Neural Computation Laboratory

UNIVERSITY OF LEEDS

UNIVERSITY OF BIRMINGHAM

Lancaster University

EUROPEAN CONFERENCE ON COMPUTER VISION

Lab:

Code:

https://github.com/wurining/Vi-ST

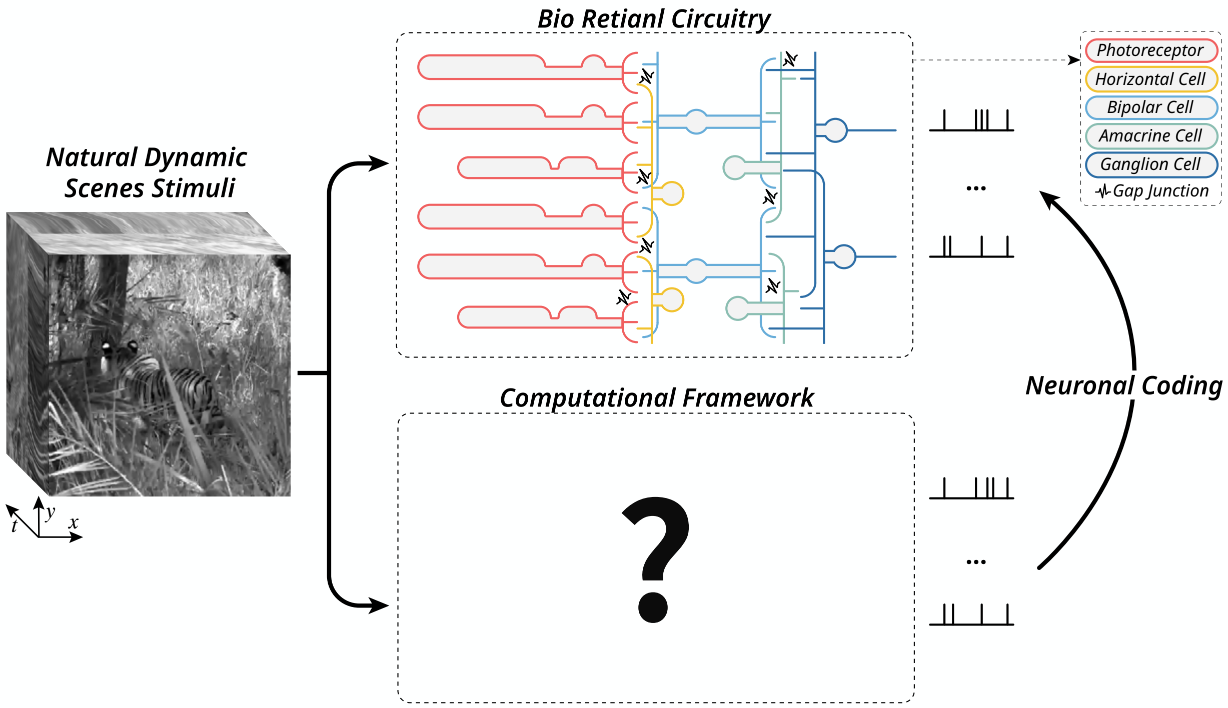
Aligning Neuronal Coding of Dynamic Visual Scenes with Foundation Vision Models

Rining Wu[1,2], Feixiang Zhou[3], Ziwei Yin[2], Jian K. Liu[1,2]

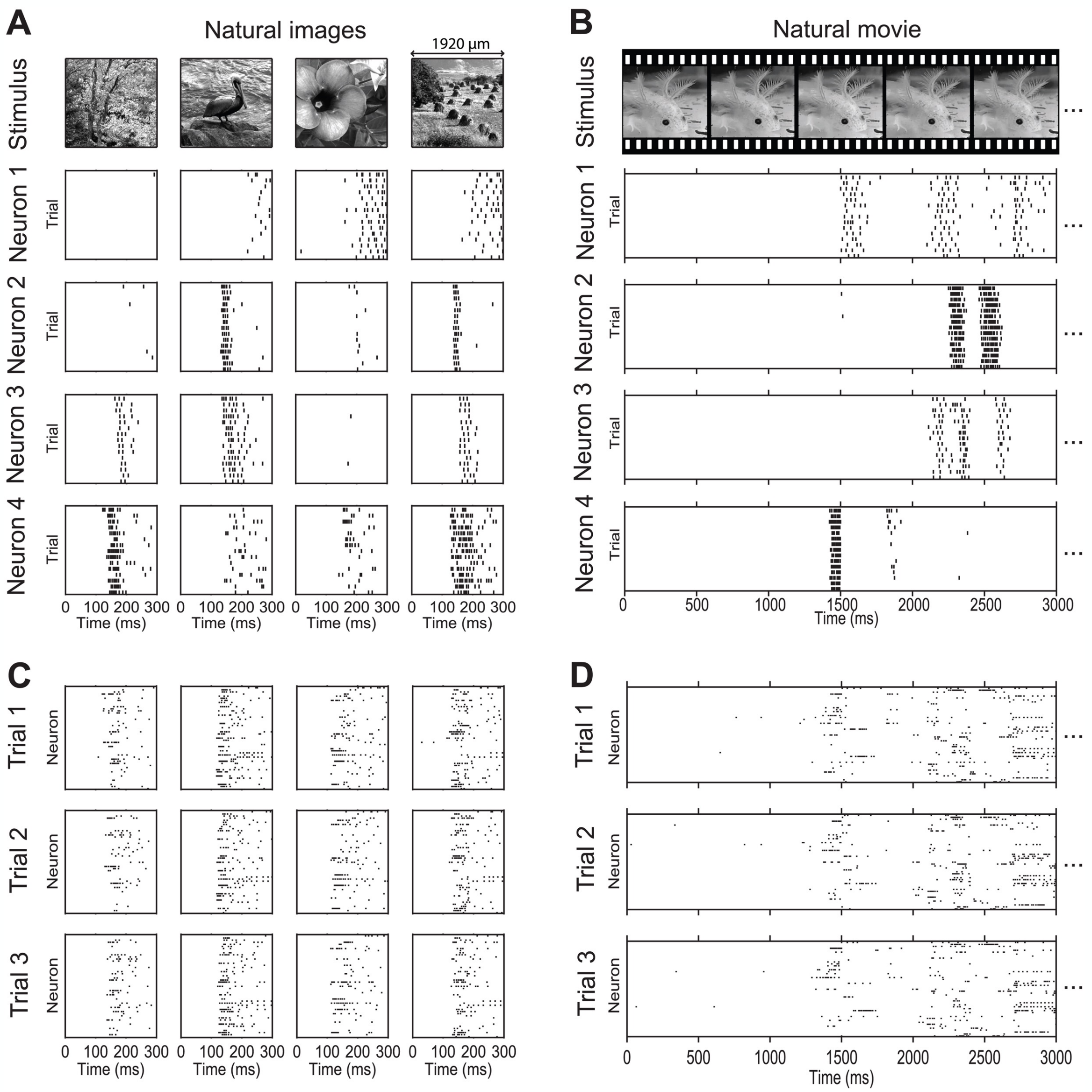
[1] University of Leeds, [2] University of Birmingham, [3] Lancaster University

Motivations

- Unraveling visual encoding of dynamic visual scenes is an important topic
- Foundation vision models have paved an advanced way of understanding image pixels
- Exploring a new perspective on the quantitative analysis of retina's capabilities



Salamander Retina Ganglion Cells (RGC) Neural Spikes



Stimuli: Nature Scene Video (**30Hz, 360x360px**)

Mov1: **1800** Frames

Mov2: **1600** Frames

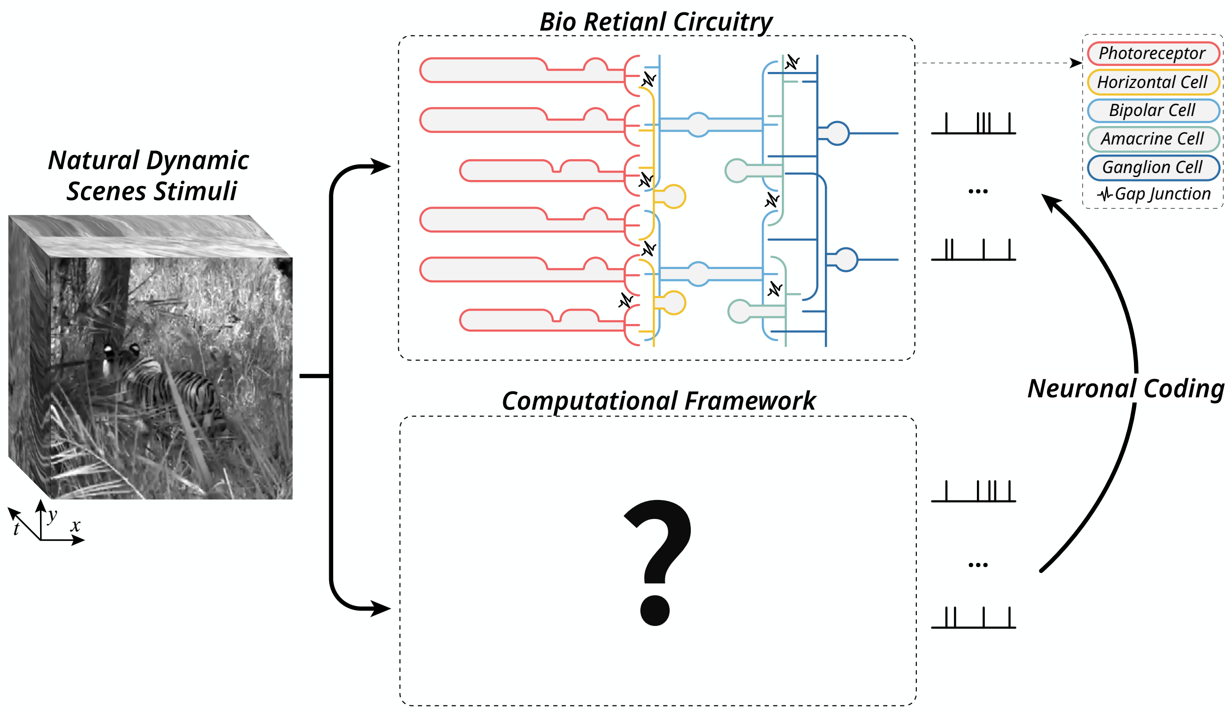
RGCs Response (Firing Rate)

- Utilize the Multielectrode recordings for **90** RGCs

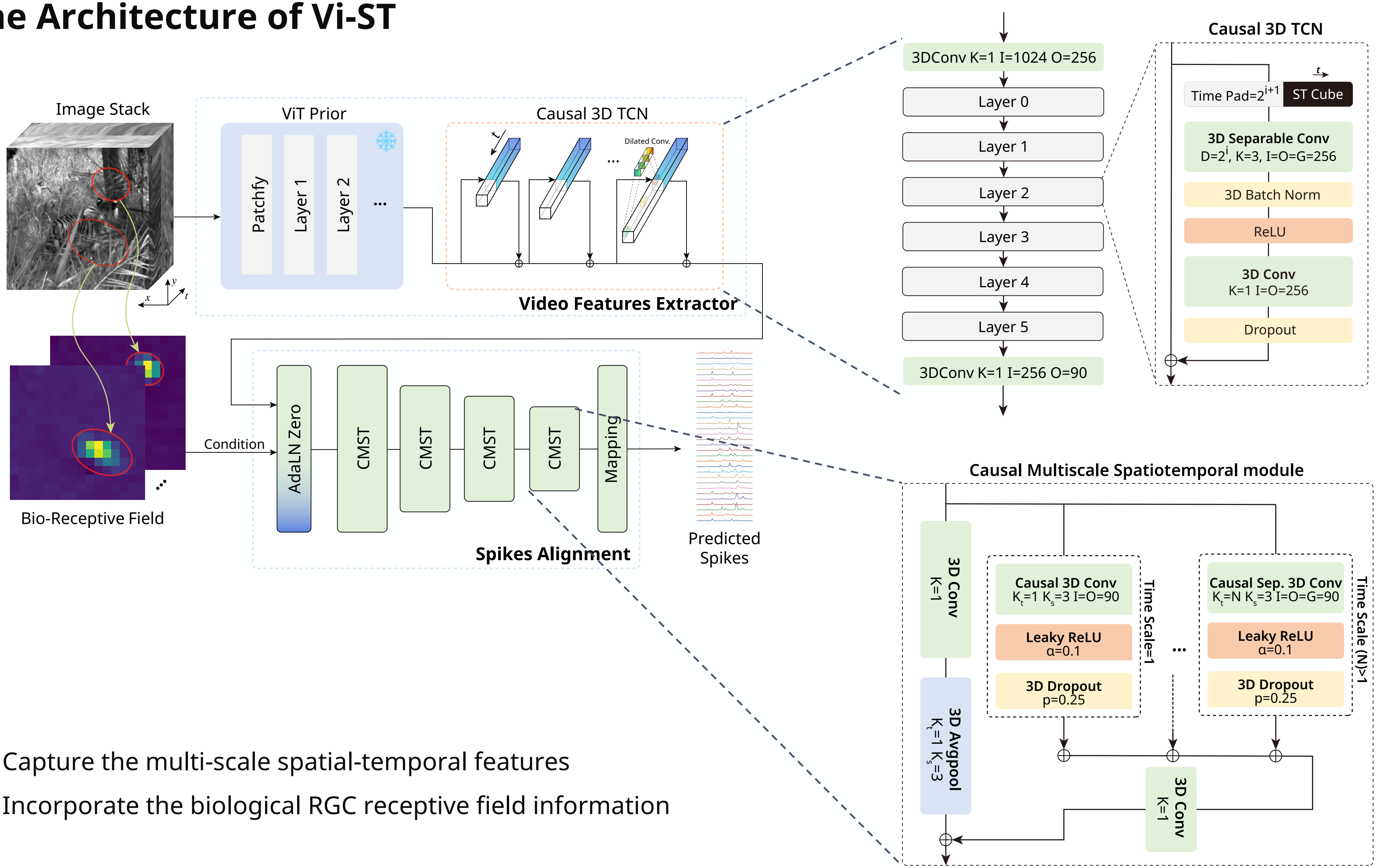
(Arno Onken, Jian K. Liu, and et al., *Using Matrix and Tensor Factorizations for the Single-Trial Analysis of Population Spike Trains*, 2016)

Highlights

- Introducing ***Vi-ST***, a ***s**patio**t**emporal convolutional network with a pre-trained ***ViT*** as a prior
- Detailed ablation experiments for demonstrating the significance of modules
- Introducing a visual coding evaluation metric, named ***SD-KL***
- Comparing the impact of different numbers of neuronal populations on complementary coding.

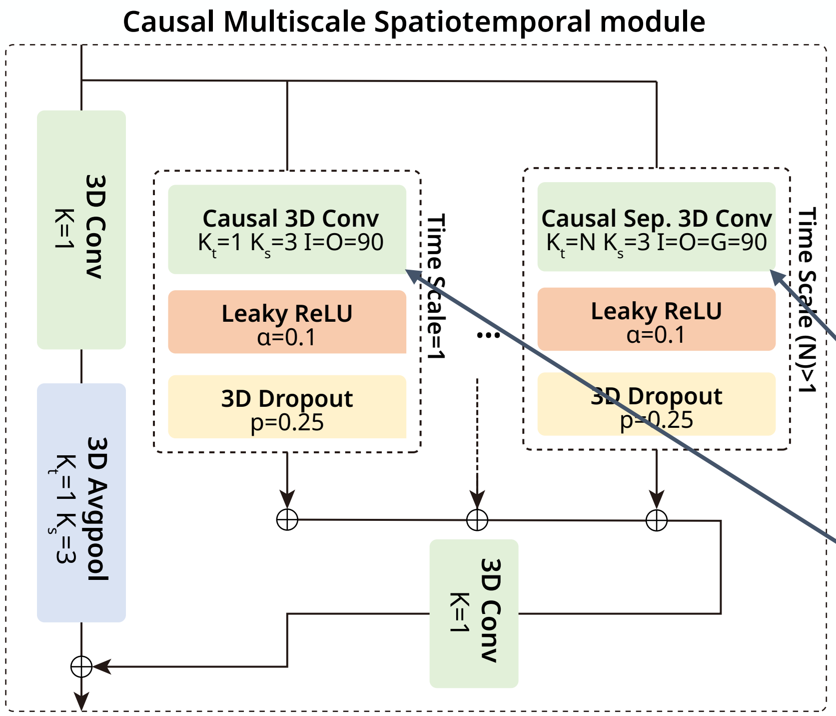


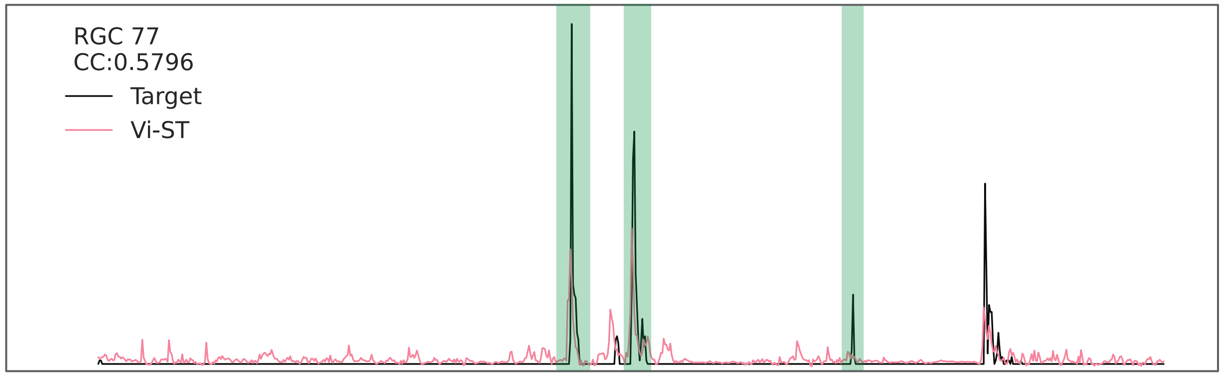
The Architecture of Vi-ST



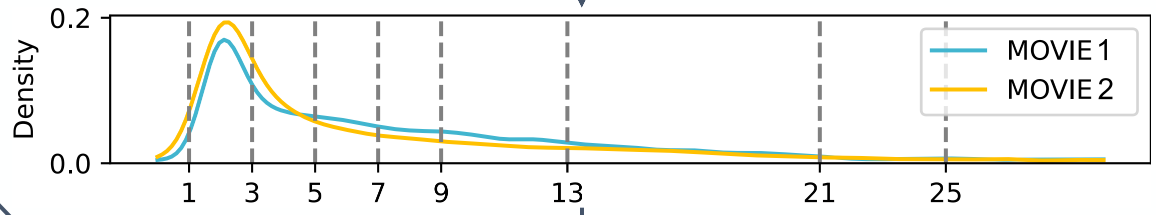
- Capture the multi-scale spatial-temporal features
- Incorporate the biological RGC receptive field information

The Architecture of Vi-ST





***Spike Duration**: from non-spike to non-spike*



CMST Level	Temporal	Spatio
1	1, 25	3
2	1, 13, 21	3
3	1, 7, 9	3
4	1, 3, 5	3

Manually identify spike duration to be the kernel size of CMST

Loss Function

$$\boxed{\mathcal{L}_{\text{Vi-ST}} = \alpha\mathcal{L}_{\text{RMSE}} + \beta\mathcal{L}_{\text{-ReLU}} + \gamma{\mathcal{L}_{\text{SoftDTW}}}^{6} + \gamma{\mathcal{L}_{\text{SoftDTW}}}^{12}}$$

α, β, and γ are hyperparameters, and we set them to 0.1, 0.5, and 5×10-6, respectively.

$$\mathcal{L}_{\text{RMSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Root Mean Square Error (RMSE): Euclidean loss

$$\mathcal{L}_{\text{-ReLU}} = \frac{1}{n}\sum_{i=1}^{n}\max(0, -\hat{y}_i)$$

Negative ReLU function: penalty term

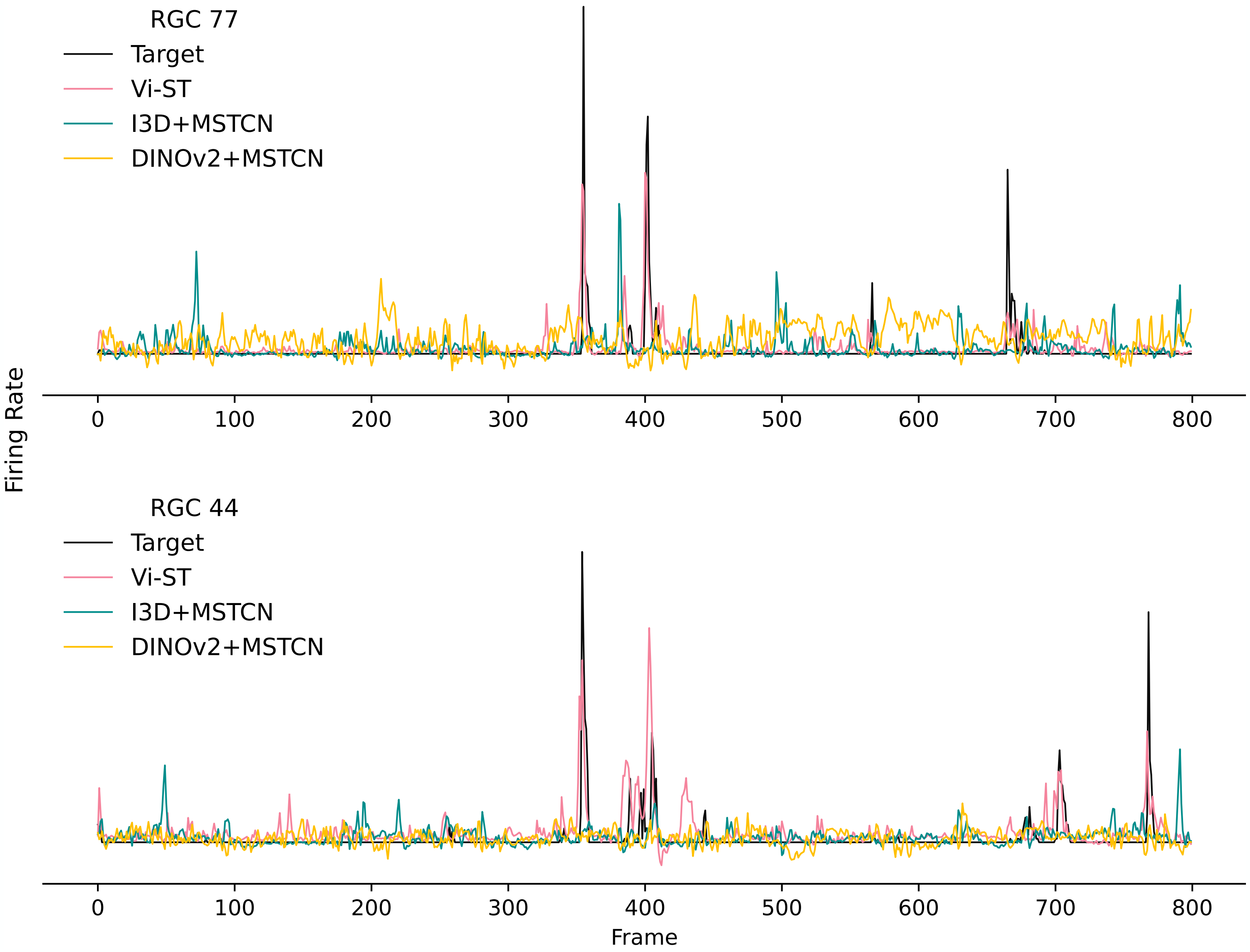
$${\mathcal{L}_{\text{SoftDTW}}}^{n} = \frac{1}{L-n}\sum_{i=1}^{L-n}\text{SoftDTW}(y_i, \hat{y}_i), i \in \{1, 2, \ldots, L-n\}$$

Soft Dynamic Time Warping (SoftDTW)

- Unlike Euclidean losses such as RMSE, considers potential time shifts or variations of length of durations
- Using rolling windows to avoid predicting longer time windows may lead to distortion and difficulty in representing local abrupt changes

(Cuturi, M., Blondel, M.: Soft-DTW: a Differentiable Loss Function for Time-Series)

Better Generalization



Model*	Mov1 → Mov1	Mov2 → Mov2
CRNN	*0.857*	*0.718*
I3D+MSTCN	0.846	0.668
DINOv2+MSTCN	0.849	0.672
Vi-ST	0.789	0.570
Model**	**Mov1 → Mov2**	**Mov2 → Mov1**
I3D+MSTCN	0.108	0.074
DINOv2+MSTCN	0.101	0.100
Vi-ST	*0.334*	*0.281*

** training and testing data are taken from the same video where pixel context is conserved*

*** training and testing data are taken from the different video*

Vi-ST gives better the generalization ability

Metrics

Pearson correlation coefficient (CC) :

While CC considers the macro trends of the entire sequence, it lacks an attention for temporal information

Spike Duration - Kullback-Leibler Divergence (SD-KL) :

Consider the detailed consideration of temporal information or dynamics over time

Algorithm 1 Pseudocode of the SD-KL

Input: $\hat{y} \in \mathbb{R}^{N\times F}, y \in \mathbb{R}^{N\times F}, \alpha = 0.3, \beta = 1.0$

Output: score

$\hat{\mathcal{D}} \leftarrow \textbf{peak widths}(\min(\max(0, \hat{y}), \beta)), \hat{\mathcal{D}} \subset \mathbb{R}$

$\mathcal{D} \leftarrow \textbf{peak widths}(\min(\max(0, y), \beta)), \mathcal{D} \subset \mathbb{R}$

$Var_{\cup} \leftarrow \frac{\alpha}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2, x_i \in \{D, \hat{D}\}$

$\mathcal{L}_{\cup} \leftarrow \left\{(lower_{\cup} - 3 * Var_{\cup}) + i \cdot \frac{(upper_{\cup}+3*Var_{\cup})-(lower_{\cup}-3*Var_{\cup})}{199} \mid i = 0, 1, \ldots, 199\right\}$

$pdf_{\mathcal{D}} \leftarrow \textbf{KDE}(\mathcal{D}, \alpha)$

$pdf_{\hat{\mathcal{D}}} \leftarrow \textbf{KDE}(\hat{\mathcal{D}}, \alpha)$

$P_{\mathcal{D}} = \left\{\left(x_i, \frac{pdf_{\mathcal{D}}(x_i)}{\sum_{j=1}^{N} pdf_{\mathcal{D}}(x_j)}\right) \mid x_i \in \mathcal{L}_{\cup}\right\}$

$P_{\hat{\mathcal{D}}} = \left\{\left(x_i, \frac{pdf_{\hat{\mathcal{D}}}(x_i)}{\sum_{j=1}^{N} pdf_{\hat{\mathcal{D}}}(x_j)}\right) \mid x_i \in \mathcal{L}_{\cup}\right\}$

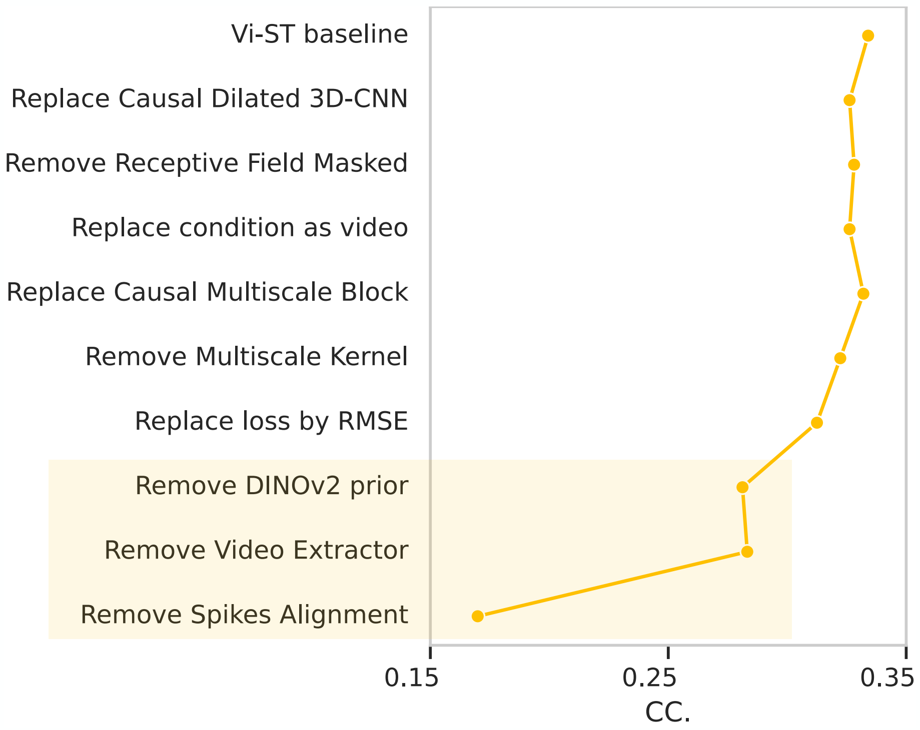
score $\leftarrow D_{KL}(P_{\hat{\mathcal{D}}} || P_{\mathcal{D}})$

score $\leftarrow \min(\max(0, \text{score}), 1000)$

return score

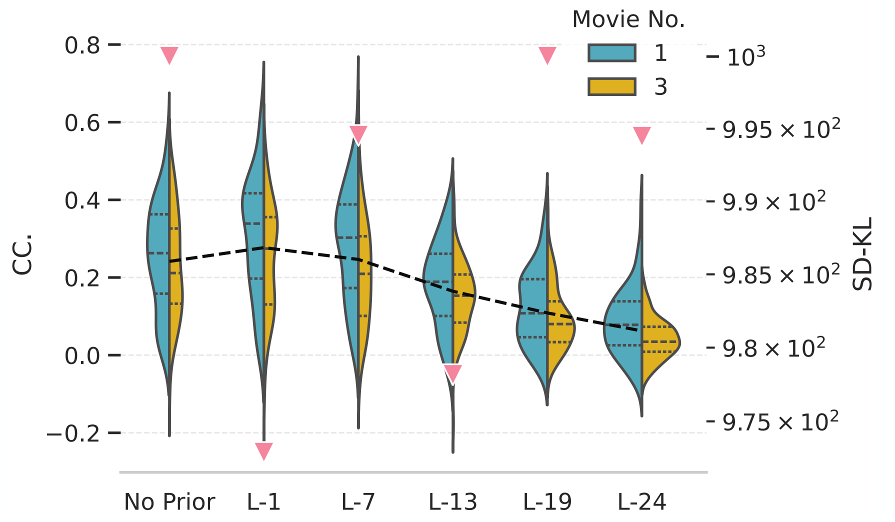
- Selects the lengths of corresponding subsequences in the response sequence, representing the duration of a complete neural response (from non-spike to non-spike).
- Then, compare the similarity of distribution which are calculated by Kernel Density Estimation, by KL divergence.

Discussion

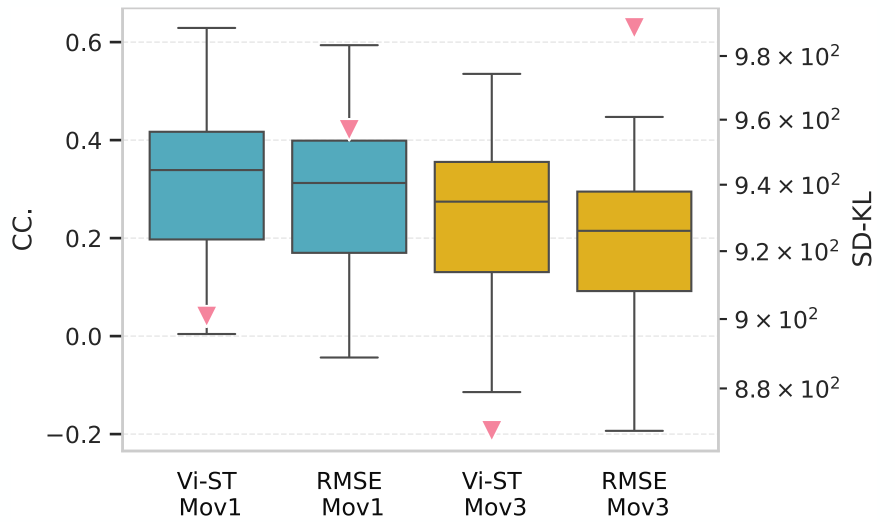


(a) Ablation Study (Mov1 → Mov2)

SD-KL: Smaller is better



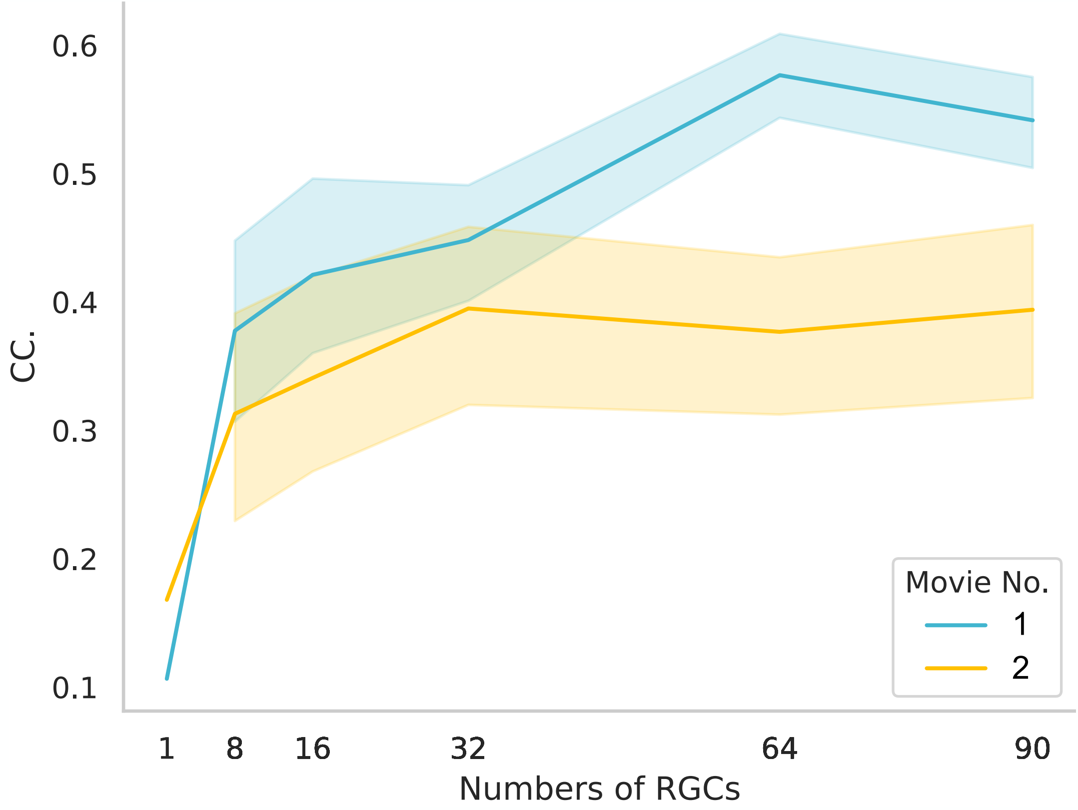
(b) The layer number of DINOv2 are represented as ***L - n***, e.g. ***L - 1*** represents the first layer



(c) Comparison of Euclidean and non-Euclidean loss

Discussion: Comparison of benefits of complementary coding

Is it optimal to construct an end-to-end model capable of simultaneously predicting all neural responses?



1. The CC of 90 RGCs predicted by the model are sorted, focusing on **the top 8 RGCs;**

2. The experiment uses encodings of 90, 64, 32, 16, 8, and 1 to make predictions;

3. The top 8 RGCs'CC from step 1 are then compared;

4. The results represent the average CC of the top 8 RGCs

(Ding, X., Lee, D., Melander, J.B., Sivulka, G., Ganguli, S., Baccus, S.A.: Information Geometry of the Retinal Representation Manifold)