北京交通大學

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

ICCV 2021

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, Canada [2]York University, Canada [3]University of Guelph, Canada

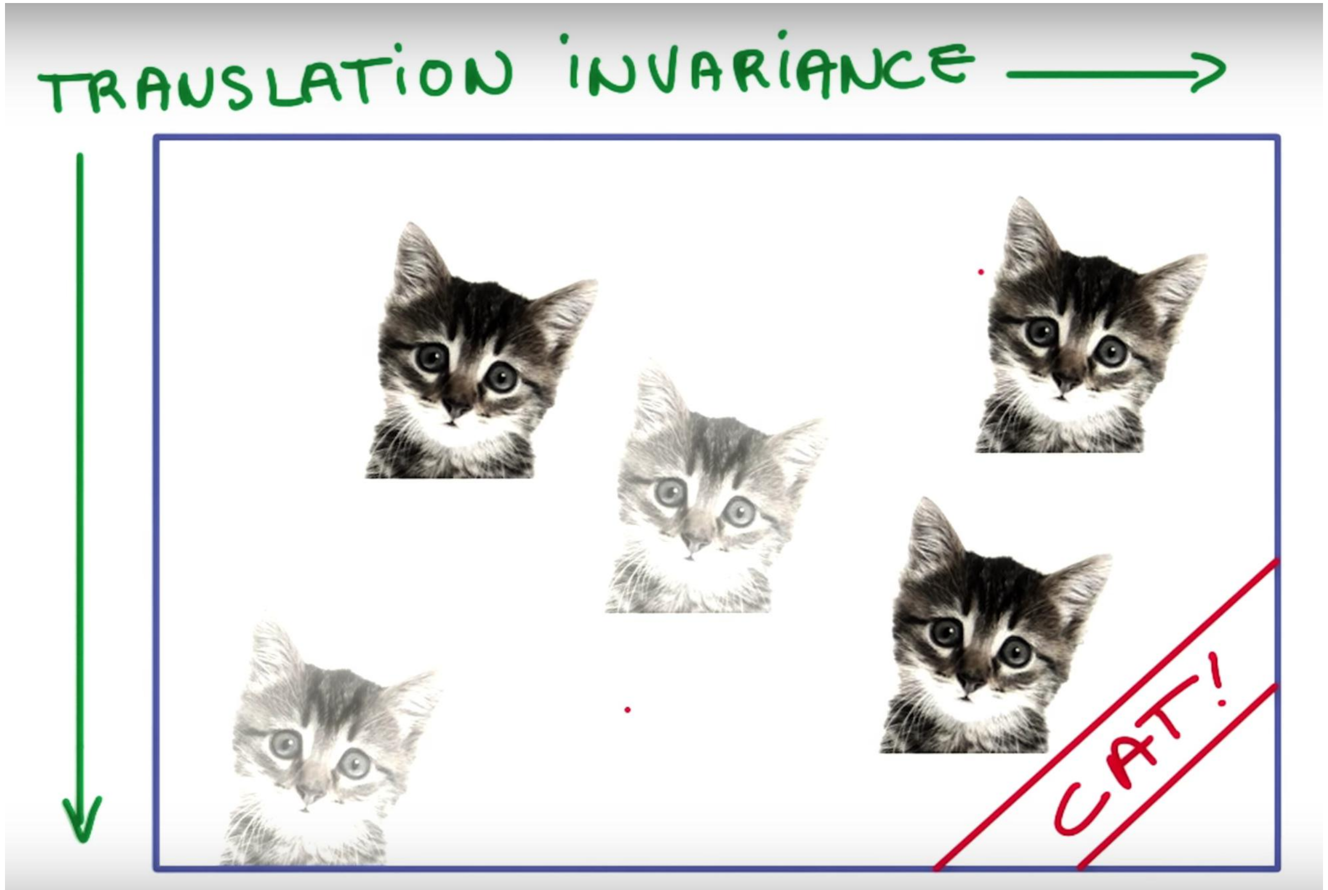
[4]Toronto AI Lab, LG Electronics [5]Samsung AI Centre Toronto, Canada [6]Vector Institute for AI, Canada

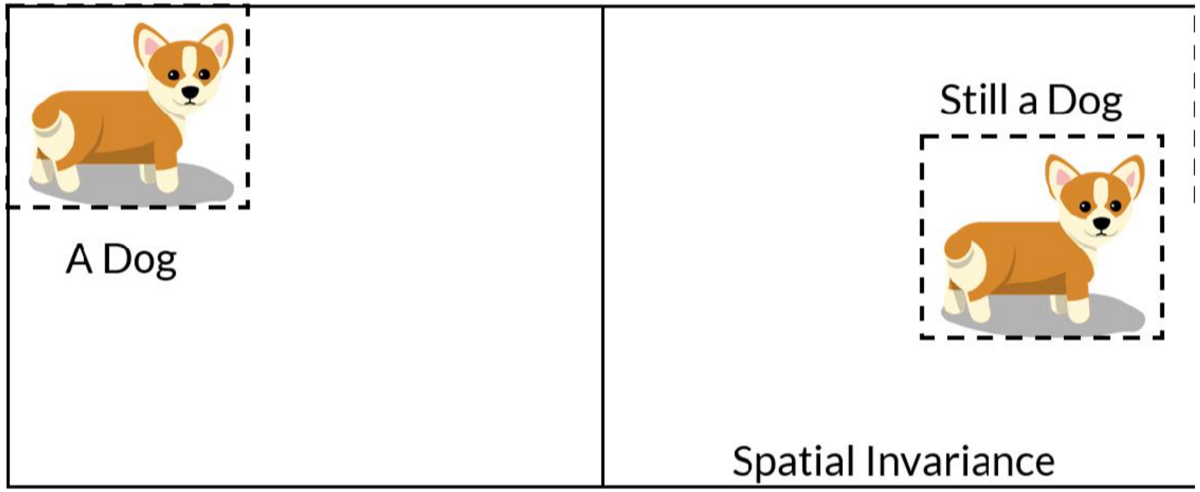
amirul@cs.ryerson.ca, {m2kowal,kosta}@eecs.yorku.ca, sen.jia@lge.com, brucen@uoguelph.ca

数字媒体信息处理研究中心
Center of Digital Media Information Processing

Kang Liao

Motivation

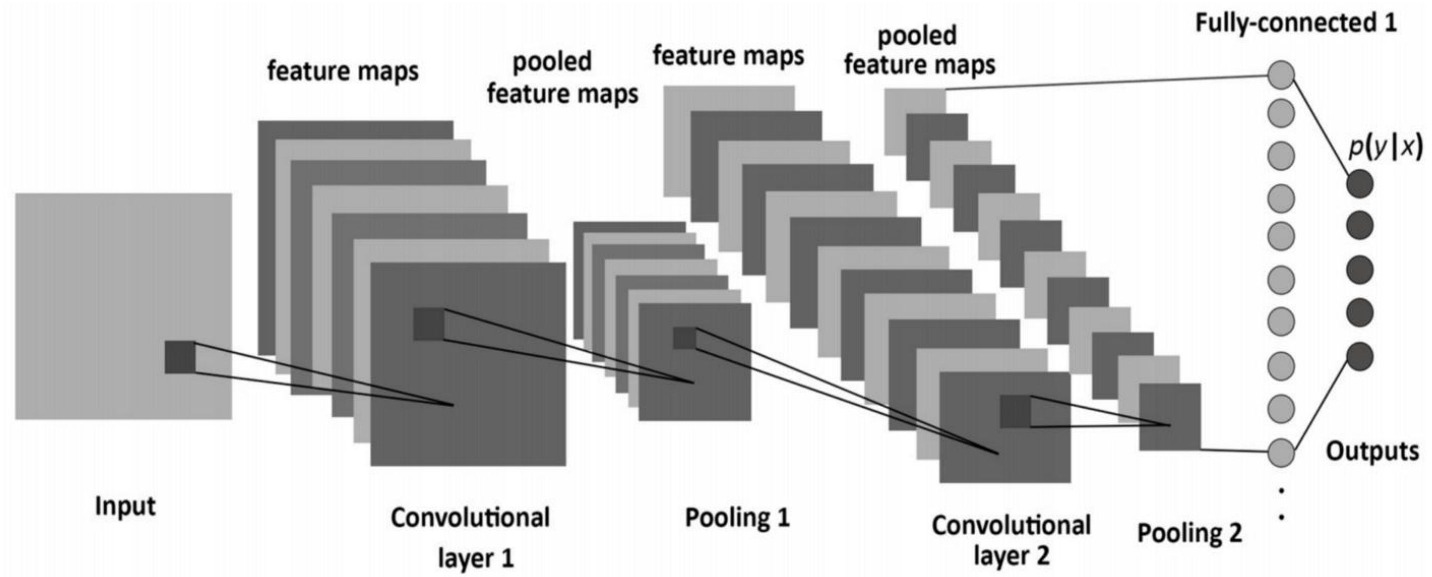




Translation Invariance

Motivation

Class 1: Top-left   Class 2: Bottom-right



1 or 2?

Surprisingly, CNNs can classify perfectly.

Framework: Explore the encoded position information in CNNs

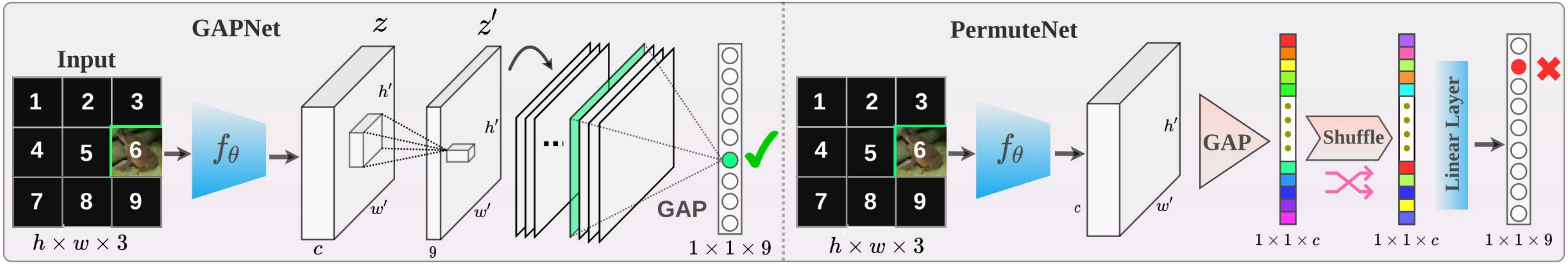
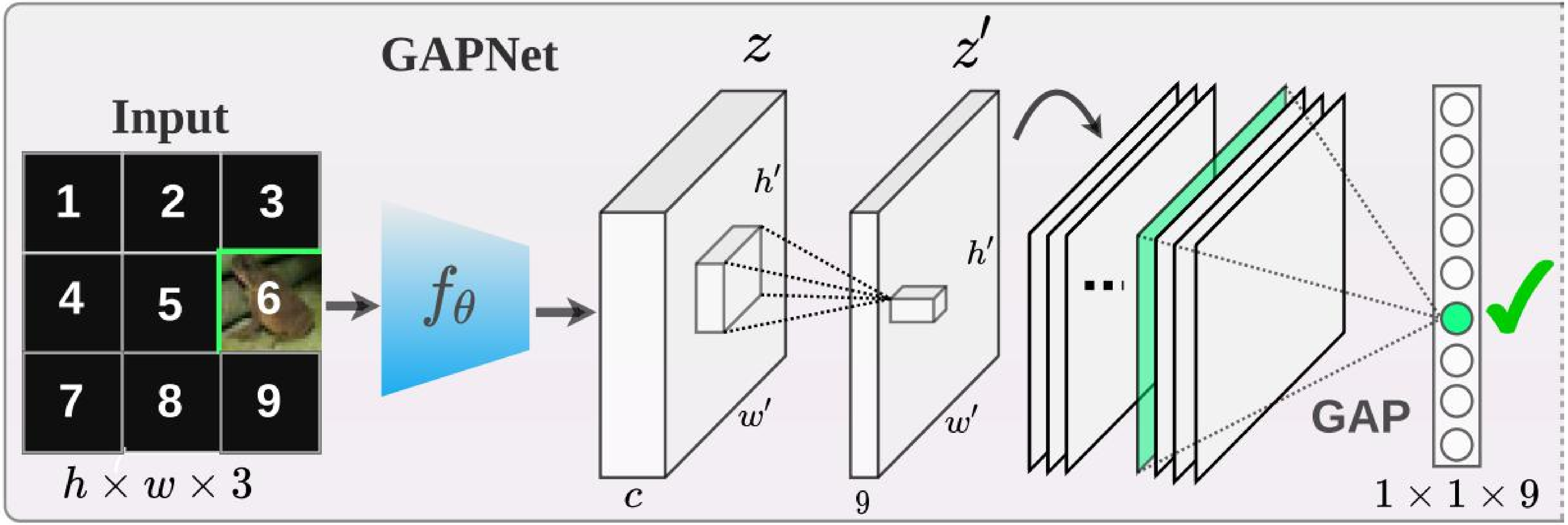


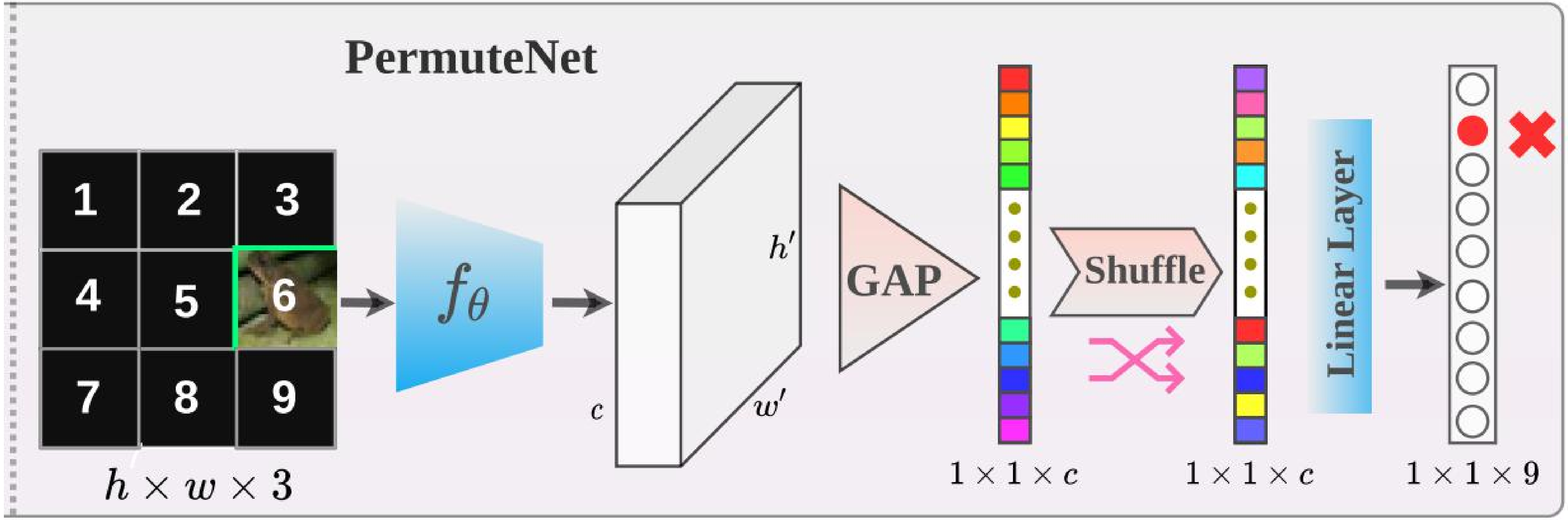
Figure 1. An illustration of our GAPNet (left) and PermuteNet (right) architectures used to determine the existence of channel-wise positional encodings in CNNs. **Left:** We feed a grid based input image to the encoder, f_θ, of standard CNNs (e.g., ResNet-18 [13]) to obtain a latent representation, z. z is then transformed to a representation, z', through the last convolutional layer which has the output channel dimension set to the number of locations in the input grid (e.g., 9 in the above example). This enforces the global average pooling (GAP) layer to output the number of locations. The network is then trained to predict the location of the image patch. **Right:** PermuteNet follows the same structure of a standard CNN except we shuffle the dimensions of the latent representation to verify whether obfuscating the channel ordering hurts the positional encoding capacity.

Framework: Explore the encoded position information in CNNs



GAPNet follows CNN (e.g., ResNet-18 and NIN) for object recognition, except we **remove the final fully connected layer**, such that the last layer of the network is the GAP layer.

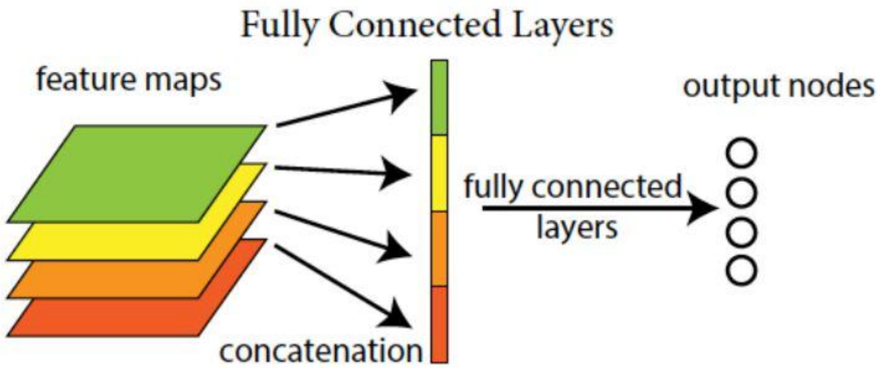
Framework: Explore the encoded position information in CNNs



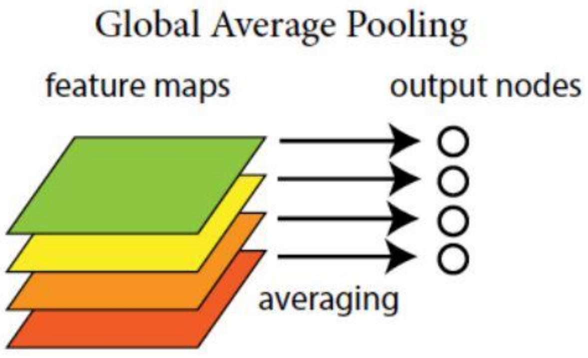
PermuteNet also follows the structure of a standard object classification network, except for a **single shuffle operation** which occurs between the GAP layer and the penultimate linear layer

GAP

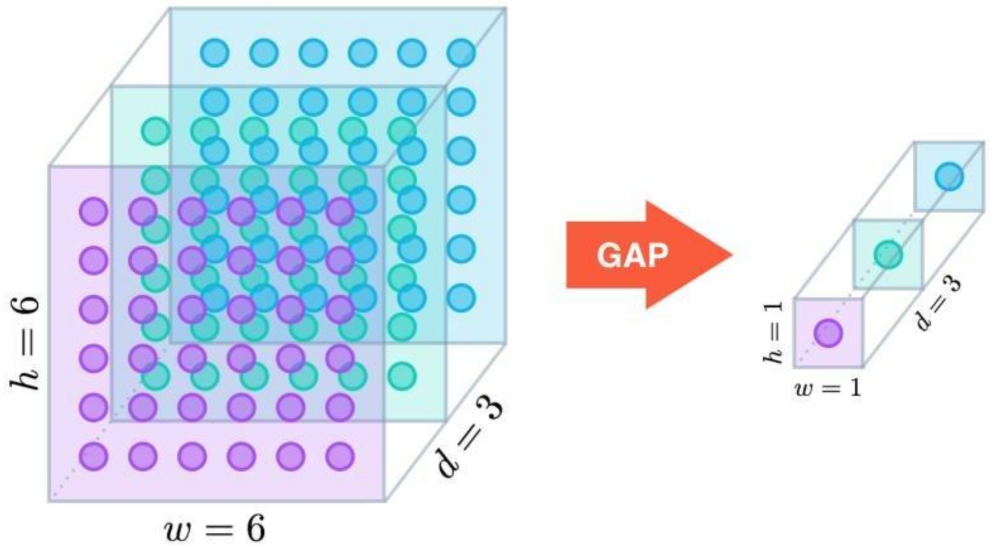
CNN



NIN



Explicitly confidence map of each category

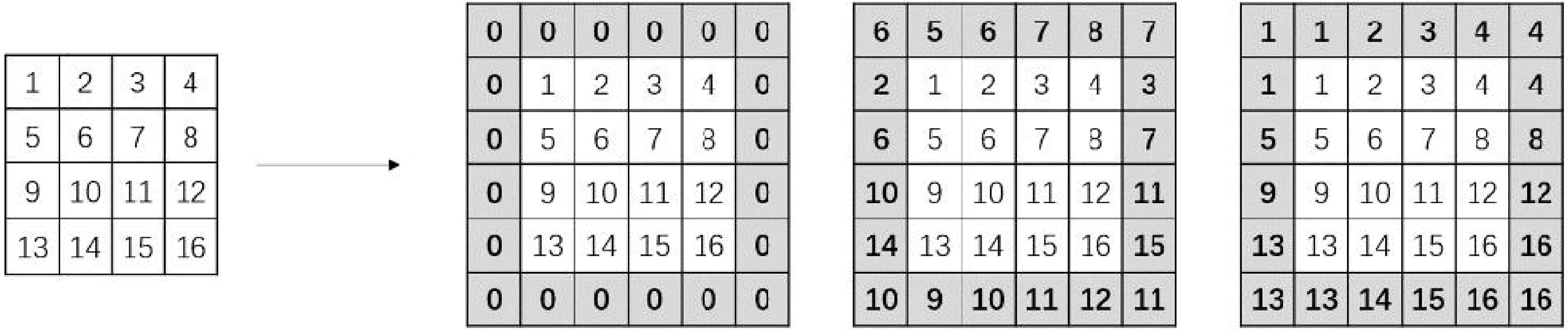


Experimental results

Padding	Network	Loc. Cls. Acc (%)			Image Cls. Acc (%)		
		3×3	5×5	7×7	3×3	5×5	7×7
Zero	GAPNet	100	100	100	82.6	82.4	82.1
	PermuteNet	78.8	37.8	21.4	73.6	72.2	69.9
Reflect	GAPNet	100	100	100	83.8	83.4	82.9
	PermuteNet	78.3	36.3	23.2	71.1	71.4	65.7
Replicate	GAPNet	100	100	100	83.1	82.9	82.8
	PermuteNet	78.3	40.1	23.6	72.0	71.5	71.4

Position information depends mainly on the **ordering of the channels**, while semantic information does not.

Different paddings

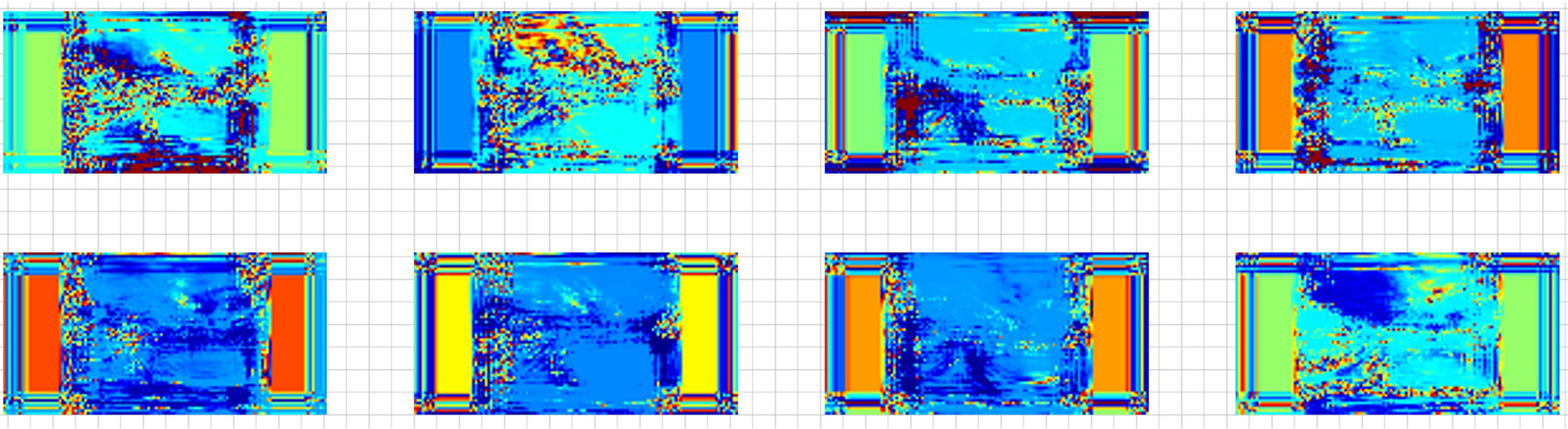


a) Zero Padding

b) Reflection Padding

c) Ruplication Padding

My experiments



Feature map visualization of different channels in a convolutional layer

Zero padding introduces the **line artifacts** in different feature maps, which indicate different position information.

Any Applications?

Learning Translation Invariant Representations

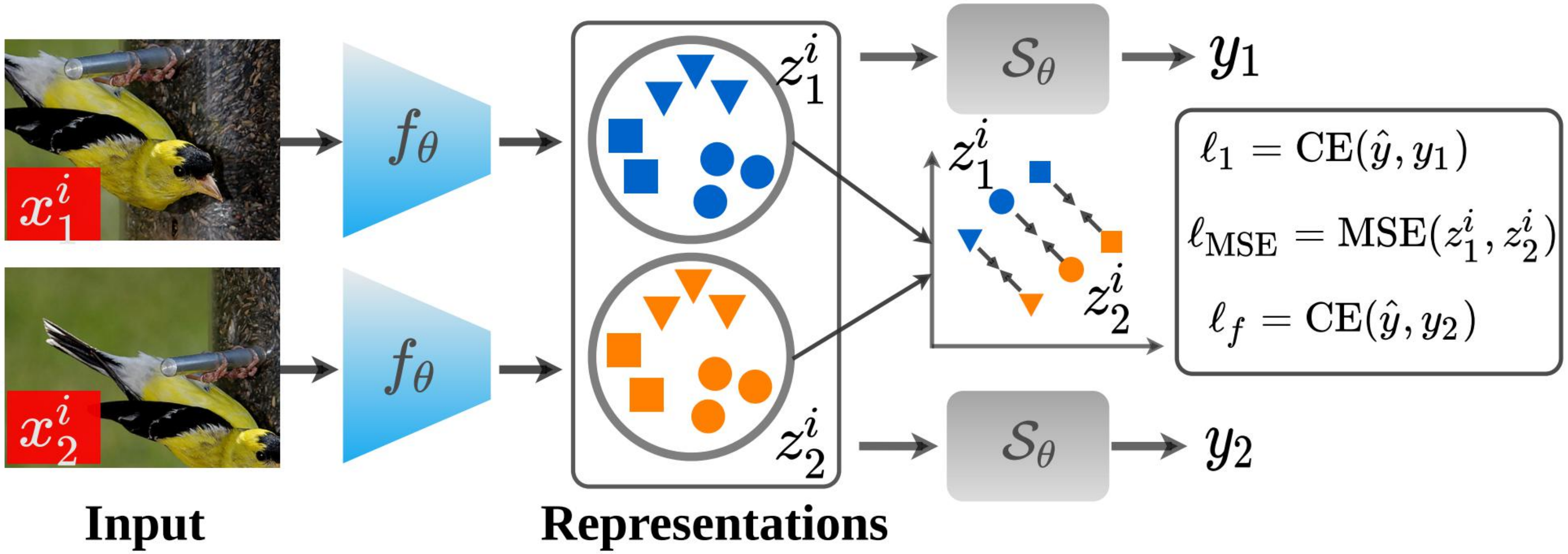


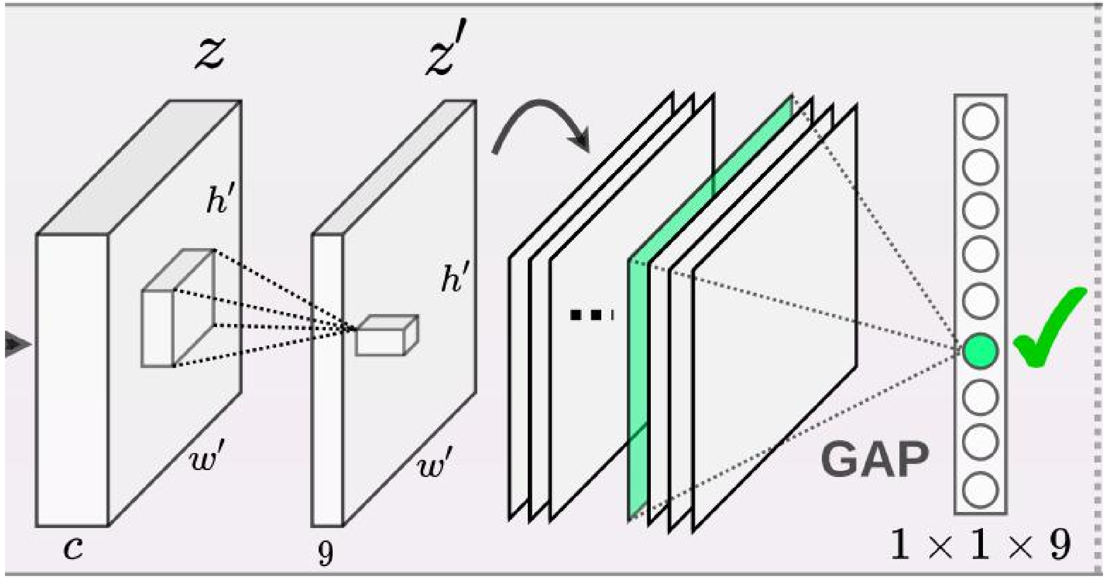
Figure 2. Illustration of the overall training pipeline of our proposed translation invariant model. We use two different crops of the input image, x^i to generate x_1^i and x_2^i, which are then both passed to a convolutional encoder network, f_θ, to obtain latent representations,

Learning Translation Invariant Representations

Methods	CIFAR-100 [20]		CIFAR-10 [20]		ImageNet [8]		
	Top-1	Cons.8	Top-1	Cons.8	Top-1	Cons.8	Cons.16
ResNet-18 [13]	72.6	70.1	93.1	90.8	69.7	89.5	87.4
+AugShift (ours)	72.6	85.6	92.1	94.8	70.1	90.2	88.2
Blurpool [34]	72.4	78.2	92.5	92.5	71.4	90.5	88.8

Cons: how often a network predicts the same category after the input image is vertically and horizontally shifted a random number of pixels: up to 8 pixels (Cons.8), and 8 pixels (Cons.16)

Attacking the Position-Encoding Channels



Rank

Remove Top-N Neurons
(Set to 0)

Attacking the Position-Encoding Channels

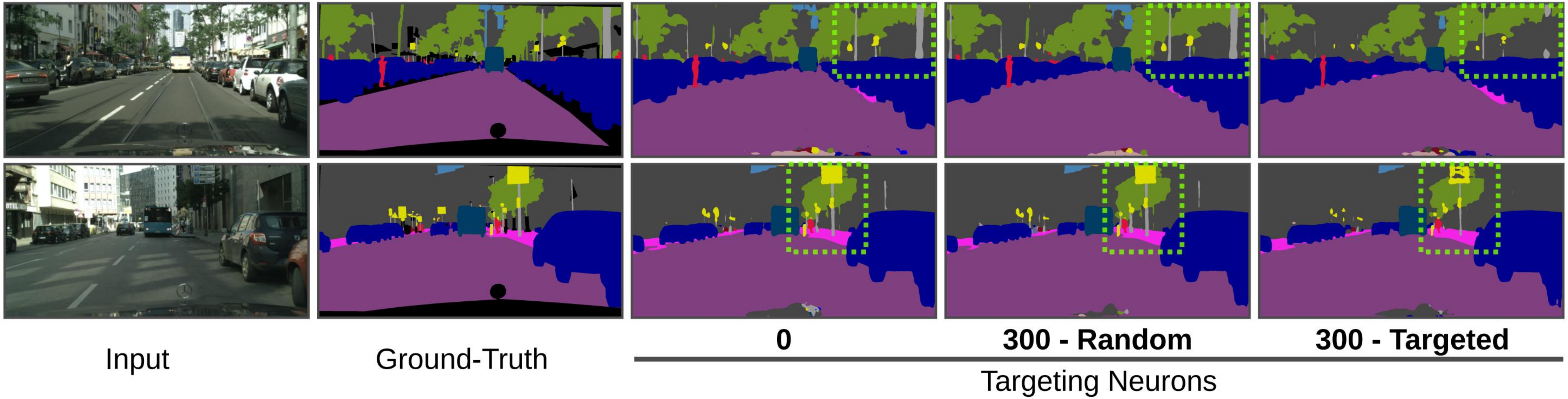
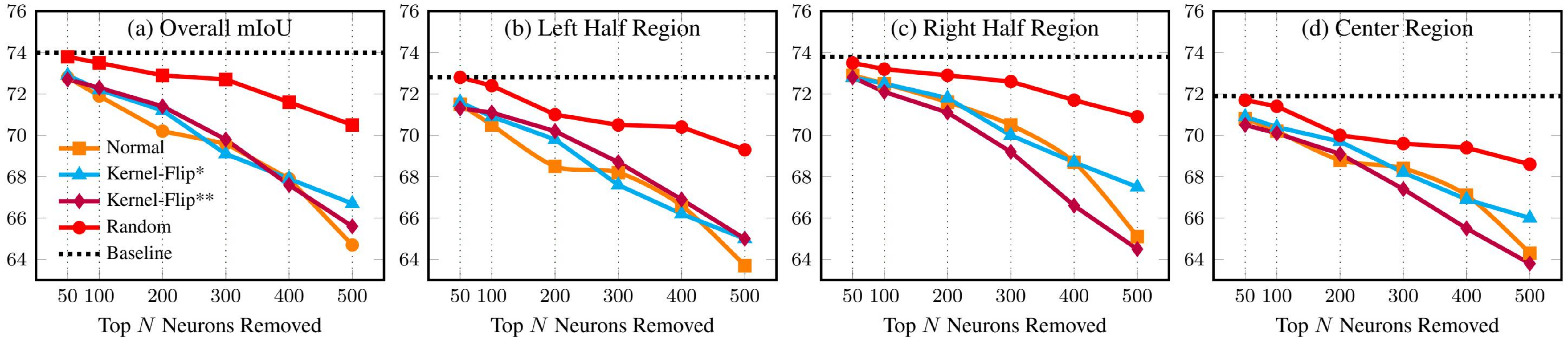


Figure 5. Qualitative comparisons between the *position-specific* and *random* neuron removal on the Cityscapes val set. Note the performance drop on objects near the periphery (highlighted in dotted box) are particularly pronounced for our position specific neuron targeting.

Attacking the Position-Encoding Channels



Thanks

Email: kang_liao@bjtu.edu.cn