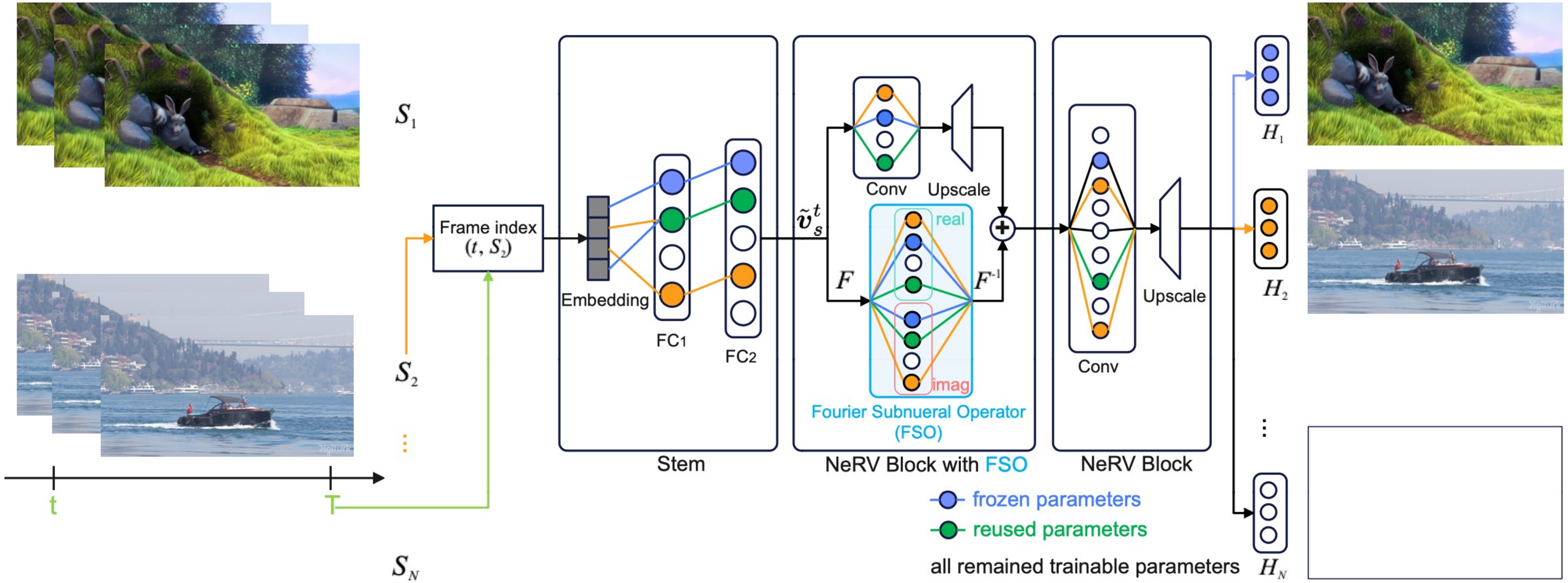
Progressive Fourier Neural Representation for Sequential Video Compilation

Haeyong Kang, Jaehong Yoon, DaHyun Kim, Sungju Hwang, and Chang D. Yoo

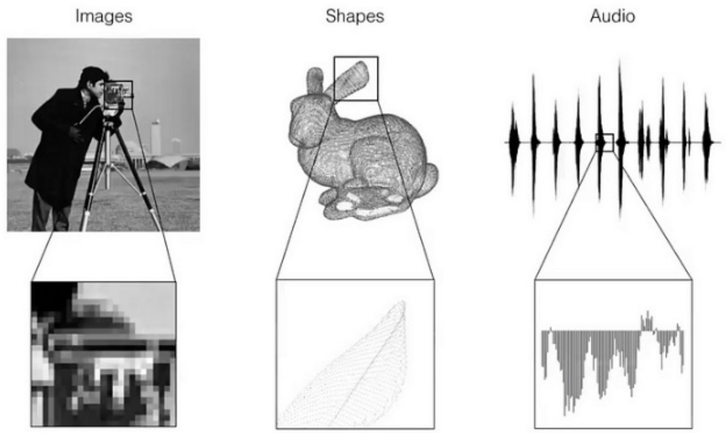
Motivation:
(1) It has not been explored sequential video so far in Continual Learning.
(2) There is also no robust continual baseline in Video Continual Learning.



➔ We propose a new method to show the effectiveness of reused winning tickets (WSN, ICML2022) in Video Continual Representations.

Backgrounds of Neural Implicit Representation (NIR)

- Backgrounds of Neural Implicit Rep.



- The world around us is not discrete,
- Yet, we choose to represent real-world signals such as images or sound in a discrete manner.

(-) Discrete only contain a discrete amount of information regarding the signal.

(-) Given a 256x256 pixel grid for an image, we are not able to scale it up to a 512x512 image.

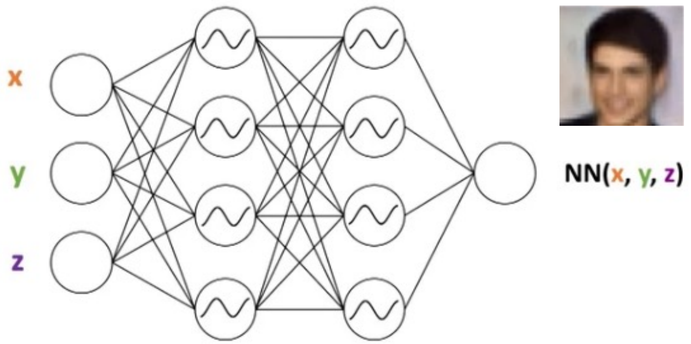
← Not enough information

- **Neural Implicit Representations (NIR)** are neural networks (e.g. MLPs) that estimate the function ***f*** that represents a signal continuously, by training on discretely represented samples of the same signal.

- **NIP** learns how to estimate the underlying (continuous) function ***f*** (denoted ***F*** below):

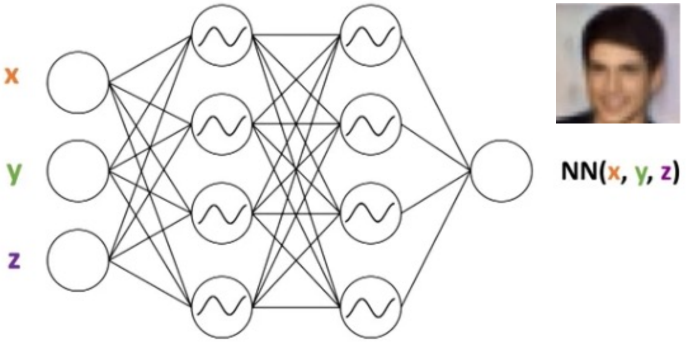
$$F(x, \Phi, \nabla_x \Phi, \nabla_x^2 \Phi, \cdots) = 0, \ \ \Phi : x \mapsto \Phi(x)$$

- The network parameterizes Φ. After training on the discretely represented samples, the estimated ***f*** would be *implicitly* encoded in the network, hence the name "**Neural Implicit Representation**".



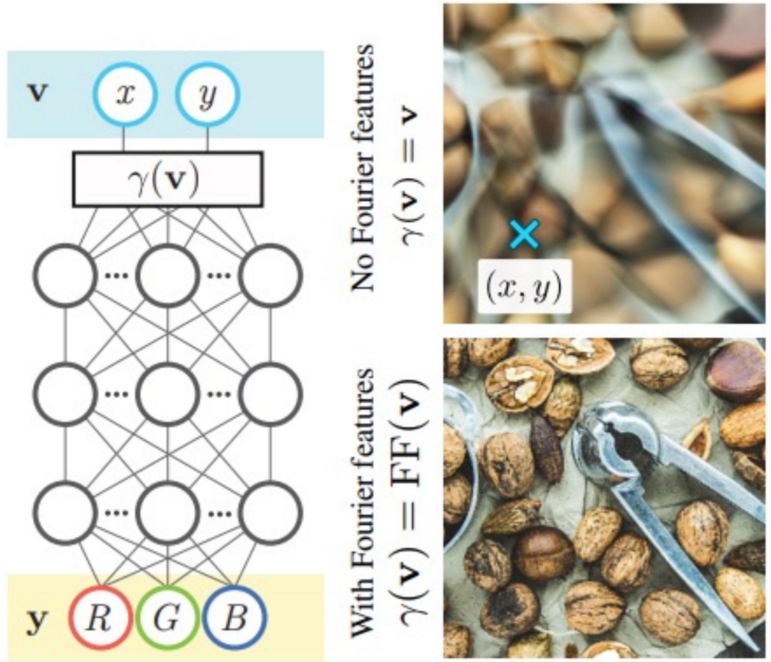
Discrete representations of various signals (SIREN: Sinusoidal Representation Networks, NeuralPS2020)

Coordinate-based Methods



SIREN: Sinusoidal Representation Networks, NeuralPS2020

(+) map coordinates to images.

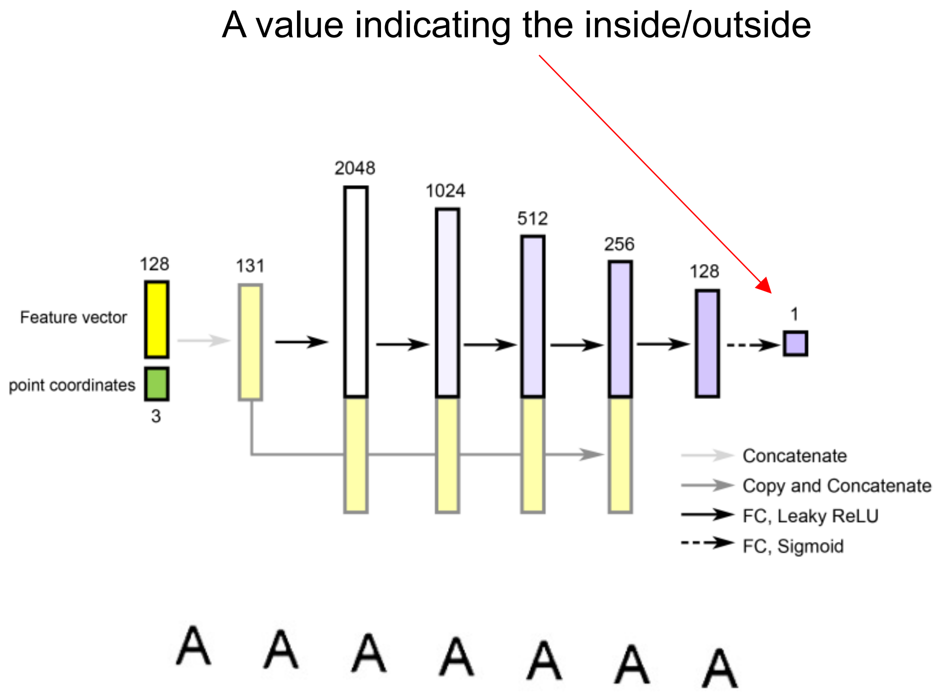


(a) Coordinate-based MLP

(b) Image regression $(x,y) \rightarrow$ RGB

Fourier features, NeuralPS2020

(+) to learn representation with high frequencies

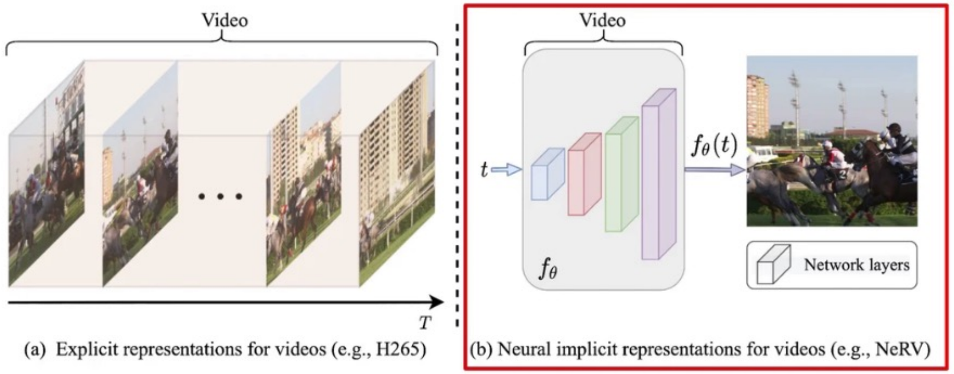


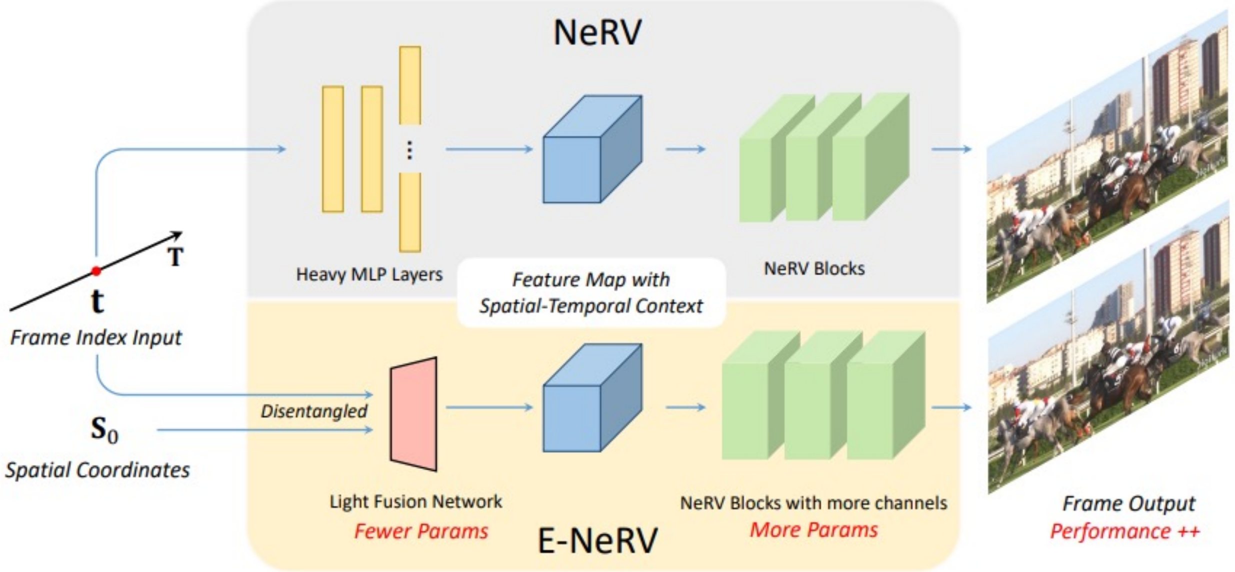
IMNET, CVPR2019

(+) Feature + coordinate led to cleaner interpolation results.

Index-based Methods (1)

(+) Faster and more accurate than Coordinate-based Methods





NeRV, NeuralPS2021

E-NeRV, ECCV2022

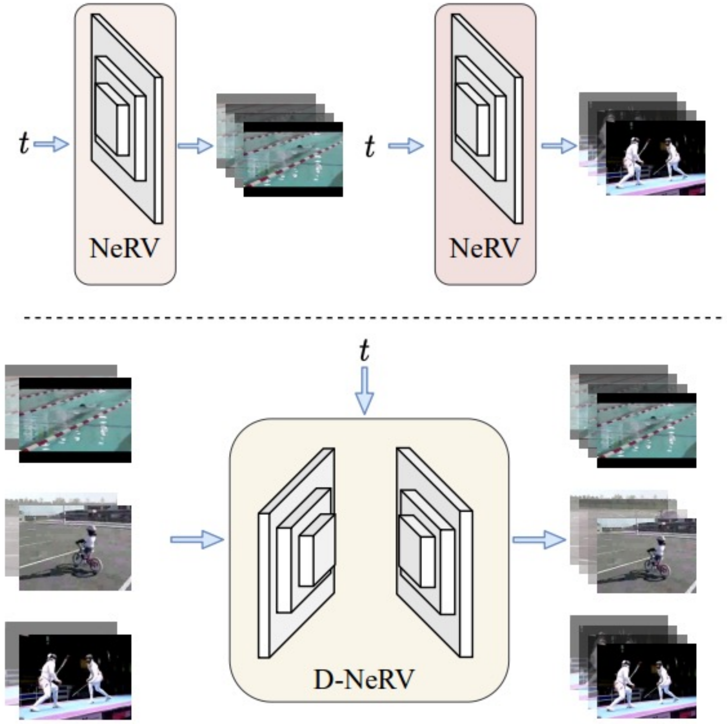
(+) NeRV represents videos or image datasets as neural neworks, **taking an image index as input**, and **outputting the whole image**.

(+) **Faster than coordinate-based methods**.

(+) The size of the parameters were reduced **by introducing disentangled spatial-temporal representations** with a light network, while maintaining the majority of performance.

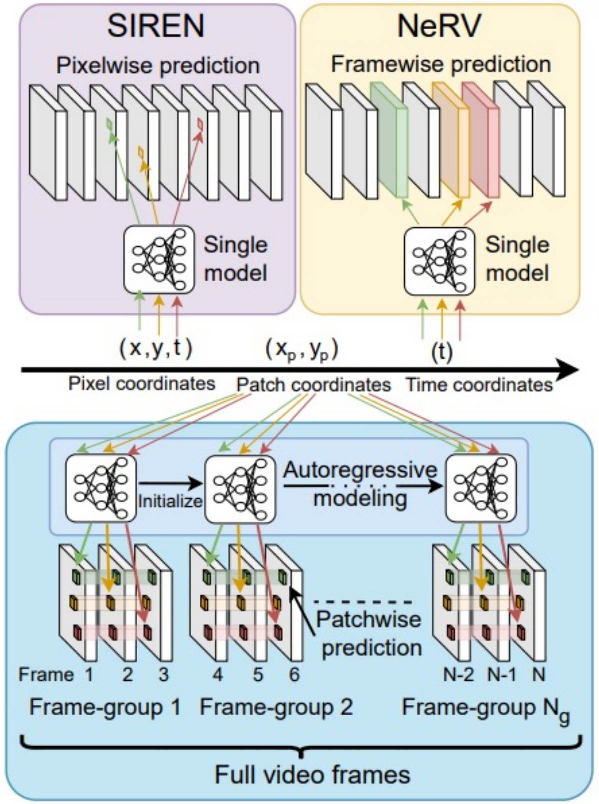
(+) They distributed the saved parameters to increase channel dimensions in convolution blocks, resulting in an E-NeRV model with similar or fewer parameters but much better performance.

Index-based Methods (2)



D-NeRV, CVPR2023

(+) NeRV optimizes representation to every video independently while **D-NeRV encodes all videos by a shared model**.



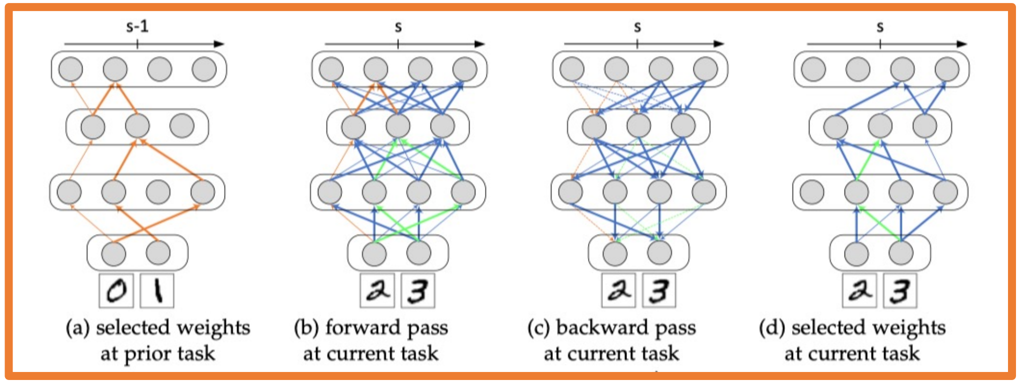
NIRVANA, CVPR2023

(+) **NIRVANA performs spatio-temporal patch-wise prediction** and **fits individual neural networks to groups of frames (clips)** which are initialized using networks trained on the previous group.

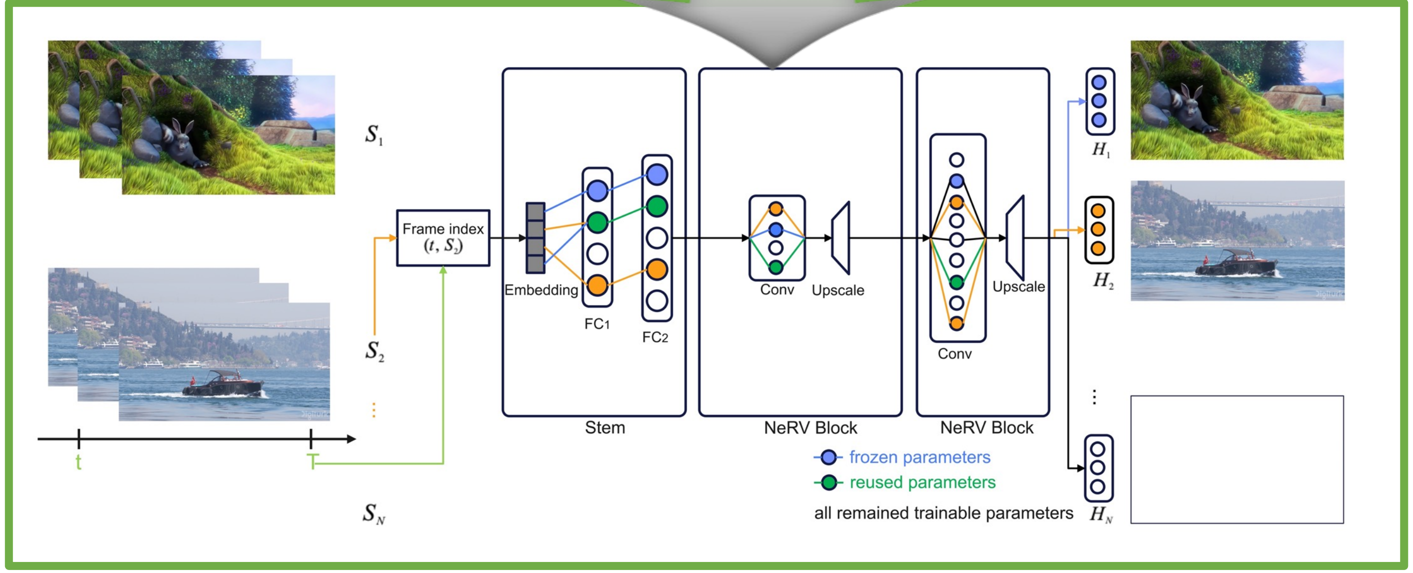
(+) Such an autoregressive patch-wise approach exploits both spatial and temporal redundancies present in videos while promoting scalability and adaptability to varying video content, resolution or duration.

Index-based sequential Neural Implicit Representation (sequential NIR)

WSN, ICML2022



(+) Providing forget-free continual learning solutions.
(+) Reused Subnetworks Adaptably tailer to tasks.
(+) Fast Convergence .



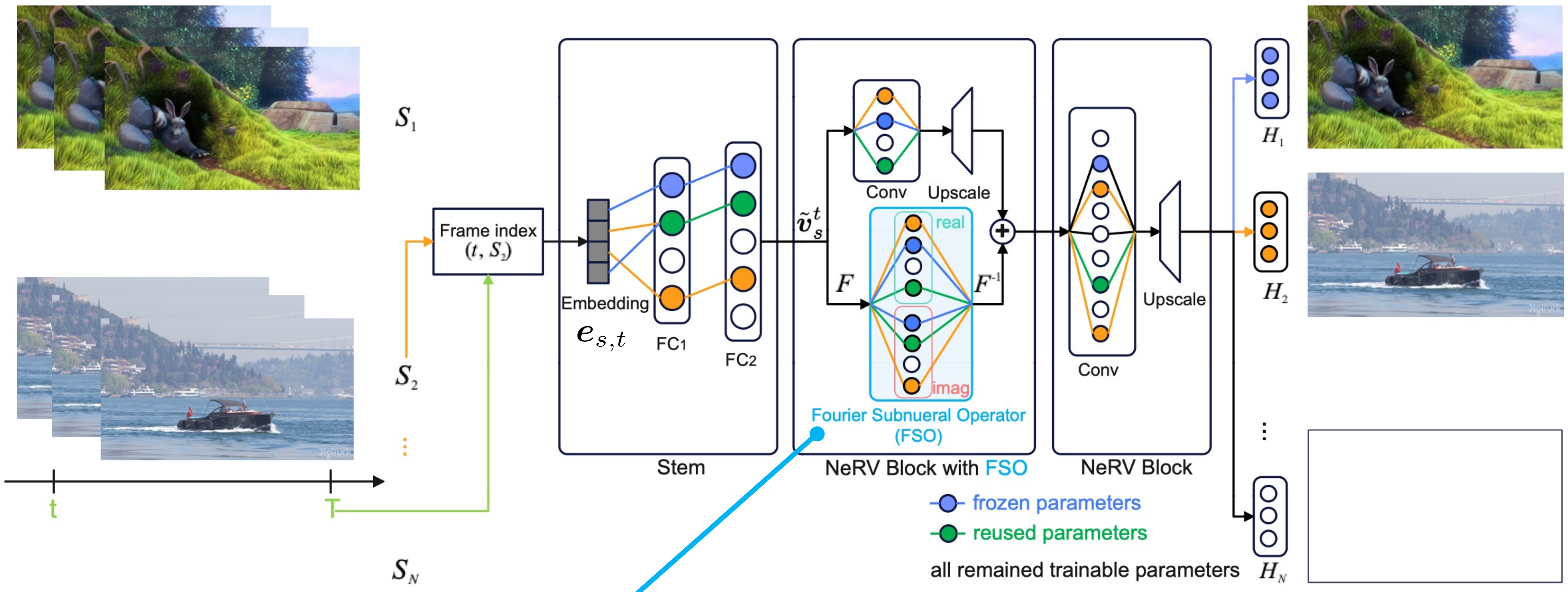
(+) Providing forget-free video continual learning.
(+) Fast Convergence.

(-) WSN does not have enough parameters for complex video representations.

➔ Needs more parameters to deal with the issue.

Index-based sequential Neural Implicit Representation (sequential NIR)

(+) Fourier Subneural Operator (FSO): Enough Reusable parameters in Frequency domain



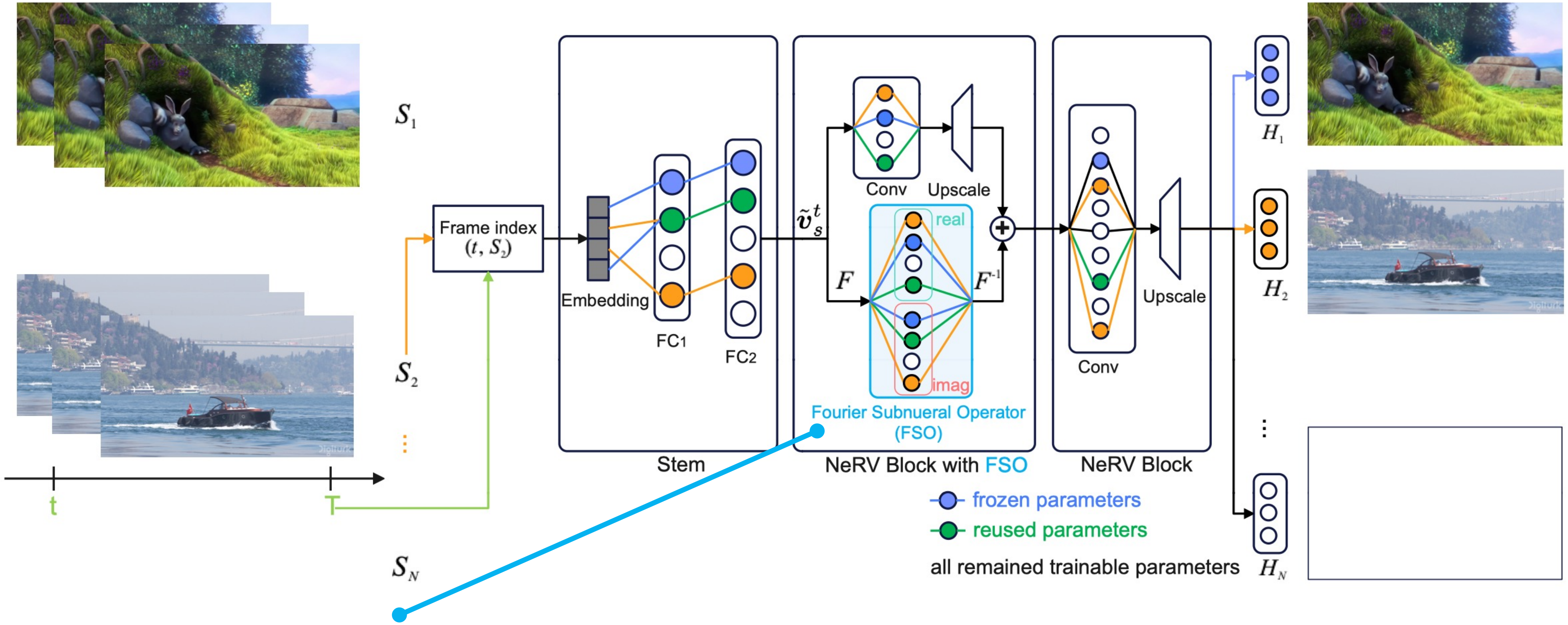
$$(\mathcal{K}(\phi)\tilde{\boldsymbol{v}}_t^s)\,(\boldsymbol{e}_{s,t}) = \mathcal{F}^{-1}(R_\phi \cdot (\mathcal{F}\tilde{\boldsymbol{v}}_t^s))(\boldsymbol{e}_{s,t}),$$

the Fourier transform of a periodic subnetwork function

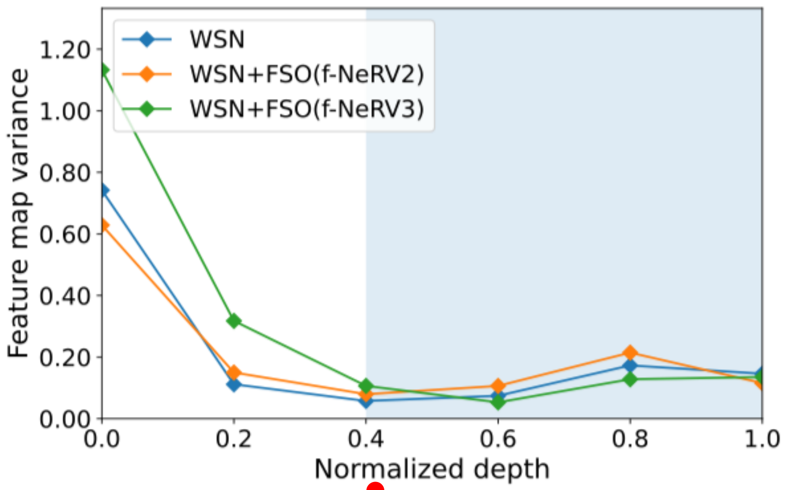
R_ϕ of $(\boldsymbol{\theta}^{real} \odot \boldsymbol{m}_s^{real})$ and $(\boldsymbol{\theta}^{imag} \odot \boldsymbol{m}_s^{imag})$

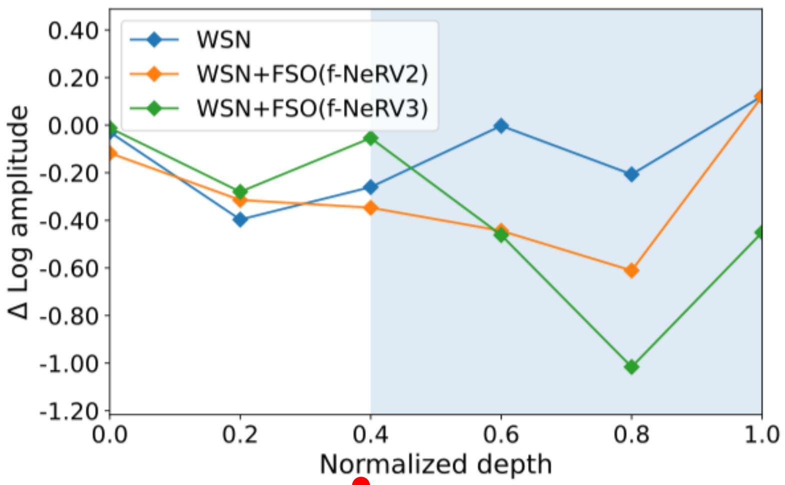
Index-based sequential Neural Implicit Representation (sequential NIR)

(+) Fourier Subneural Operator (FSO): Enough Reusable parameters in Frequency domain

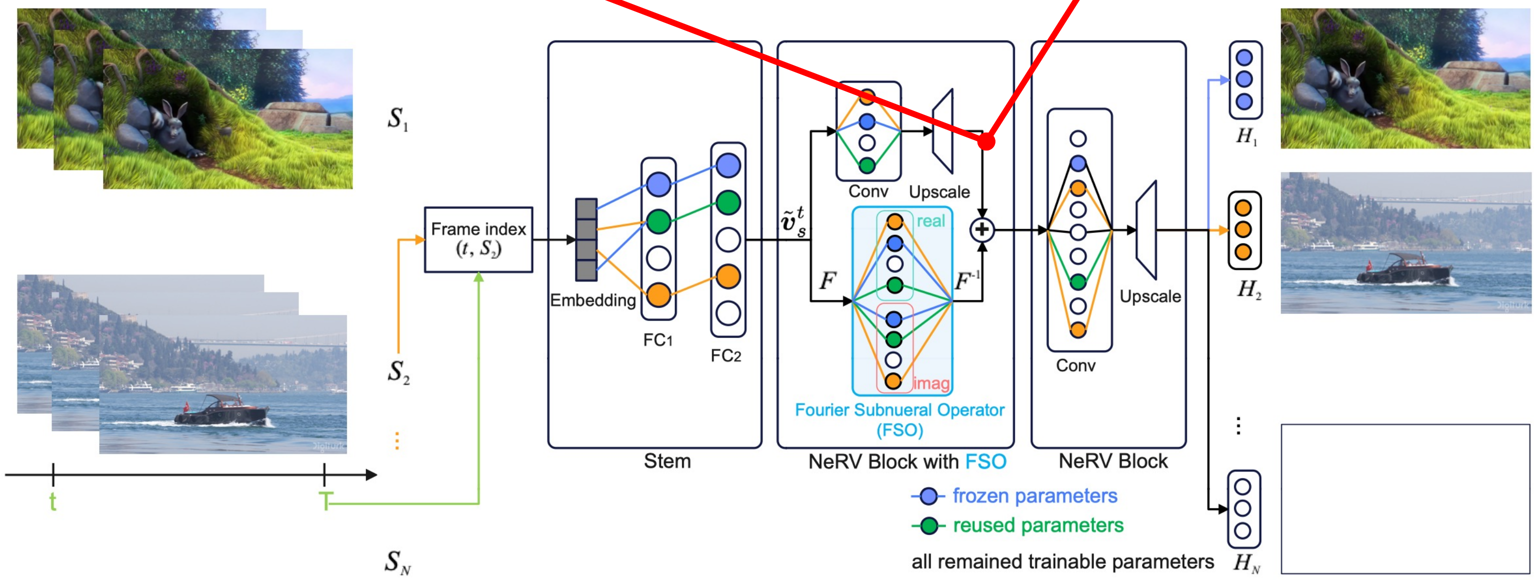


(+) Providing forget-free video continual learning
(+) Fast Convergence
(+) Enough Reusable parameters in Frequency domain for complex video representation



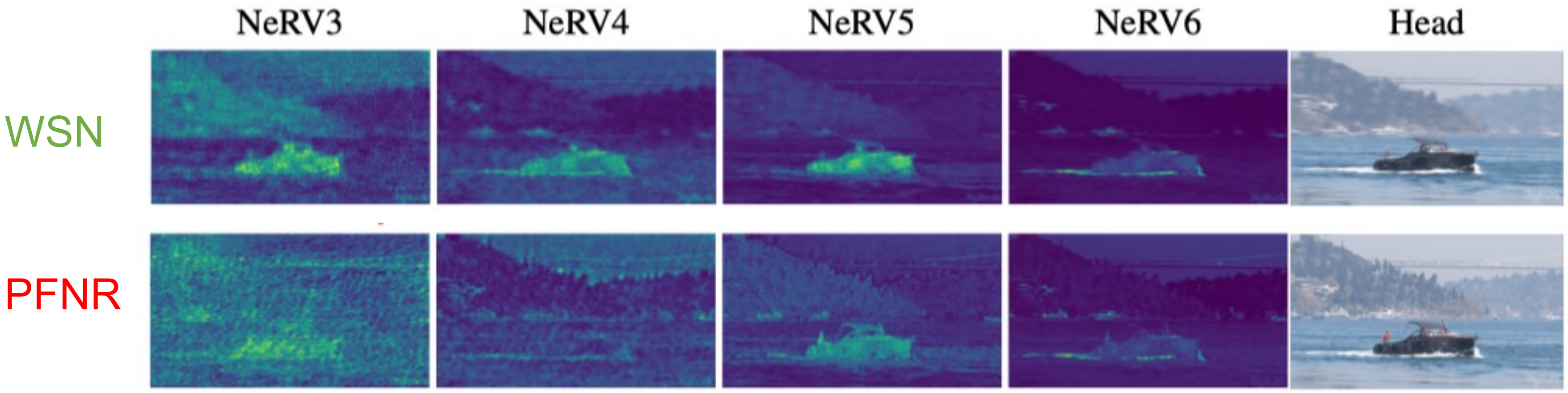


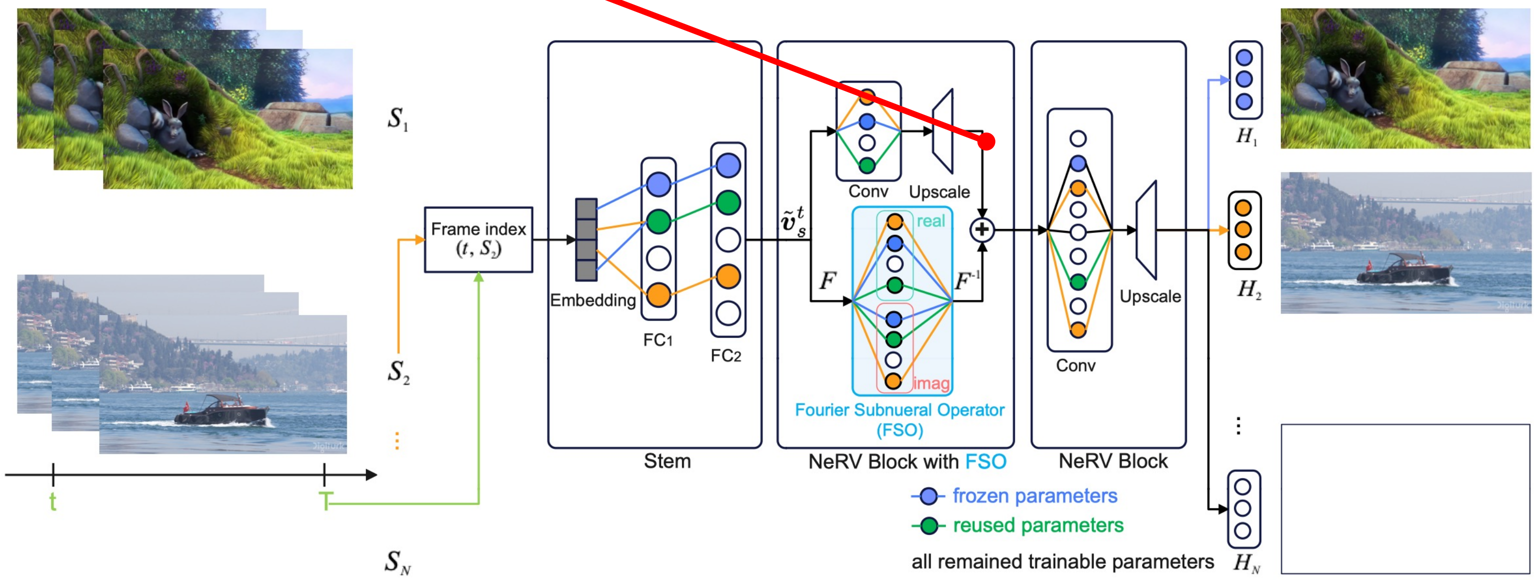
Continual Learning: Forget-free Winning Subnetworks for Video Representations – Haeyong Kang, 2024.



(+) High Feature Variance
(+) High Frequency Components

PFNR's Statistics of Variance and Frequency





(+) High Feature Variance
(+) High Frequency Components

➔ High Quality Video Represen.

PFNR for VCL

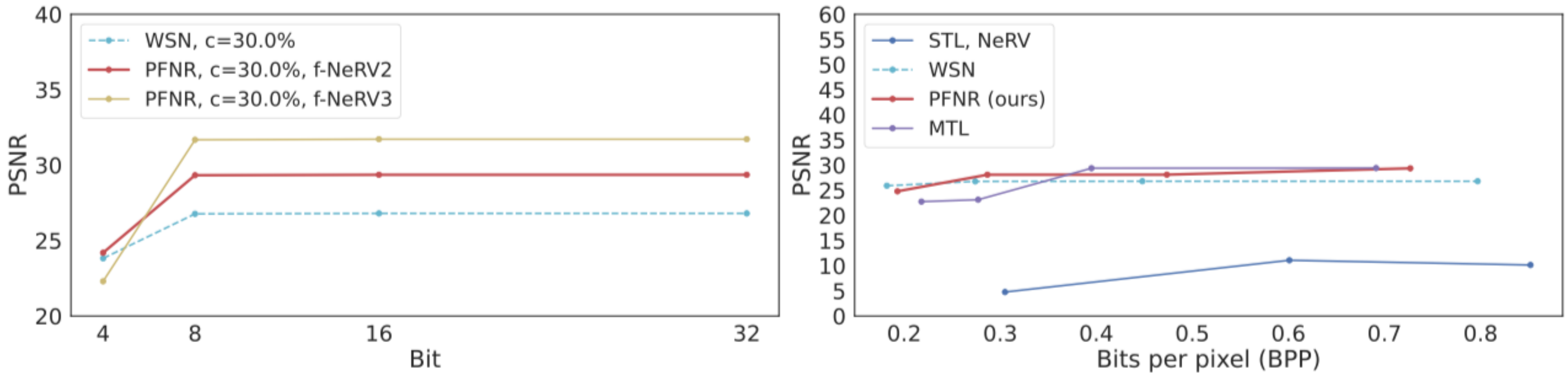
Algorithm 1 Progressive Fourier Neural Representation (PFNR) for VCL

input: $\{\mathcal{D}_s\}_{s=1}^{N}$, model weights of FSO $\boldsymbol{\theta}_* = \{\boldsymbol{\theta}, \boldsymbol{\phi}_{FSO}\}$, score weights of FSO $\boldsymbol{\rho}_* = \{\boldsymbol{\rho}, \boldsymbol{\rho}_{FSO}\}$, binary mask $\mathbf{M}_0 = \{\mathbf{0}^{|\boldsymbol{\theta}|}, \mathbf{0}^{|\boldsymbol{\theta}_{FSO}|}\}$, and layer-wise capacity $c\%$.

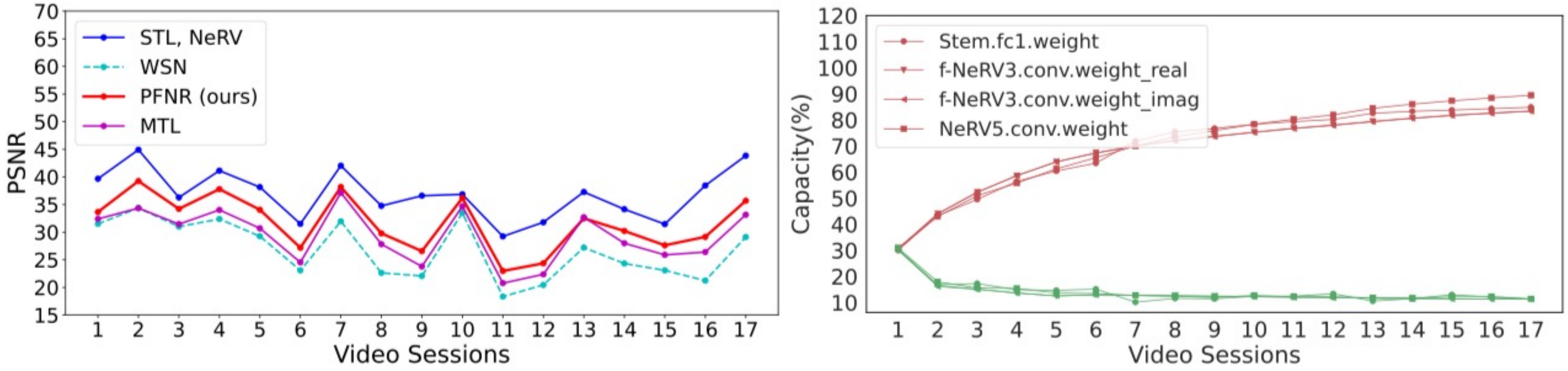
randomly initialize $\boldsymbol{\theta}_*$ and $\boldsymbol{\rho}_*$.
for session $s = 1, \cdots, |\mathcal{S}|$ **do**
  if $s > 1$ **then**
    randomly re-initialize $\boldsymbol{\rho}_*$.
  end if
  for batch $\mathbf{b}_t \sim \mathcal{D}_s$ **do**
    obtain mask $\mathbf{m}_s$ of the top-$c\%$ scores $\boldsymbol{\rho}_*$ at each layer
    compute $\mathcal{L}\left(f(\boldsymbol{e}_{s,t}; \boldsymbol{\theta}_* \odot \mathbf{m}_s), \mathbf{b}_t\right)$, where input embedding, $\boldsymbol{e}_{s,t} = [\boldsymbol{e}_s; \boldsymbol{e}_t]$.
    $\boldsymbol{\theta}_* \leftarrow \boldsymbol{\theta}_* - \eta\left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_*} \odot (\mathbf{1} - \mathbf{M}_{s-1})\right)$ ▷ trainable weight update
    $\boldsymbol{\rho}_* \leftarrow \boldsymbol{\rho}_* - \eta(\frac{\partial \mathcal{L}}{\partial \boldsymbol{\rho}_*})$ ▷ weight score update
  end for
  $\hat{\boldsymbol{\theta}}_s = \boldsymbol{\theta}_* \odot \boldsymbol{m}_s$
  $\mathbf{M}_s \leftarrow \mathbf{M}_{s-1} \vee \mathbf{m}_s$ ▷ accumulate binary mask
end for

output: $\{\hat{\boldsymbol{\theta}}_s\}_{s=1}^{N}$

Quantization and Bits per pixel (BPP)

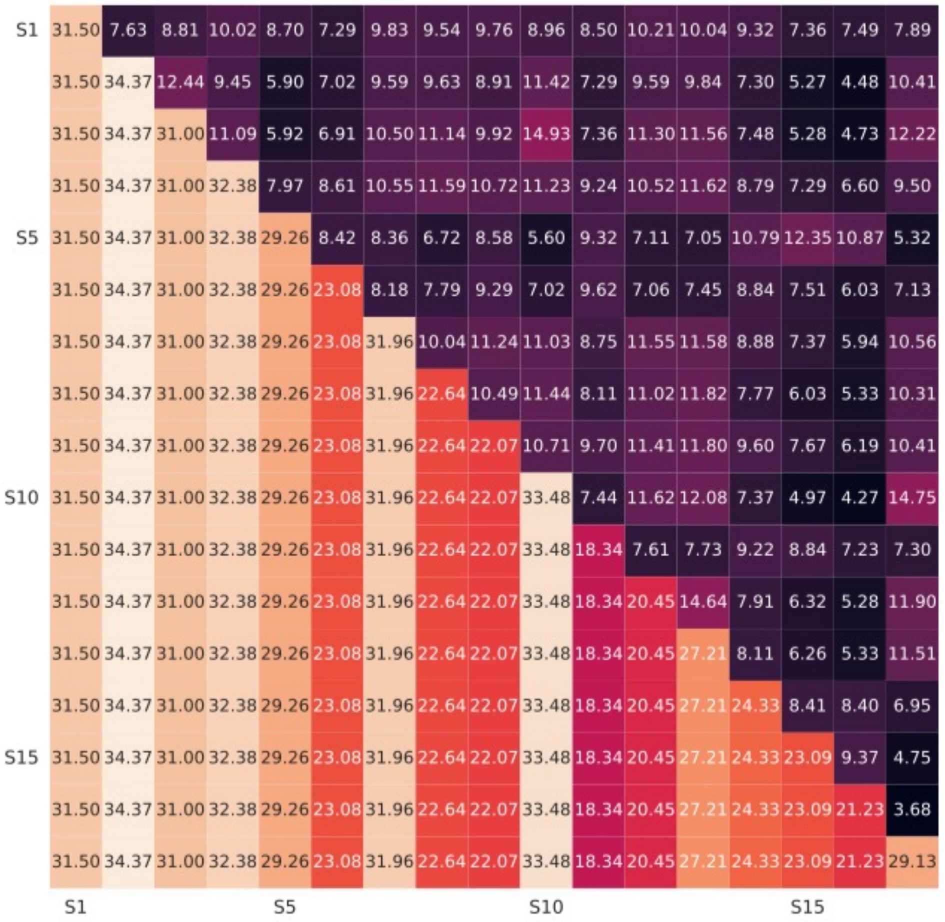


(+) Quantization of PFNR & Bits per pixel (BPP)

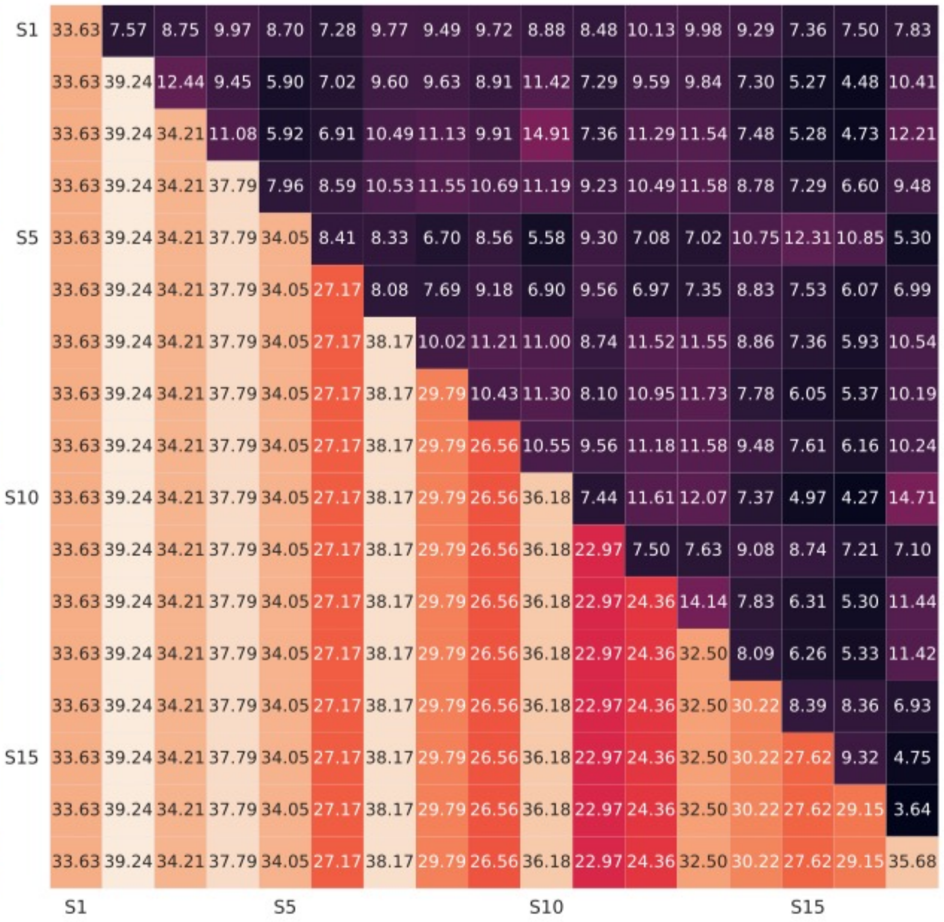


(+) PFNR's Performances and Progressive Model Capacity

PFNR’s Transfer Matrixes on the UVG17 datasets



(a) WSN, $c = 30.0\%$



(b) PFNR, $c = 30.0\%$, f-NeRV3

(+) PFNR’s Forget-free Video Continual Learning

PFNR's Performances on the UVG17 datasets

WSN (29.26, PSNR)

PFNR, *f*-NeRV2 (31.24, PSNR)

PFNR, *f*-NeRV3 (34.05, PSNR)

(+) PFNR's Video Generation

PFNR's Performances on the UVG17 datasets

(+) 8-bit PFNR's Video Generation

Conclusions

- Neural Implicit Representation (NIR) has recently gained significant attention due to its remarkable ability to encode complex and high-dimensional data into representation space and easily reconstruct it through a trainable mapping function.

- However, NIR methods assume a one-to-one mapping between the target data and representation models regardless of data relevancy or similarity. This results in poor generalization over multiple complex data and limits their efficiency and scalability.

- Motivated by continual learning, this work investigates how to accumulate and transfer neural implicit representations for multiple complex video data over sequential encoding sessions. To overcome the limitation of NIR, we propose a novel method, Progressive Fourier Neural Representation (PFNR), that aims to find an adaptive and compact sub-module in Fourier space to encode videos in each training session.

- This sparsified neural encoding allows the neural network to hold free weights, enabling an improved adaptation for future videos. In addition, when learning a representation for a new video, PFNR transfers the representation of previous videos with frozen weights.

- This design allows the model to continuously accumulate high-quality neural representations for multiple videos while ensuring lossless decoding that perfectly preserves the learned representations for previous videos. We validate our PFNR method on the UVG8/17 and DAVIS50 video sequence benchmarks and achieve impressive performance gains over strong continual learning baselines.