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

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

## S1. Decoding Absolute Location From Pre-trained Models

We have shown in Sec. 3 that Global Average Pooling (GAP) layers can admit absolute position information by means of the *ordering* of the channel dimensions. Now we explore how much absolute position information can be decoded from various *pre-trained* models which are not explicitly trained for location classification. We first explore an ImageNet [4] pretrained ResNet-18 model [7], $f_{\text{enc}}$. As input, we use the same images as described in Sec. 3 of the main manuscript: we place a CIFAR-10 [8] image on a black canvas in a location (note there is no overlapping with other locations), where each location has a unique index (see Fig. 1 in the main manuscript for a visual example). We feed this grid-based input image, $I$, to $f_{\text{enc}}$ and obtain the latent representation, $z$. Next, we apply a $1 \times 1$ convolution on $z$ to produce a representation, $z'$, which has the same number of channel dimensions as the number of classification logits. Then we apply the GAP operation which collapses the spatial dimension, resulting in the final classification logits, $\hat{y}$. Note that we freeze the classification network as we are interested in validating how much absolute location can be decoded from the latent representation of pre-trained model for image classification. We can formalize the operations as follows:

$$z = f_{\text{enc}}(I), \quad z' = \text{Conv}_{1\times1}(z), \quad \hat{y} = \text{GAP}(z'). \tag{S1}$$

We also decode absolute position information from the latent representation of a ResNet-18 model trained for the task of *semantic segmentation* on the PASCAL VOC 2012 dataset [5]. The same method is applied as above, using a simple $1 \times 1$ convolution on the latent representation $z$, followed by a GAP layer to output the number of location classes.

*Equal Contribution

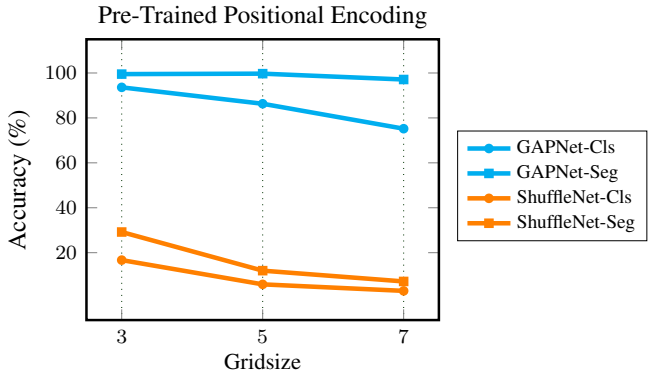


Figure S1. Decoding absolute location from ImageNet [4] pre-trained classification and PASCAL VOC 2012 [5] pretrained segmentation models (ResNet-18 [7] backbone) using *GAPNet* and *ShuffleNet*. Note that we *freeze* the classification and segmentation pretrained models and only train the GAP or linear layer which predicts the output logits. It is clear that *GAPNet* can decode positional information from a model trained for classification or semantic segmentation, while *ShuffleNet* fails to correctly decode locations. This demonstrates that positions are encoded channel-wise in the latent representation.

We provide the location classification results from image classification and semantic segmentation pretrained models in Fig. S1. These results are consistent with the results in Sec. 3 of the main manuscript and further demonstrate unequivocally that rich positional information is contained in the channels of CNNs. Furthermore, as shown by the degradation of performance when a shuffling operation is applied (*ShuffleNet*), that this information is based on the *ordering* of the channels.

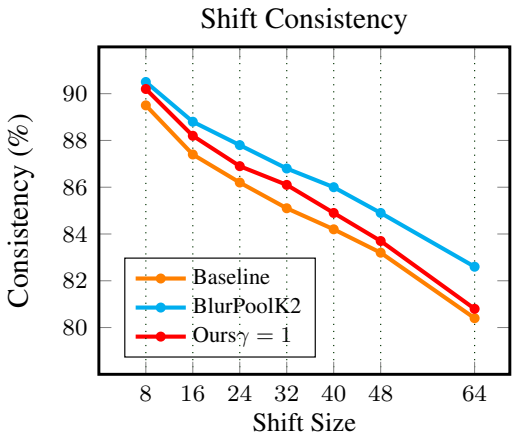


Figure S2. Comparison of shifting consistency with increasing pixel shift sizes across different methods trained on ImageNet [4] .

## S2. Shift Invariance Results

In Table 2 of the main manuscript we presented shift consistency results for various networks. We show additional shift-consistency results in Figure S2. We compare three networks, a standard ResNet-50 [7], a ResNet-50 with BlurPool-k2 [12], and our *AugShift* method. Note that we train each model on ImageNet and use the validation set to validate the consistency for pixel shifts $= \{8, 16, 32, 40, 48, 64\}$. Our method consistently outperforms the ResNet-50 baseline and reveals a useful adjunct strategy when compared with BlurPool.

## S3. Targeting Region-Specific Channels

In Sec. 4.2.2 of the main manuscript, we have shown it is possible for specific channels in the latent representation of a CNN to encode specific regions contained in an image, and furthermore, that suppressing these activations can harm the performance in *specific regions* in the image. We now show an overall comparison of the difference in performance, in terms of mean intersection over union (mIoU), between the left and right halves of the image, when either the left-encoding or right-encoding channels are suppressed. Figure S3 shows the change in mIoU when evaluated for the left and right halves of the Cityscapes [3] validation image when the *right*-encoding channels are turned off. As expected, we see a moderate but consistent decrease in performance on the *right half* of the image. Figure S4 shows the same results but when targeting the *left*-encoding channels. Similar to the right-encoding channels, we see a moderate but consistent decrease in performance on the *left half* of the image.

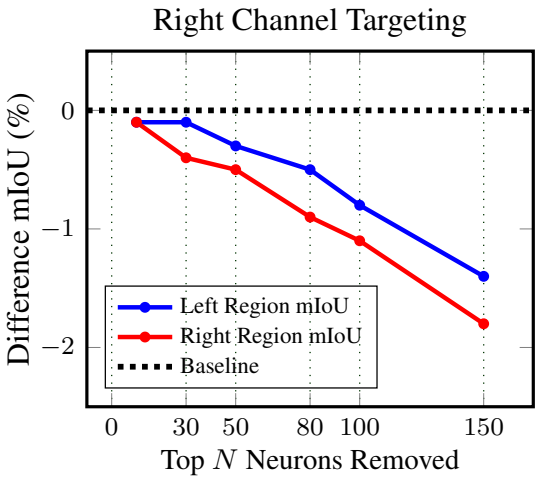


Figure S3. Relative performance drop in terms of mIoU when the top $N$ *right-specific* neurons are removed from the left and right regions. Note that we evaluate on either the *left half or right half of the image* for DeepLabv3-ResNet-50's [2] trained on Cityscapes [3] when the top $N$ *region*-specific channels are removed from the latent representation during inference.

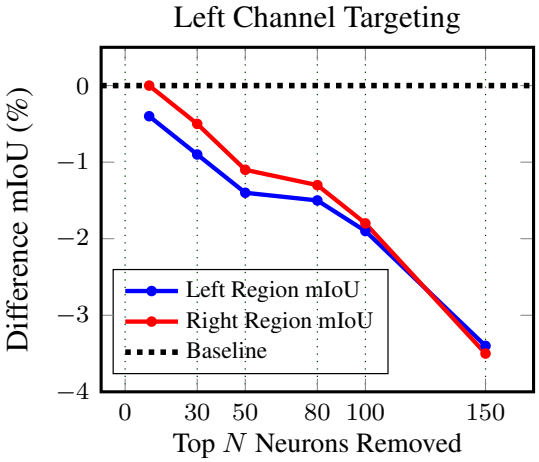


Figure S4. Relative performance drop when the top $N$ *left-specific* neurons are removed from the left and right regions. Note that we evaluate on either the *left half or right half of the image* for DeepLabv3-ResNet-50's [2] trained on Cityscapes [3] when the top $N$ *region*-specific channels are removed from the latent representation during inference.

## S4. Targeting Pedestrian Detection Networks

We are interested in whether position-specific neurons are important for object-centric position-dependant tasks. Our hypothesis is that removing position-specific neurons may harm detection performance more than removing random neurons as position is an important factor in the successful detection of objects in a scene. To this end, we now target the overall position-specific channels of a pedestrian

| Methods | Reasonable ↓ | Small ↓ | Heavy↓ | All↓ |
|---|---|---|---|---|
| Faster-RCNN [10] | 10.3 | 11.59 | 33.07 | 30.34 |
| + Random | 10.83 | 11.87 | 33.78 | 31.92 |
| + **Targeted** | **12.51** | **12.77** | **36.95** | **34.36** |
| Cascade-RCNN [1] | 7.55 | 8.55 | 27.47 | 26.89 |
| + Random | 7.85 | 8.39 | 28.31 | 27.17 |
| + **Targeted** | **8.44** | **9.3** | **30.59** | **28.57** |
| CSP [9] | 11.05 | 14.76 | 41.35 | 37.57 |
| + Random | 11.05 | 15.0 | 41.3 | 38.05 |
| + **Targeted** | **11.14** | **16.7** | **41.24** | **38.82** |

Table S1. Targeting pedestrian detection models with position-specific neurons. We remove the top 100 neurons from the latent presentation of the detection models. Targeting the position-specific channels has more influence on the overall pedestrian detection performance compared to the random targeting. Note that lower is better for the reported metrics.

detection model trained on the CityPerson [3] dataset. The CityPerson dataset is based on Cityscapes [3] but only uses the bounding box annotations of the *person* category and is used for the task of pedestrian detection. We choose the following three recent pedestrian detection models trained on CityPerson (available in [6]): (i) Faster-RCNN [10] (ii) Cascade-RCNN [1] with the HRNet [11] backbone, and (iii) CSP-ResNet-50 [9]. Similar to the experiment in Sec. 4.2.2, we identify the top $N$ overall position-encoding channels (using Eq. 2) and remove these dimensions before passing the latent representation to the detection head.

Table S1 presents the pedestrian detection results when the top 100 position-specific neurons are removed from a pedestrian detection model trained on CityPerson (we choose $N = 100$ as the latent dimension of the networks used are relatively small (e.g., 256 for HRNet [11])). Note that we follow the standard benchmark metric, mean average-precision (mAP), to report the detection results under four different settings. The results are consistent with the semantic segmentation results (Sec. 4.2.2): removing the top 100 position-encoding channels degrades the performance more than choosing 100 random neurons. For example, for the Faster-RCNN network, targeting the position encoding neurons decreases the performance by 4.02%, while targeting random neurons admits a 1.58% drop.

## References


[1] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018. 3

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017. 2

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html, 2010. 1

[6] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Pedestrian detection: The elephant in the room. *arXiv preprint arXiv:2003.08799*, 2020. 3

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2

[8] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The CIFAR-10 dataset. *online: http://www. cs. toronto. edu/kriz/-cifar. html*, 2014. 1

[9] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *CVPR*, 2019. 3

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, 2015. 3

[11] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, and Xinggang Wang. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2020. 3

[12] Richard Zhang. Making convolutional networks shift-invariant again. In *ICML*, 2019. 2