Global Pooling, More than Meets the Eye: Position Information is Encoded Channel-Wise in CNNs

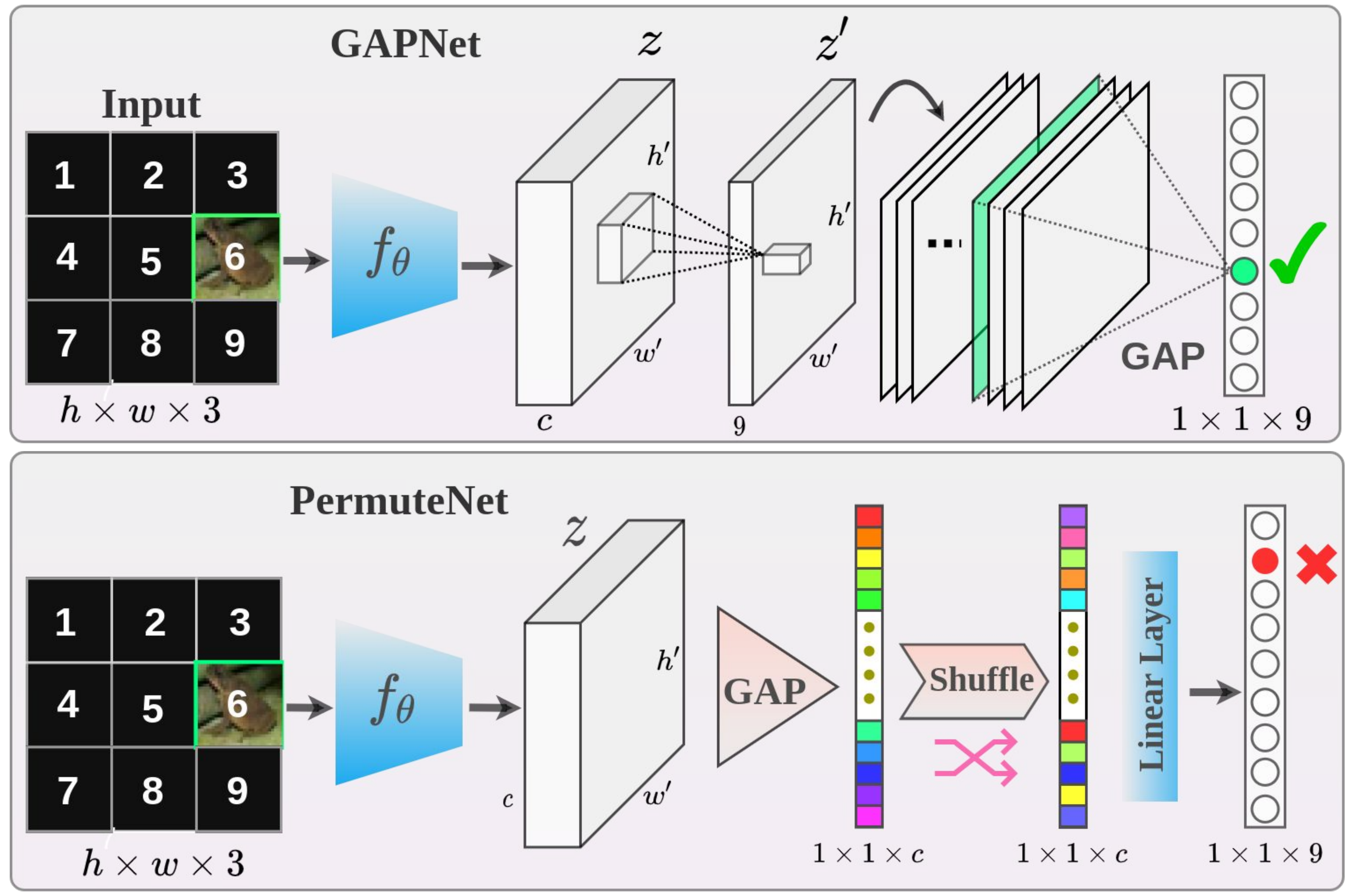
Md Amirul Islam*[1,6] Matthew Kowal*[2,6] Sen Jia[4]
Konstantinos G. Derpanis[2,5,6] Neil D. B. Bruce[3,6]
[1]Ryerson University, [2]York University, [3]University of Guelph, [4]LG Electronics, [5]Samsung AI Centre Toronto, [6]Vector Institute

2021 ICCV OCTOBER 11-17 VIRTUAL

Motivation

- ➢ We challenge the common assumption that collapsing the spatial dimensions of a 3D tensor into a vector via global pooling removes all spatial information.
- ➢ How can a CNN contain positional information in the representations if there exists a global average pooling layer in the forward pass?
- ➢ We hypothesize that the position information is encoded within the ordering of the channel dimensions.

Learning Positions with a GAP Layer



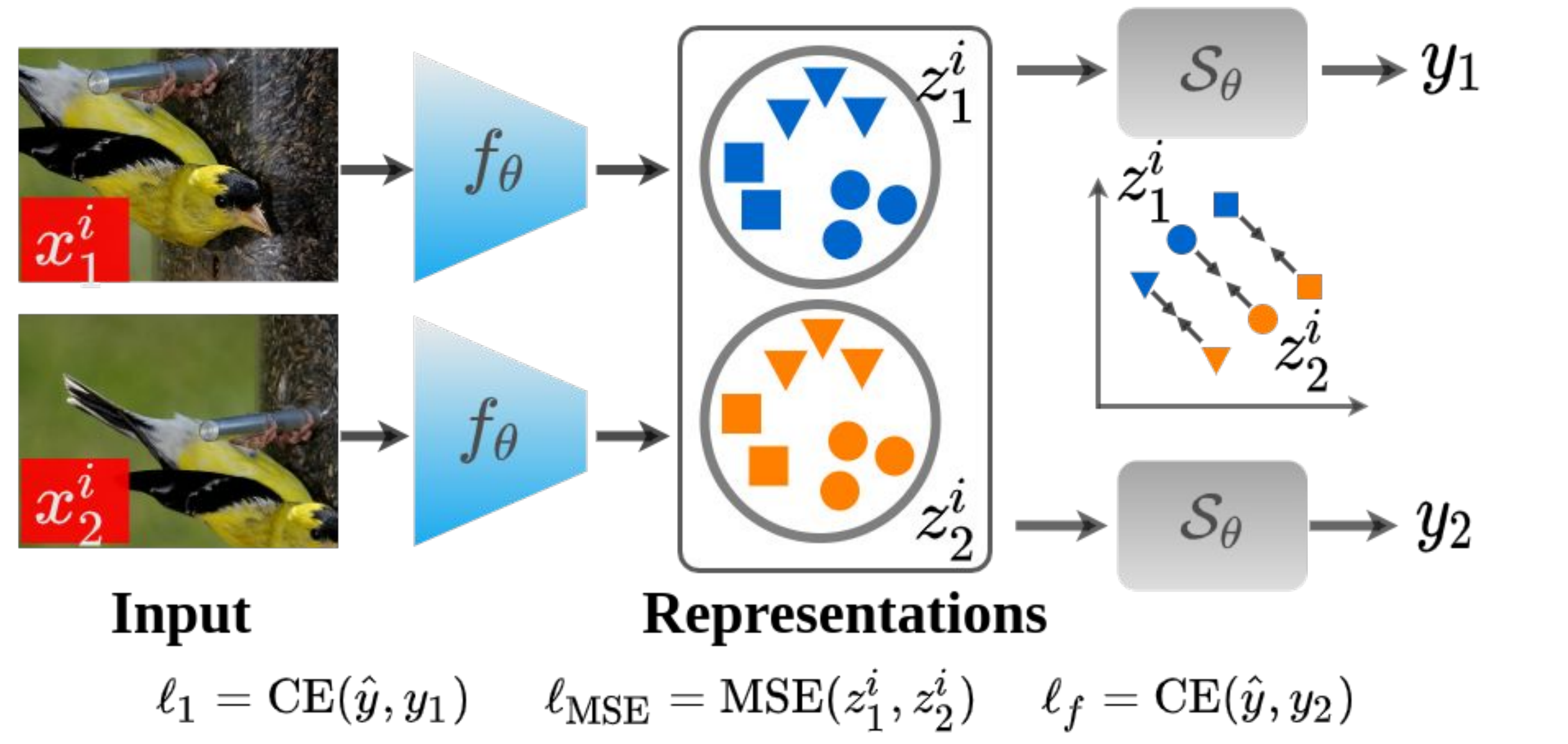
- ➢ GAPNet transforms the latent representation to a representation where *number of channels* matches with the *number of locations in the grid* and **outputs the location of the image patch** placed on a black canvas.
- ➢ PermuteNet follows the structure of a standard classification network and applies a single random **permutation** of the channels between the GAP *and the prediction layer*.

Channel-wise Position Encoding

Network	*Loc. Classification*		*Image Classification*	
	3x3	*7x7*	*3x3*	*7x7*
GAPNet	100	100	82.6	82.1
PermuteNet	78.8	21.4	73.6	69.9

- ➢ **GapNet** achieves 100% location accuracy while **PermuteNet** fails to correctly classify locations.
- ➢ *The order of the channel* dimensions is the main representational capacity which allows for the GAP layer to admit absolute position information.

Translation Invariance in CNNs

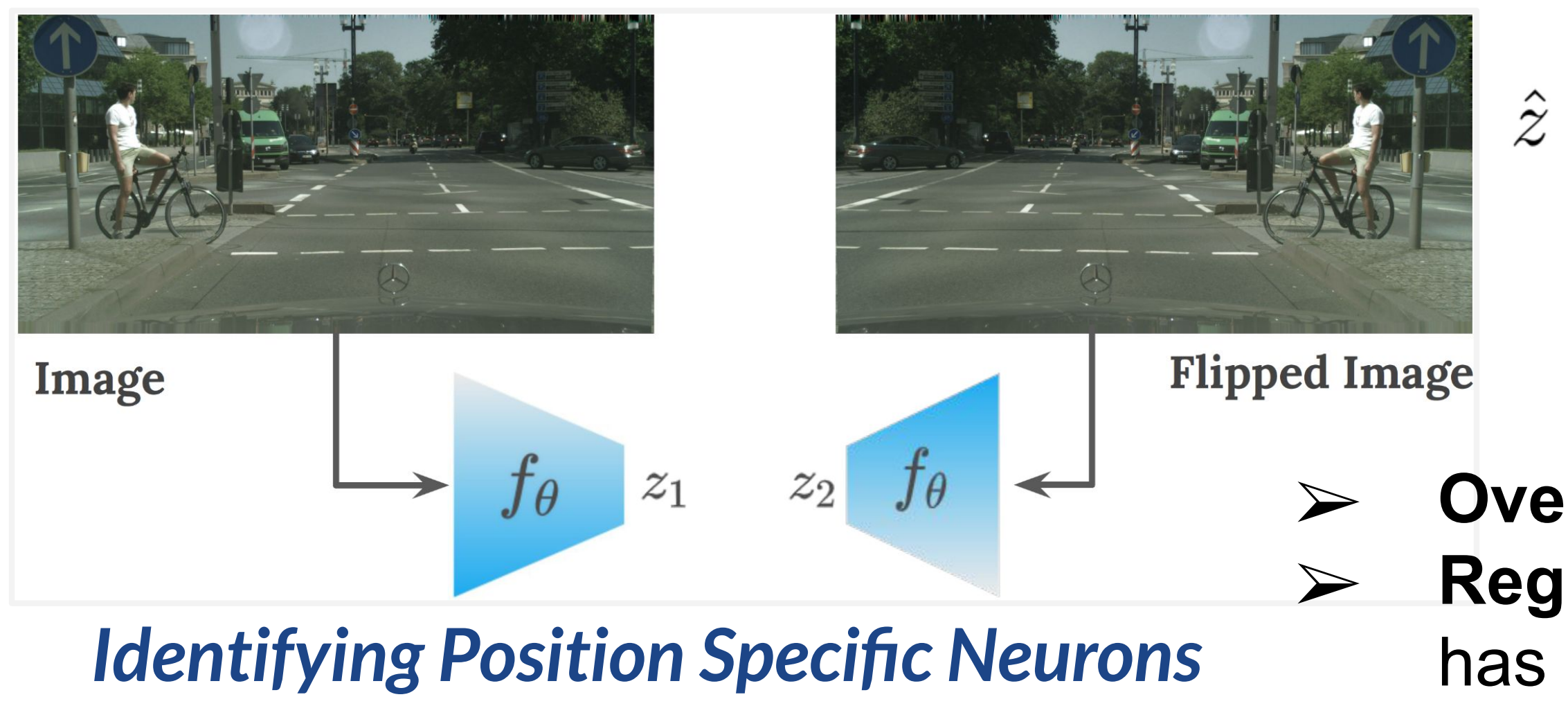

x_1^i
x_2^i
f_θ
f_θ
z_1^i
z_2^i
S_θ
S_θ
y_1
y_2
Input
Representations
$\ell_1 = \mathrm{CE}(\hat{y}, y_1)$
$\ell_{\mathrm{MSE}} = \mathrm{MSE}(z_1^i, z_2^i)$
$\ell_f = \mathrm{CE}(\hat{y}, y_2)$

- ➢ *Minimize* the distance between two globally pooled latent representations of **different shifts** of the same image

Network	*CIFAR-10*		*CIFAR-100*	
	Top-1 Acc.	*Consistency*	*Top-1 Acc.*	*Consistency*
ResNet-18	93.1	90.8	72.6	70.1
Blurpool	92.5	92.5	72.4	78.2
AugShift (Ours)	92.1	94.8	72.6	85.6

- ➢ *AugShift* **improves** the overall classification performance and the shift consistency on **CIFAR-10** and **CIFAR-100**.

Attacking the Position Encoding Channels: Overall & Region-Specific

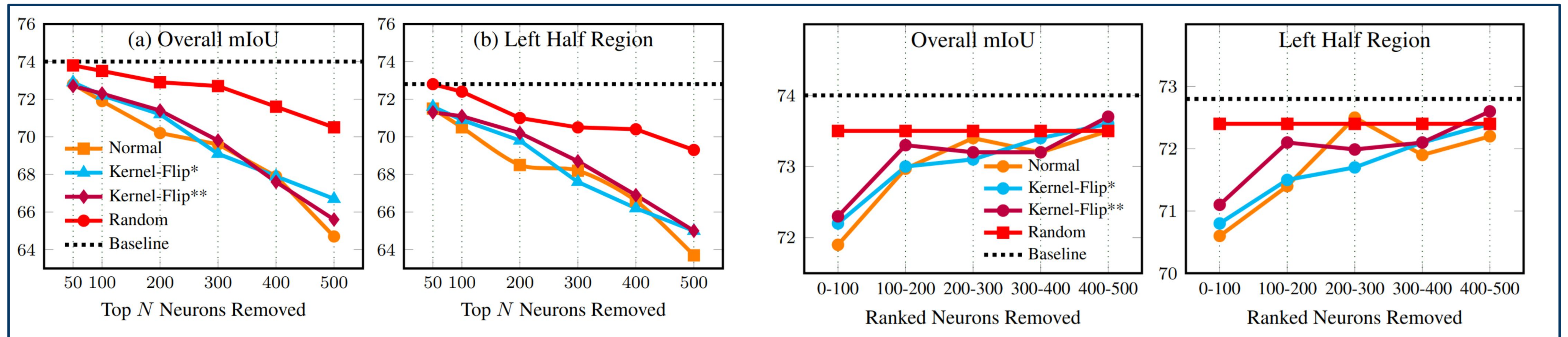


Identifying Position Specific Neurons

$$\underbrace{\hat{z} = \mathtt{argsort}_{j \in C}\left[\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}|\Delta z_i|\right]}_{\text{Overall (Absolute Difference)}} \qquad \underbrace{\hat{z}^l = \mathtt{argsort}_{j \in C}\left[\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\Delta z_i\right]}_{\text{Region-specific (Signed Difference)}}$$

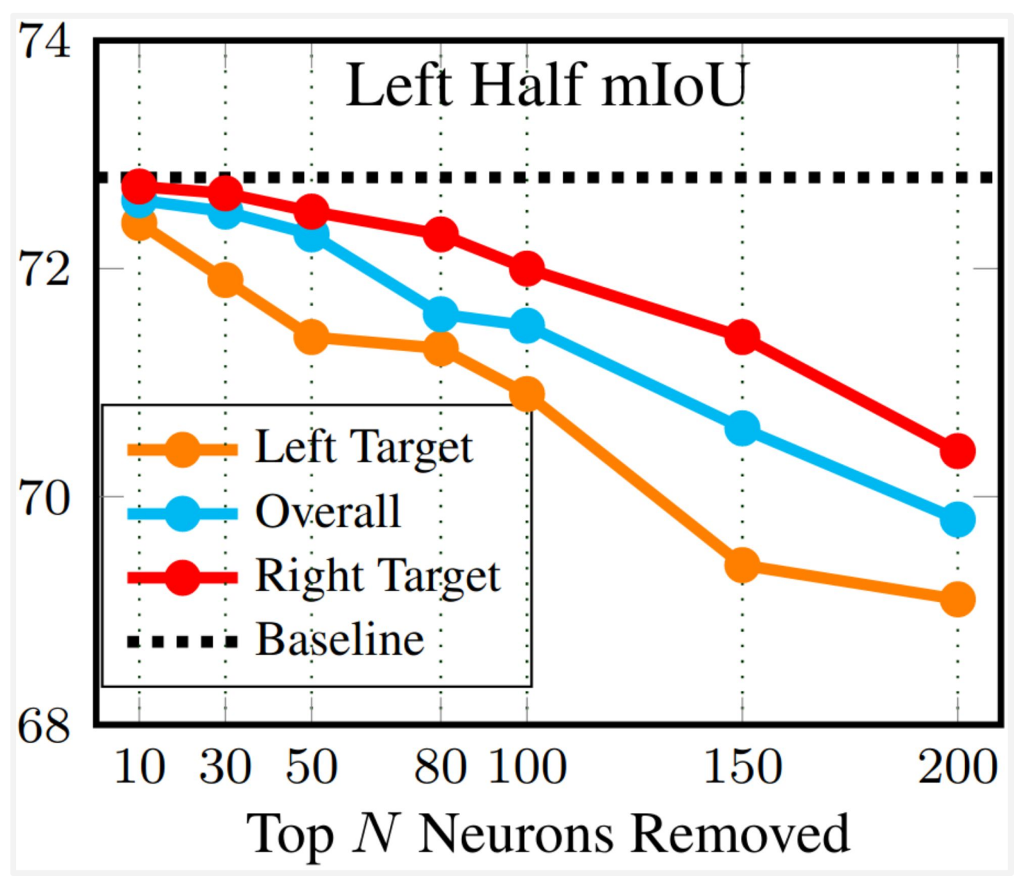
- ➢ **Overall:** Position-specific channels will have higher difference in activation.
- ➢ **Region-specific:** Highly activated when an object is in a particular location, and has a low activation otherwise.

Targeting Overall Position Specific Neurons: Semantic Segmentation



- ➢ Gradually removing the *position-specific* neurons **hurts** the overall mIoU significantly more than *random* neurons

Targeting Region-Specific Neurons



- ➢ Segmentation results when top N region specific channels are removed during **inference**
- ➢ *Left-encoding* channels should harm the mIoU on the **left half** more than the *right-encoding* channels.

Conclusions

- ➢ Position information is encoded based on the *ordering* of the channels while semantic information is largely not.
- ➢ Introduced a simple data augmentation strategy to improve translation invariance of CNNs.
- ➢ Introduced a intuitive technique to identify the position-specific channels in a network's latent representation.