

# Revisiting Deep Video Motion Magnification for Real-time Applications

\*Hyunwoo Ha<sup>1</sup> \*Oh Hyun-Bin<sup>1</sup> Kim Jun-Seong<sup>1</sup> Kwon Byung-Ki<sup>1</sup> Kim Sung-Bin<sup>1</sup>  
Ji-Yun Kim<sup>1</sup> Sung-Ho Bae<sup>2</sup> Tae-Hyun Oh<sup>1</sup>  
<sup>1</sup>POSTECH    <sup>2</sup>Kyung Hee University

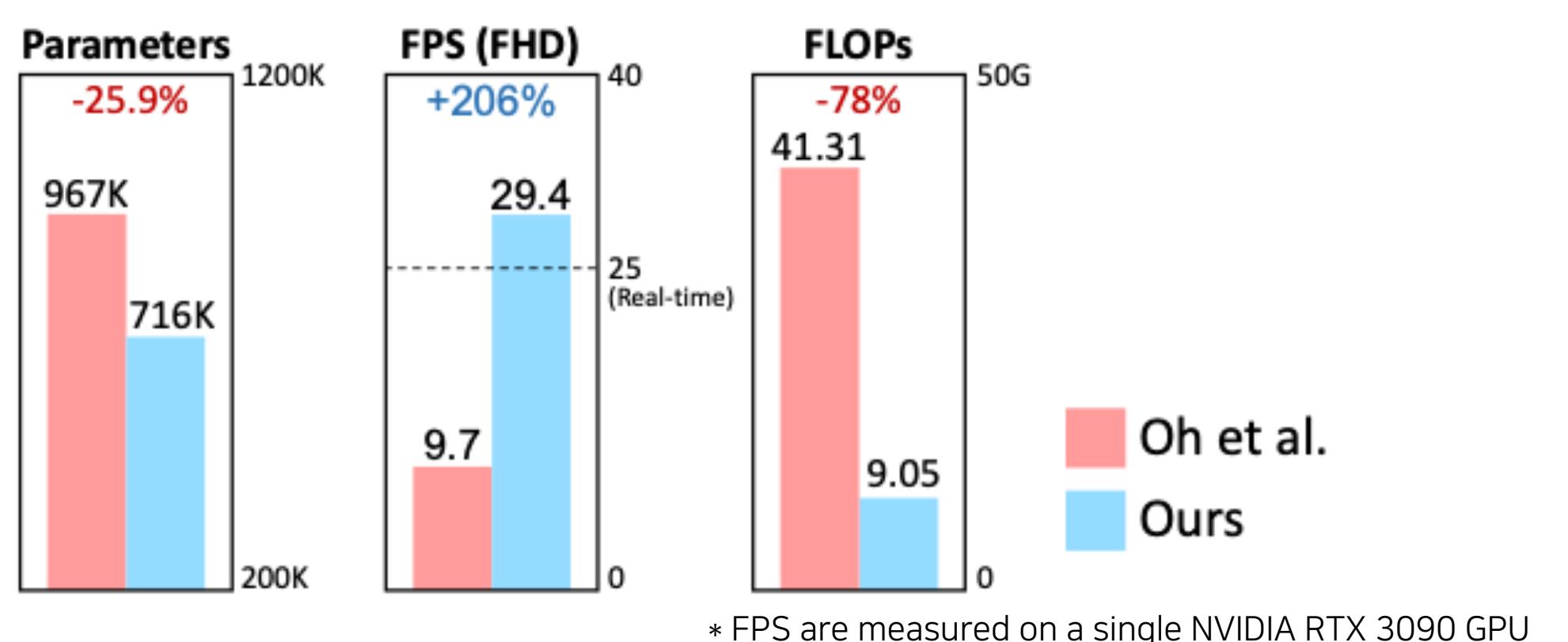
## Motivation

- The de-facto standard deep motion magnification model, *Oh et al.*, still falls behind real-time performance
- Real-time processing would open-up online applications, i.e., monitoring for safety or medical purpose



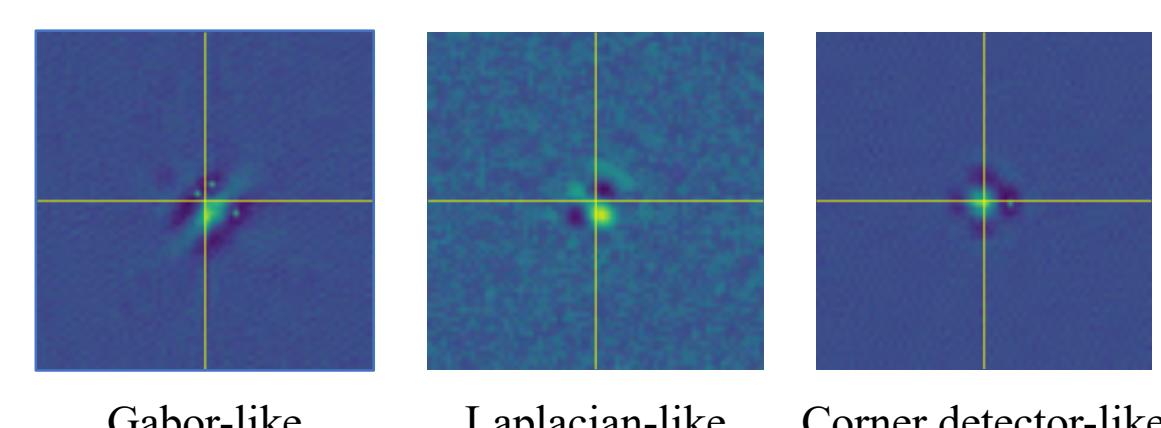
## Contribution 🚀

- Introduce an efficient motion magnification model achieving **real-time\* performance on Full-HD videos**
- Propose a two-stage architecture exploration framework for inhomogeneous architectures



## Findings & Interesting Points 🔎

- A **linear neural network** is sufficient for **Encoder**
- An extremely **asymmetric architecture** is more preferred for real-time video motion magnification



## Two-stage Architecture Exploration

### Stage1. Learning-to-remove Method

- Parameterize:  

$$F(x) = (1 - \omega)A_O(x) + \omega \cdot x$$
  - \*  $\omega$  is learnable switch parameter
- Add bias loss term:  

$$\mathcal{L}_{rm} = \frac{1}{K} \sum_{k \in K} (1 - \omega_k^p)$$
- Minimize:  

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda_{rm}\mathcal{L}_{rm}$$

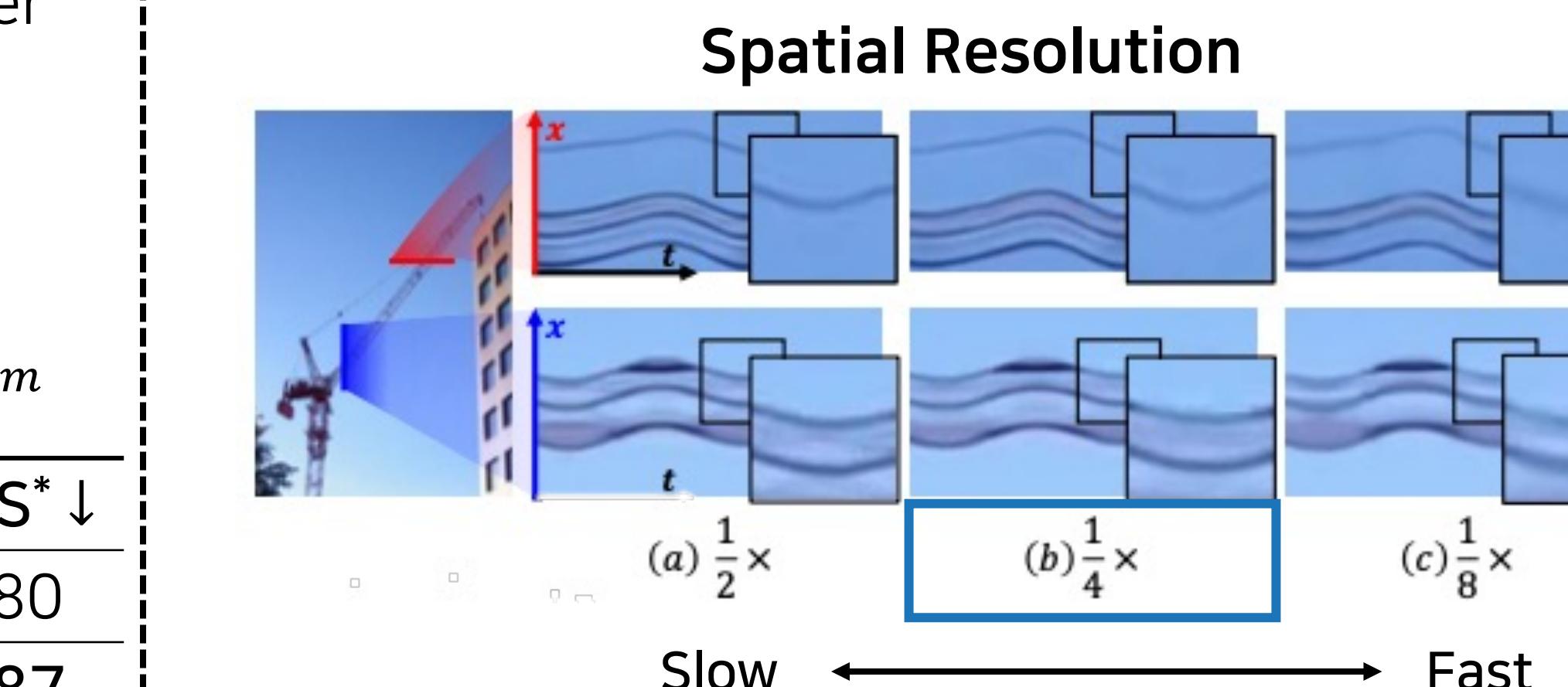
Module	ReLU	Block	GFLOPs*	SSIM*↑	LPIPS*↓
[Oh et al.]	—	—	41.3	0.932	0.180
Encoder	X	—	41.3	0.929	0.187
—	—	X	37.6	0.928	0.186
Manipulator	X	—	41.3	0.930	0.181
—	—	X	40.6	0.930	0.182
Decoder	△	—	41.3	0.902	0.252
—	—	△	25.0	0.864	0.317

\* Metrics are measured on synthetic data of resolution 384x384

**Discard residual blocks in the encoder & manipulator!**

### Stage2. Hyperparameter Analysis

The decoder is vulnerable to removal of layers...  
So, we analyze details of architectural hyperparameters



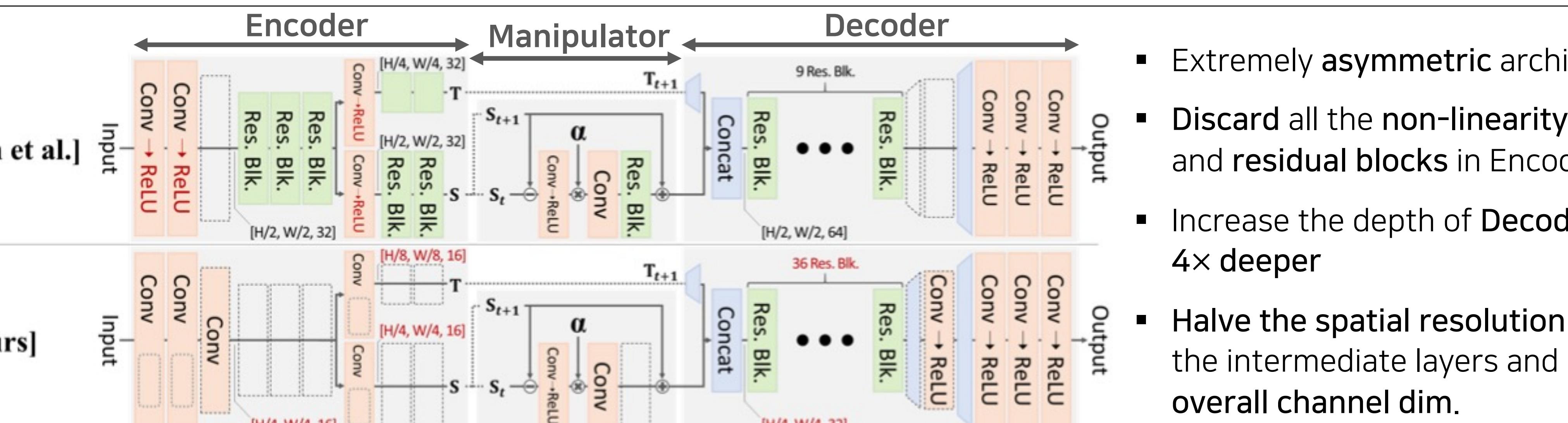
### Balance b/w Depth & Channel

Depth	Channel	GFLOPs*	SSIM*↑	LPIPS*↓	FPS*
9	64	41.3	0.921	0.191	83.2
36	32	41.5	0.928	0.172	85.5
144	16	41.6	0.925	0.175	46.9

\* Metrics are measured on synthetic data of resolution 384x384

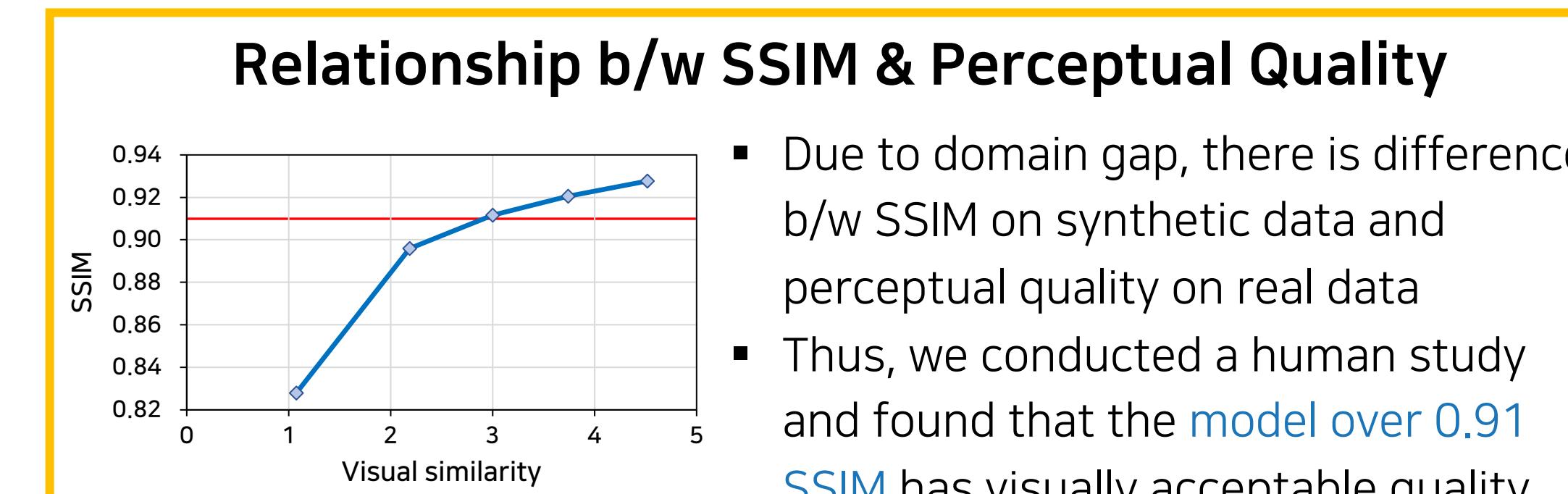
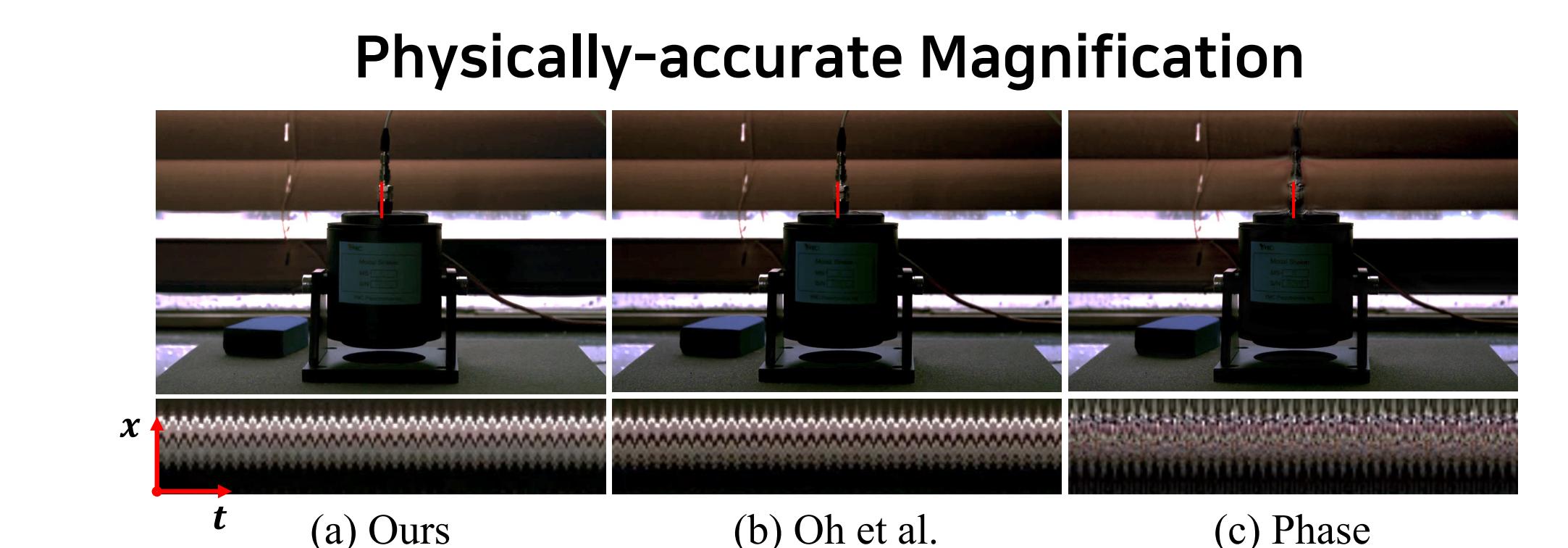
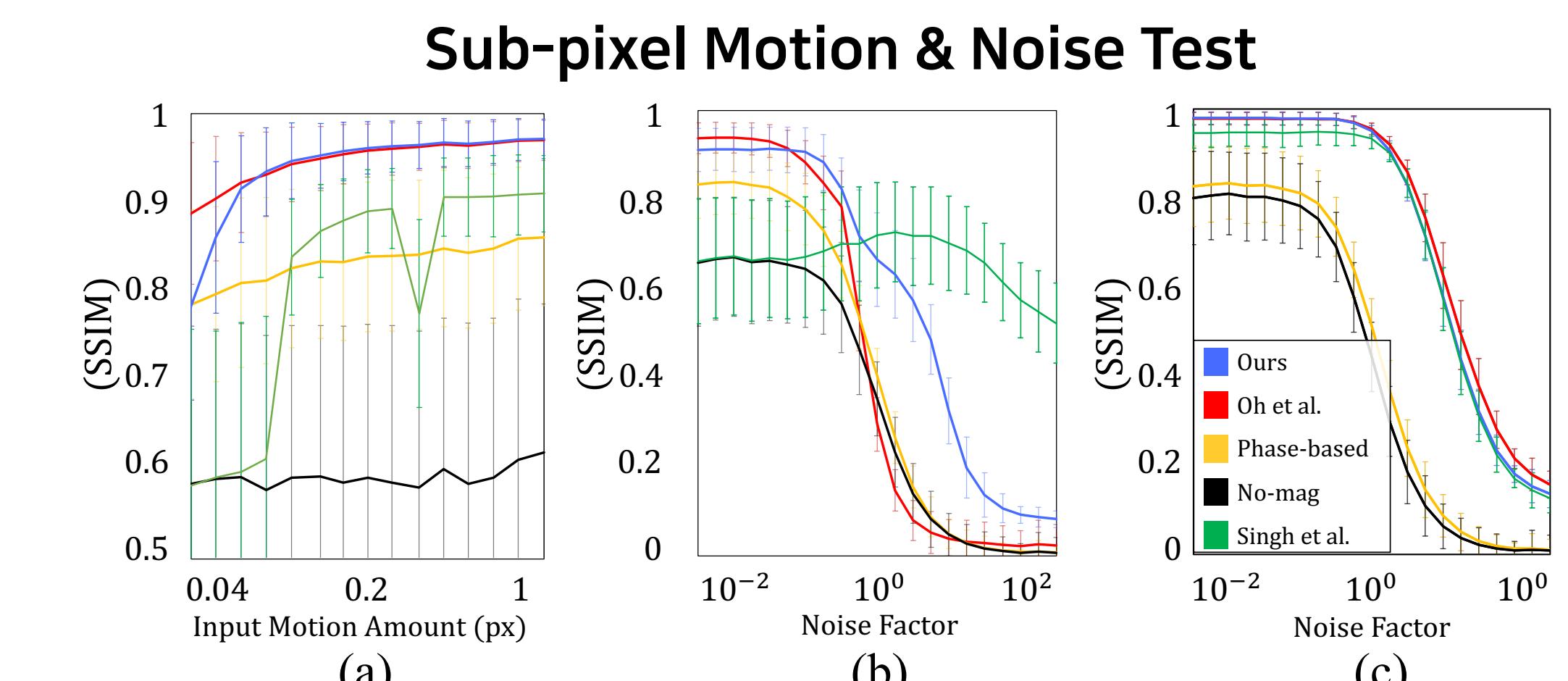
**Adopt above found hyperparameters!**

## Proposed Architecture



- Extremely asymmetric archi.
- Discard all the non-linearity and residual blocks in Encoder
- Increase the depth of Decoder 4x deeper
- Halve the spatial resolution of the intermediate layers and overall channel dim.

## Results



**Acknowledgment.** This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00290, Visual Intelligence for Space-Time Understanding and Generation based on Multi-layered Visual Common Sense). This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.NRF-2021R1C1C1006799).