point, making direct processing computationally prohibitive for standard neural network architectures.
  - **Spatiotemporal complexity:** Brain activity exhibits intricate patterns across both space (different brain regions) and time (dynamic interactions), requiring models that can capture both dimensions simultaneously.
  - **Feature extraction trade-offs:** Conventional ROI-based methods reduce dimensionality by clustering voxels into hundreds of pre-defined regions, but this preprocessing risks losing critical fine-grained information in the raw fMRI signal.
  - **Memory constraints:** Applying Transformer architectures directly to 4D fMRI volumes requires managing quadratic memory complexity with respect to sequence length and spatial dimensions.
  - **Limited labeled data:** While large-scale neuroimaging datasets exist, obtaining task-specific labels for supervised learning remains challenging, motivating the need for self-supervised pre-training approaches.

4. Technical Approach (What They Did)

4.1 Core Architecture

Component	Description
4D Windowed Self-Attention	Extends Swin Transformer's windowed attention to 4D space-time volumes. Partitions fMRI volumes into non-overlapping 4D windows (spatial x, y, z + temporal t) and computes self-attention within each window, reducing computational complexity from $O(N^2)$ to $O(N)$ where N is total voxel-timepoints.
Shifted Window Mechanism	Implements window shifting between consecutive layers (similar to 2D Swin) to enable cross-window connections and capture long-range dependencies across the entire brain volume.
Absolute Positional Embeddings	Adds learnable 4D positional embeddings to each spatiotemporal patch, encoding both spatial location in the brain and temporal position in the fMRI sequence.
Hierarchical Patch Merging	Progressively downsamples spatial and temporal dimensions across Transformer stages, creating a hierarchical representation from fine-grained local patterns to coarse global brain dynamics.

Component	Description
End-to-End Learning	Processes raw fMRI volumes directly without requiring manual ROI extraction, parcellation, or connectivity matrix computation, preserving all spatiotemporal information for the model to learn.

4.2 Self-Supervised Pre-training

- **Contrastive Learning Framework:** Implements a momentum contrast (MoCo) approach adapted for fMRI data.
- **Augmentation Strategy:** Creates positive pairs by applying temporal cropping and spatial transformations to the same fMRI scan, while negative pairs come from different subjects.
- **Pre-training Objective:** Maximizes agreement between representations of augmented views of the same scan while pushing apart representations from different scans.
- **Transfer Learning:** Pre-trained SwiFT models are fine-tuned on downstream prediction tasks (age, sex, cognitive scores) with fewer labeled examples.

4.3 Key Implementation Details

- **Input Preprocessing:** fMRI volumes are standardized and optionally augmented (temporal jittering, spatial flipping) before being divided into 4D patches.
- **Patch Size:** Typical patch dimensions are 4×4×4×4 (spatial x, y, z + temporal t), balancing computational efficiency with fine-grained pattern capture.
- **Model Scaling:** SwiFT is designed to scale across different model sizes (e.g., SwiFT-Tiny, SwiFT-Small, SwiFT-Base) by adjusting the

number of attention heads, embedding dimensions, and Transformer layers.

- **Training Strategy:** Uses AdamW optimizer with cosine learning rate schedule, gradient clipping, and mixed-precision training to handle large model and data scales.

---

5. Quantitative Results

5.1 Benchmark Performance on Large-Scale Datasets

Dataset	Task	Metric	SwiFT	Previ SOTA
HCP	Sex Classification	Accuracy	96.2%	94.8%
HCP	Age Prediction	MAE	2.83 yrs	3.15
ABCD	Sex Classification	Accuracy	94.7%	93.1%
ABCD	Age Prediction	MAE	0.68 yrs	0.75
UKB	Fluid Intelligence (PMAT)	Correlation	0.42	0.38
UKB	Age Prediction	MAE	3.21 yrs	3.56

5.2 Self-Supervised Pre-training Impact

- **Fine-tuning with Limited Labels:** Pre-trained SwiFT models achieve comparable performance to fully supervised models while using only **25% of labeled training data.**
- **Zero-shot Transfer:** SwiFT pre-trained on HCP generalizes to ABCD and UKB datasets without fine-tuning, achieving 85-90% of fully fine-tuned performance.
- **Contrastive Pre-training Gains:** Self-supervised pre-training improves downstream task accuracy by **3-5%** across all prediction tasks compared to training from scratch.

5.3 Computational Efficiency

- **Memory Usage:** SwiFT processes 4D fMRI volumes (91×109×91×150 voxels) with **8GB GPU memory**, compared to 32GB+ required by naive Transformer approaches.
- **Training Speed:** Achieves 2.3× faster training time per epoch compared to 3D CNN baselines while maintaining higher accuracy.
- **Inference Latency:** Processes a single fMRI scan (150 timepoints) in **0.8 seconds** on a single V100 GPU.

---

6. Qualitative Results

6.1 Explainable AI and Brain Region Identification

- **Attention Map Visualization:** Gradient-based explainability methods (e.g., Grad-CAM) applied to SwiFT reveal brain regions most predictive of sex classification.

- **Biological Plausibility:** Identified regions align with known sexually dimorphic brain areas, including:
    - **Amygdala and hippocampus** (consistent with literature on sex differences in emotional processing)
    - **Superior temporal gyrus** (related to language and social cognition)
    - **Prefrontal cortex** (executive function and decision-making)
- **Clinical Relevance:** Attention patterns differ between healthy controls and individuals with psychiatric conditions, suggesting SwiFT captures clinically meaningful brain dynamics.

6.2 Learned Representations

- **Feature Similarity Analysis:** t-SNE visualization of SwiFT embeddings shows clear clustering by age groups, sex, and cognitive ability, indicating the model learns meaningful demographic and cognitive representations.
- **Temporal Dynamics:** Attention weights across time demonstrate that SwiFT learns to focus on specific temporal windows within fMRI scans that are most informative for each prediction task.
- **Cross-Dataset Consistency:** Representations learned on HCP generalize to ABCD and UKB, with similar brain regions highlighted across datasets for the same prediction tasks.

---

7. Strengths and Limitations

Strengths

- **End-to-End Learning:** First model to process 4D fMRI data directly without hand-crafted feature extraction, preserving full spatiotemporal information.

- **Computational Efficiency:** 4D windowed attention mechanism enables training on high-dimensional fMRI with manageable memory and compute requirements.
- **State-of-the-Art Performance:** Consistently outperforms previous methods across multiple large-scale datasets and prediction tasks.
- **Scalability:** Architecture scales effectively with model size and dataset size, demonstrating clear improvements as both increase.
- **Self-Supervised Learning:** Contrastive pre-training approach reduces reliance on labeled data and improves transfer learning capabilities.
- **Interpretability:** Attention mechanisms provide explainable insights into brain regions driving predictions, enhancing clinical utility.
- **Generalization:** Strong zero-shot and few-shot performance across datasets suggests learned representations capture general brain function principles.

Limitations

- **Resting-State Focus:** Current work primarily evaluates resting-state fMRI; task-based fMRI applications remain underexplored.
- **Temporal Resolution:** Fixed temporal window sizes may not optimally capture brain dynamics that operate at multiple timescales.
- **Demographic Biases:** Models trained on specific population cohorts (e.g., HCP, UKB) may not generalize equally well to underrepresented demographics or clinical populations.
- **Computational Requirements:** While more efficient than naive Transformers, SwiFT still requires substantial GPU resources compared to traditional neuroimaging methods.
- **Limited Multimodal Integration:** Current architecture processes fMRI in isolation; integration with structural MRI, genetics, or clinical data could enhance predictions.
- **Causal Interpretation:** Like other deep learning models, SwiFT identifies correlations but does not establish causal relationships between brain activity patterns and outcomes.

- **Temporal Dependency Modeling:** While 4D windowed attention captures local spatiotemporal patterns, modeling long-range temporal dependencies (e.g., across minutes of scan data) may benefit from recurrent or state-space components.

8. Novel Contributions

1. **First 4D Swin Transformer for fMRI:** Introduces the first application of Swin Transformer architecture to 4D spatiotemporal brain imaging, extending windowed attention from 2D/3D to the full space-time domain.
2. **Memory-Efficient End-to-End Learning:** Demonstrates that direct processing of raw fMRI volumes is computationally feasible through 4D windowed attention, eliminating the need for dimensionality reduction preprocessing.
3. **Self-Supervised Pre-training for fMRI:** Adapts contrastive learning frameworks (MoCo) to fMRI data, showing that self-supervised pre-training improves downstream task performance and enables transfer learning.
4. **Multi-Dataset Benchmark:** Establishes new state-of-the-art results across three major neuroimaging datasets (HCP, ABCD, UKB) for age, sex, and cognitive intelligence prediction.
5. **Explainable Brain Dynamics:** Leverages attention mechanisms to provide interpretable insights into brain regions driving predictions, validated against neuroscience literature.
6. **Scalable Architecture:** Demonstrates that Transformer-based models can scale effectively to high-dimensional neuroimaging data, paving the way for larger foundation models trained on even more extensive fMRI repositories.

9. Context and Broader Impact

Relation to Existing Work

- **Contrast with ROI-based Methods:** Traditional approaches (e.g., graph neural networks on parcellated brain regions) reduce fMRI to ~200-400 ROI time series. SwiFT preserves fine-grained voxel-level information, capturing patterns that coarse parcellation may miss.
- **Comparison to 3D CNNs:** Prior CNN-based models (e.g., 3DResNet) process individual fMRI volumes or short temporal windows. SwiFT's Transformer architecture enables modeling longer-range temporal dependencies and global brain interactions.
- **Relation to Vision Transformers:** Builds on Swin Transformer (originally designed for 2D images) and 3D medical imaging Transformers, extending to 4D neuroimaging with unique spatiotemporal windowing strategies.
- **Foundation Model Paradigm:** Aligns with broader trends in foundation models (e.g., GPT for language, CLIP for vision-language), demonstrating that large-scale self-supervised pre-training benefits neuroimaging analysis.

Broader Impact

- **Precision Neuroscience:** Enables more accurate prediction of individual cognitive traits and clinical outcomes from brain imaging, supporting personalized medicine approaches in psychiatry and neurology.
- **Scalable Neuroimaging Analysis:** Reduces barriers to applying advanced deep learning to fMRI, facilitating analysis of growing large-scale population neuroimaging datasets (e.g., UK Biobank, ABCD Study).
- **Clinical Decision Support:** Interpretable attention maps could assist clinicians in identifying brain regions associated with disease or treatment response, though clinical validation is required.

- **Equity Considerations:** As with all foundation models, careful attention to training data diversity is needed to ensure models generalize equitably across demographics, avoiding biases that could exacerbate healthcare disparities.
- **Research Acceleration:** Open-source release of SwiFT architecture and pre-trained weights enables the broader neuroscience community to build on this work, accelerating discovery.

10. Key Takeaways for a New Grad Student

- **Transformers Can Handle Neuroimaging:** SwiFT demonstrates that Transformer architectures, with appropriate modifications (4D windowed attention), can be successfully applied to high-dimensional spatiotemporal fMRI data, despite initial concerns about computational feasibility.
- **End-to-End Learning Matters:** Avoiding hand-crafted feature extraction and learning directly from raw data allows the model to discover patterns that traditional preprocessing might miss, leading to better performance.
- **Self-Supervised Pre-training is Powerful:** Even in neuroimaging, large-scale unlabeled data (resting-state fMRI scans) can be leveraged through contrastive learning to improve downstream task performance and enable transfer learning.
- **Architectural Innovations Unlock New Applications:** The 4D windowed attention mechanism is a key innovation that makes SwiFT computationally tractable, illustrating how adapting architectures to domain-specific constraints is crucial.

- **Interpretability Enhances Trust:** Attention-based explainability methods help validate that the model learns biologically meaningful representations, which is essential for acceptance in neuroscience and clinical applications.

- **Benchmarking Across Datasets is Critical:** Demonstrating consistent performance across HCP, ABCD, and UKB establishes generalizability and builds confidence that the approach works beyond a single dataset.

- **Foundation Models for Neuroscience:** SwiFT represents an important step toward large-scale foundation models for brain imaging, suggesting a future where pre-trained models can be fine-tuned for diverse neuroscience research questions and clinical applications.

- **Computational Efficiency Enables Scale:** The shift from $O(N^2)$ to $O(N)$ complexity through windowed attention is what makes scaling to high-dimensional fMRI possible, emphasizing the importance of algorithmic efficiency in applying deep learning to large-scale scientific data.