Pixel-Level Understanding of a World in Motion within a Neural Encoding Framework

Mai Gamal[1] Mohamed Rashad[2] Eman Ehab[3] Seif Eldawlatly[4,5] Mennatullah Siam[6]

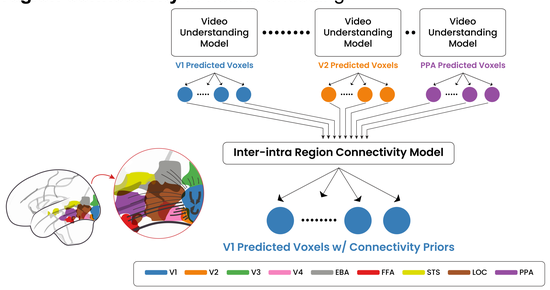
[1]German University in Cairo [2]Ain Shams University [3]Nile University [4]American University in Cairo [5]Ain Shams University [6]University of British Columbia

Scientific Reports (2026) 16:12188 · doi.org/10.1038/s41598-025-34141-w

INTRO & MOTIVATION

Neural Encoding of a World in Motion

- Deep networks' representations can be used to predict human visual cortical fMRI responses (neural encoding framework).
- Prior encoding studies use only **image- or video-level** (coarse) models' representations. We study pixel-level dense ones.
- We also study functional connectivity where we learn **inter–intra region connectivity** in neural encoding.



Inter–intra region connectivity encoding (predicting V1 as an example)

RESEARCH QUESTIONS

Q1 Scientific Reports
How do **pixel-level understanding** tasks and representations benefit the neural encoding of human brain responses?
- **pixel-level (dense)** vs. **image/video-level (coarse)** models
- **class-aware** vs. **class-agnostic** models

Q2 NeuroAI
Do video models that encode **dynamics** predict the visual cortex better than image models?

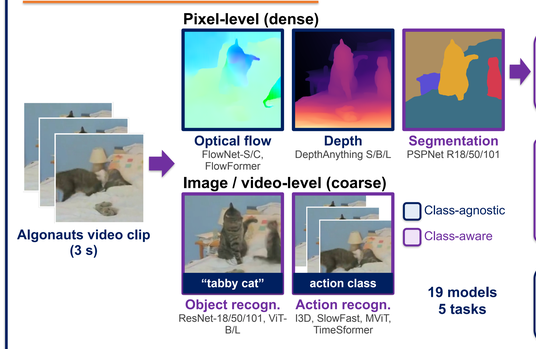
Q3 NeuroAI
Does modelling **inter– and intra–region connectivity** improve neural encoding?

We present the **first comprehensive study of pixel-level understanding models** (optical flow, depth, segmentation) for neural encoding of the visual cortex with video stimuli.

CONTRIBUTIONS

- **Dense vs. coarse:** pixel-level vs. image/video-level understanding.
- **Class-aware vs. class-agnostic** tasks.
- **19 models, 5 tasks:** CNNs & transformers, ventral & dorsal regions.
- **Layer-wise & connectivity analysis:** layer contribution per region and learned **inter– and intra–region connectivity** (NeuroAI).
- **New state-of-the-art** on BMD and Mini-Algonauts datasets.

METHOD OVERVIEW



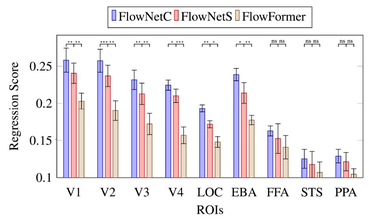
Pixel- vs. image/video-level

- Frozen features from 19 models, averaged over time.
- **Pixel-level:** optical flow, depth, segmentation.
- **Image/video-level:** object & action recognition.
- **Class-aware** tasks are supervised with categories; **class-agnostic** (flow, depth) use no class labels.

Layer-weighted regression

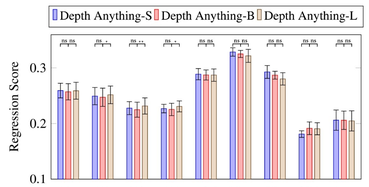
- Weighted sum over all layers (enforce sparsity on layer weights).
- Regression score **r** = Pearson correlation between predicted and measured fMRI responses (4-fold CV, paired t-test).
- **Layer contribution** = r using ***one layer*** / r using ***all layers***.
- **Connectivity** (NeuroAI): predictions refined with learned inter– and intra–region connectivity.

PIXEL-LEVEL UNDERSTANDING



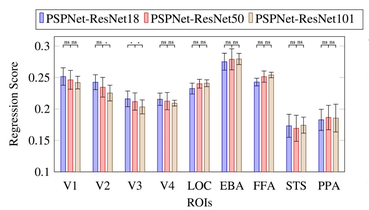
Optical flow

- CNN-based **FlowNet** outperforms the transformer-based **FlowFormer** in V1–V4.
- **no significant difference** in most late cortical regions.



Depth

- **Strong** in late cortical regions.
- Model size has little effect.

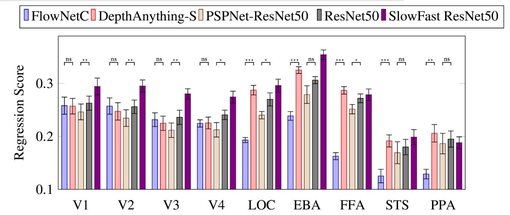


Segmentation

- **Better than optical flow** in late cortical regions; on par with or less than using depth models' representations.
- Model size has little effect.

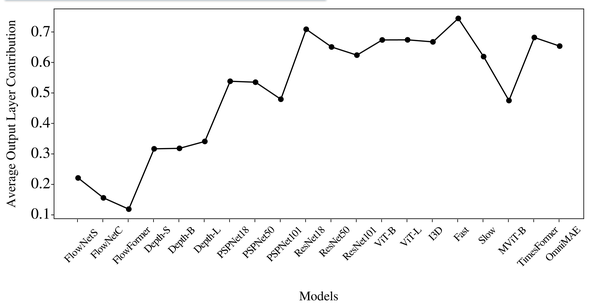
Intermediate layers, not the flow/depth output, drive neural encoding.

COMPARISON ACROSS TASKS



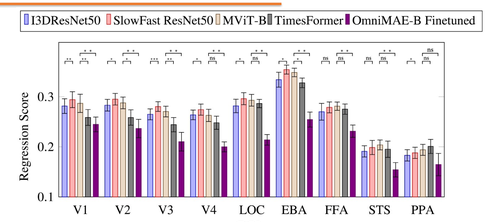
- **Depth** is strong in late cortical regions (LOC–PPA); **optical flow** is the weakest there.
- **Object recognition (coarse) representations outperforms segmentation (pixel-level dense) representations** with the same backbone.
- **SlowFast** is the best model in all regions except PPA.

LAYERS' CONTRIBUTION



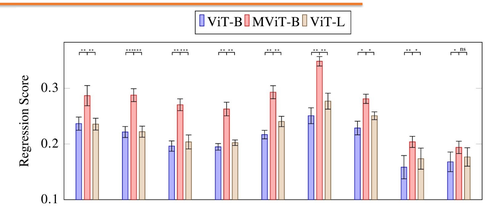
- **Class-agnostic** (flow, depth): low output-layer contribution.
- **Class-aware** (segmentation, object/action recognition): output layers contribute more.

IMAGE / VIDEO-LEVEL



- **CNNs outperform transformers** in early regions (V1–V4), except MViT which behaves closer to CNNs. Refer to artificial network target results for ablations removing any confounding factors.

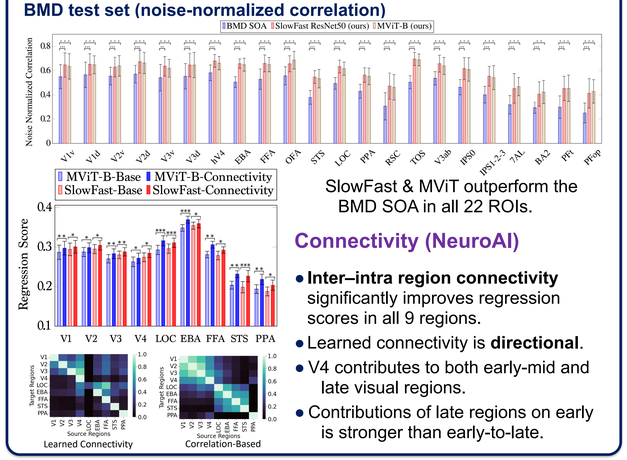
MULTISCALE TRANSFORMERS



- Multiscale processing improves transformers in encoding early-mid regions.
- V1-V4, EBA & FFA regions align better with video models' representations. STS & PPA have little/no diff.

STATE OF THE ART & CONNECTIVITY

BMD test set (noise-normalized correlation)



SlowFast & MViT outperform the BMD SOA in all 22 ROIs.

Connectivity (NeuroAI)

- **Inter–intra region connectivity** significantly improves regression scores in all 9 regions.
- Learned connectivity is **directional**.
- V4 contributes to both early-mid and late visual regions.
- Contributions of late regions on early is stronger than early-to-late.

TAKEAWAYS

- **Intermediate layers in depth models encode semantics**: depth is strong in late regions.
- **Optical flow models are the weakest** in predicting late regions that focus on high-level semantics.
- **Class-agnostic models rely on intermediate layers**, not their output; class-aware output layers contribute more.
- **CNNs surpass transformers** in early-to-mid regions, except multiscale **MViT**. This might relate to limitations of transformers in capturing high frequency components vs. convolutional models.
- **Object recognition (image-level coarse) surpasses segmentation (pixel-level dense)** in most regions.
- **Inter–intra region connectivity** improves encoding in all regions; the learned connectivity is directional (NeuroAI). Our learned connectivity differs from simple correlation-based ones.
- **SlowFast & MViT** set a new state of the art on BMD.