Neurocomputing 506 (2022) 158–167

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Polarized self-attention: Towards high-quality pixel-wise mapping

Huajun Liu [a,1], Fuqiang Liu [b], Xinyi Fan [a], Dong Huang [b,*]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[b] Robotics Institute, Carnegie Mellon University, United States

ARTICLE INFO

Article history:
Received 12 February 2022
Revised 1 July 2022
Accepted 12 July 2022
Available online 16 July 2022
Communicated by Zidong Wang

Keywords:
Pixel-wise mapping
Self-attention
Polarization
Convolution

ABSTRACT

We address the pixel-wise mapping problem that commonly exists in the fine-grained computer vision tasks, such as estimating keypoint heatmaps and segmentation masks. These tasks require, at low computation overheads, modeling the long-range dependencies among high-resolution inputs and estimating the highly nonlinear pixel-wise outputs. While the attention mechanism added to Deep Convolutional Neural Networks (DCNNs) can boost long-range dependencies, the element-specific attention, such as the Nonlocal block, is highly complex and noise-sensitive to learn, and most of the simplified attention blocks are designed for image-wise classification purposes and simply applied to pixel-wise tasks. In this paper, we present the **Polarized Self-Attention (PSA)** block targeting the high-quality pixel-wise mapping with: (1) Polarized filtering: keeping high internal resolution in both channel and spatial attention computation while completely collapsing input tensors along their counterpart dimensions. (2) Enhancement: composing non-linearity that directly fits the output distribution of typical pixel-wise mappings, such as the 2D Gaussian distribution (keypoint heatmaps), or the 2D Binormial distribution (binary segmentation masks). Experimental results show that PSA boosts standard baselines by 2–4 points, and boosts state-of-the-arts by 1–2 points on 2D pose estimation and semantic segmentation benchmarks. Codes are available at (https://github.com/DeLightCMU/PSA).

1. Introduction

Recent trends from the coarse-grained (such as image-wise classification [1] and bounding box detection [2]) to the fine-grained computer vision tasks (such as keypoint estimation [3] and semantic segmentation [4]) have received booming advances in both research and industrial communities. Compared to the coarse-grained tasks, perception at the pixel-wise level is increasingly appealing in autonomous driving [5], augmented reality [6], medical image processing [7], and public surveillance [8].

The goal of the pixel-wise mapping problem is to map every image pixelof the same semantics to the same scores. For instance, mapping all the background pixels to 0 and all the foreground pixels to their class indices, respectively. Two typical tasks are keypoint heatmap [9,3] and segmentation mask estimation [10,4], which estimate the pixel-wise closeness to 2D keypoints and the pixel-wise semantic labels respectively. Most DCNN models for these problems take an encoder-decoder architecture. The encoder usually consists of a backbone network, such as ResNet [11], that sequentially reduces the spatial resolution and increases the channel resolution, while the decoder usually contains de-convolution/up-sampling operations that recover the spatial resolution and decrease the channel resolution. Typically the bottleneck tensor connecting the encoder and decoder is smaller than both the input image tensor and the output tensor. The reduction of elements is necessary for computation/memory efficiency and stochastic optimization reasons [12]. However, the pixel appearances and patch shapes of the same semantics are highly nonlinear in nature and therefore difficult to be encoded with a reduced number of features. Moreover, high input–output resolutions are preferred for fine details of objects and object parts [13–15]. Compared to the image classification task where an input image is collapsed to one output vector of class indices, the pixel-wise mapping problem produces many vectors of an output tensor. From the model design perspective, the pixel-wise mapping problem faces special challenges: (1) Keeping high internal resolution at a reasonable cost; (2) Fitting output distribution such as that of the keypoint heatmaps or segmentation masks.

Based on the tremendous success in new DCNNs architectures, we focus on a plug-and-play solution that could consistently improve an existing (vanilla) network, i.e., inserting attention

* Corresponding author.
E-mail address: dghuang@andrew.cmu.edu (D. Huang).

[1] This work was partially done when Huajun Liu is a Postdoc at Carnegie Mellon University.

https://doi.org/10.1016/j.neucom.2022.07.054
0925-2312/

blocks [16–23]. Most of the above hybrids try to reach the best compromise among multiple types of tasks, for instance, image classification, object detection, as well as for instance segmentation. These generalized goals are partially the reason that channel-only attention (SE [24], GE [22] and GCNet [23]) are among the most popular blocks. Unfortunately, due to the lack of critical differences in attention designs, the channel-spatial compositional attention blocks, (e.g., DA [25], CBAM [26]), did not show significant overall advantages over the latest channel-only attentions such as GCNet [23].

In this paper, we present the Polarized Self-Attention (PSA) block (See Fig. 1) for high-quality pixel-wise mapping. PSA demonstrates the superiority of high-resolution internal feature maps and nonlinear composition of attention in the pixel-wise mapping tasks. The main contributions of this work can be summarized as follows:

1. To preserve high-resolution information needed in pixel-wise mapping tasks, we stick with high internal resolution in all attention computation.
2. To fit the typical output distribution of pixel-wise mapping tasks, we incorporate the softmax-sigmoid composition in all attention branches;
3. To demonstrate consistent performance gains of PSA over baselines and the state-of-the-arts (SOTA), we conducted extensive experiments of popular DNN structures and configurations (e.g. input size)

2. Related work

Pixel-wise Mapping Tasks: The advances of DCNNs for pixel-wise mapping are basically pursuing higher resolution. For body keypoint estimation, Simple-Baseline[27] consists of conventional components ResNet + deconvolution. HRnet[14] addresses the resolution challenge of Simple-Baseline with 4 parallel high-to-low resolution branches and their pyramid fusion. Other most recent variants, DARK-Pose[28] and UDP-Pose[29], both compensate for the loss of resolution due to the preprocessing, post-processing, and propose techniques to achieve a sub-pixel estimation of keypoints. Note that, besides the performance gain among network designs, the same models with and 388×284 inputs are usually better than that with 256×192 inputs. This constantly reminds researchers of the importance of keeping high-resolution information. For Semantic segmentation, [30] introduces atrous convolution in the decoder head of Deeplab for wide receptive field on high-resolution inputs. To overcome the limitation of ResNet backbones in Deeplab, all the latest advances are based on HRnet [15], in particular, HRNet-OCR[31] and its variants are the current state-of-the-art. There are many other multitask architecture [32–34] that include pixel-wise mapping as a component.

PSA further pursues the high-resolution goals of the above efforts from the attention perspective and further boosts the above DCNNs.

Self-attention and its Variants. Attention mechanisms have been introduced to many visual tasks to address the limitations of standard convolutions on long-range interaction [17,35–37]. In the self-attention mechanism, an input tensor is used to compute an attention tensor and is then re-weighted by this attention tensor. Self-attention [16–18] became popular component to capture long-range interactions, after it success in sequence modeling and generative modeling tasks. Cordonnier et al. [18] has proven that a multi-head self-attention layer with a sufficient number of heads is at least as expressive as any convolutional layer. In some vision tasks, such as object detection and image classification, self-attention augmented convolution models [35] or standalone self-attention models [37] have yielded remarkable gains. While most self-attention blocks were inserted after convolution blocks, attention-augmented convolution [35] demonstrates that parallelizing the convolution layer and attention block is a more powerful structure to handle both short and long-range dependency.

PSA advances self-attention for pixel-wise mapping and could also be used in other variants such as the convolution-augmented attentions.

Full-tensor and simplified attention blocks. The basic non-local block (NL) [19] and its variants, such as a residual form [38] second-order non local [20,21], and asymmetric non-local [39], produce full-tensor attentions and have successfully improved person re-identification, image super-resolution, and semantic segmentation tasks. To capture pair-wise similarities among all feature elements, the NL block computes an extremely large similarity matrix between the key feature maps and query feature maps, leading to huge memory and computational costs. EA [40] produces a low-rank approximation of NL block for computation efficiency. BAM [41], DAN [25] and CBAM [26] produce different compositions of the channel-only and spatial-only attentions. Squeeze-and-Excitation (SENet) [24], Gather-Excite [22] and GCNet [23] only re-weight feature channels using signals aggregated from global context modeling. Most of above attention blocks were designed as a compromise among multiple types of tasks, and do not address the specific challenges in pixel-wise mapping.

PSA address the specific challenges in pixel-wise mapping by keeping the highest attention resolution, and directly fitting the typical output distributions.

3. Our method

Notations:[2] Denote $\mathbf{X} \in \Re^{C_{in} \times H \times W}$ as a feature tensor of one sample (e.g., one image), where C_{in}, H, W are the number of elements along the channel, height, and width dimension of $\mathbf{X}$, respectively. $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{HW}$ where $\mathbf{x}_i \in \Re^{C_{in}}$ is a feature vector along the channel dimension. A self-attention block $\mathbf{A}(\cdot)$ takes $\mathbf{X}$ as input, and produces a tensor $\mathbf{Z}$ as output, where $\mathbf{Z} \in \Re^{C_{out} \times H \times W}$. A DCNN block is formulated as a nonlinear mapping $\Psi : \mathbf{X} \rightarrow \mathbf{Z}$. For conciseness, all the (1×1) convolution layers in attention blocks are denoted by $\mathbf{W}$. Without lose of generality, we only consider the case where the input tensor $\mathbf{X}$ and output tensor $\mathbf{Z}$ of a DCNN block have the same dimension $C \times H \times W$ (i.e., $C_{in} = C_{out}$).

3.1. Self-attention for pixel-wise mapping

A DCNN for pixel-wise mapping computes weighted combinations on input features along two dimensions: (1) channel-wise combinations for class score outputs; (2) spatial-wise combinations (e.g., convolution) for identifying pixels of the same semantics. The self-attention mechanism applied to the DCNN is expected to further highlight the critical input feature elements for both above goals.

Ideally, with a full-tensor self-attention $\mathbf{Z} = \mathbf{A}(\mathbf{X}) \odot \mathbf{X}$ (with $\mathbf{A}(\mathbf{X}) \in \Re^{C \times H \times W}$), the highlighting could potentially be achieved at the element-wise granularity ($C \times H \times W$ elements). However, the attention tensor $\mathbf{A}$ is very complex and noise-prone to learn directly. In the Non-Local self-attention block [19], $\mathbf{A}$ is calculated as,

$$\mathbf{A} = \mathbf{W}_z(F_{sm}(\mathbf{X}^T\mathbf{W}_k^T\mathbf{W}_q\mathbf{X})\mathbf{W}_v\mathbf{X}). \quad (1)$$

[2] All non-bold letters represent scalars. Bold capital letter $\mathbf{X}$ denotes a matrix; Bold lower-case letters $\mathbf{x}$ is a column vector. $\mathbf{x}_i$ represents the i^{th} column vector of the matrix $\mathbf{X}$. x_j denotes the j^{th} element of $\mathbf{x}$. $\langle \mathbf{x}, \mathbf{y} \cdot \rangle = \mathbf{x}^T\mathbf{y}$ denotes the inner-product between two vectors or metrics.

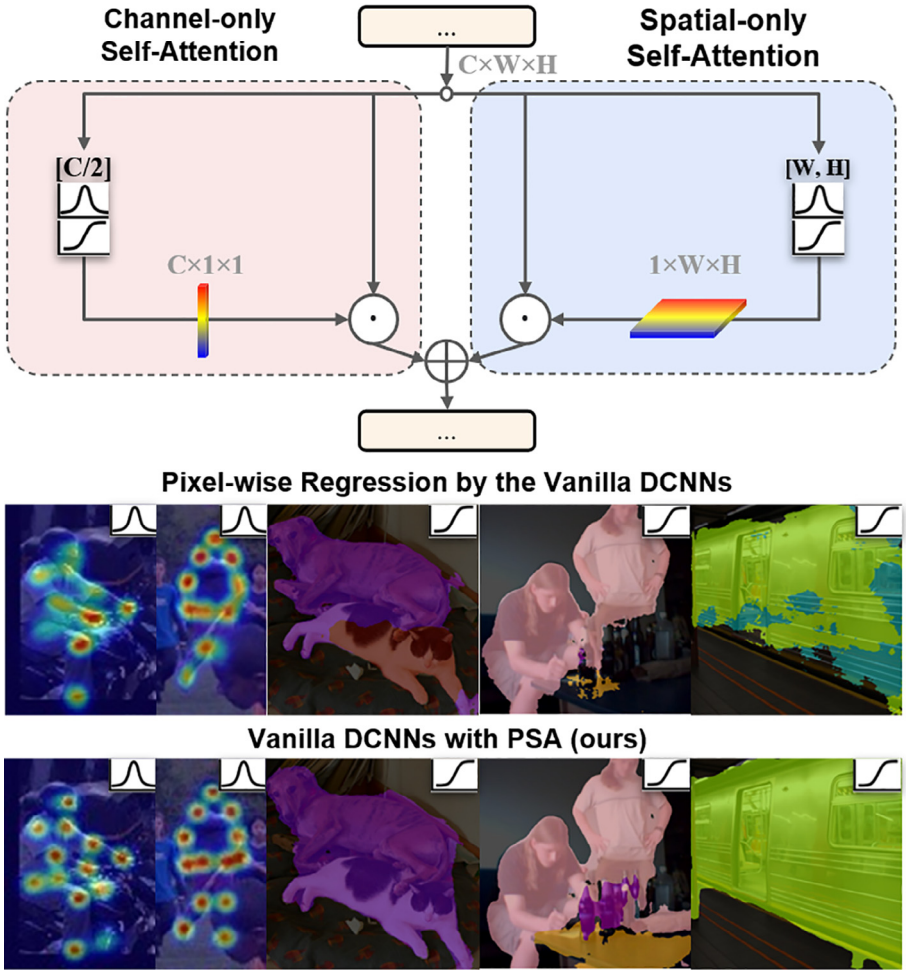


Fig. 1. Polarized Self-Attention(PSA) block (ours), keeps high internal resolution along the channel ($C/2$) and spatial dimension($[W, H]$) while collapses the input tensor $[C \times W \times H]$ along their counterparts dimensions, and fits output distributions of pixel-wise mapping with a softmax-sigmoid composition. At minor computation&memory overheads upon the vanilla DNNs, PSA produces significantly higher-quality person keypoint heatmaps and semantic segmentation masks (also see Table 2–3 for the ga.ins in metrics).

There are four (1×1) convolution kernels, i.e., $\mathbf{W}_z, \mathbf{W}_k, \mathbf{W}_q$, and $\mathbf{W}_v$, that learns the linear combination of spatial features among different channels. Within the same channels, the $HW \times HW$ outer-product between $\mathbf{W}_k\mathbf{X}$ and $\mathbf{W}_q\mathbf{X}$ activates any features at different spatial locations that have a similar intensity. The joint activation mechanism of spatial features is very likely to highlight the spatial noise. The only actual weights, $\mathbf{W}$s, are channel-specific instead of spatial-specific, making the Non-Local attention exceptionally redundant at the huge memory-consumption of the $HW \times HW$ matrix. For efficient computation, reduction of NL leads to many possibilities: Low rank approximation of $\mathbf{A}$ (EA), Channel-only self-attention $\mathbf{A}^{ch} \in \Re^{C\times 1\times 1}$ that highlight the same global context for all pixels(GC [23] and SE [22]), Spatial-only self-attention $\mathbf{A}^{sp} \in \Re^{1\times W\times H}$ not powerful enough to be recognized as a standalone model, Channel-spatial composition $\mathbf{A}^{cs}$, where the parallel composition: $\mathbf{Z} = \mathbf{A}^{ch} \odot^{ch} \mathbf{X} + \mathbf{A}^{sp} \odot^{sp} \mathbf{X}$and the sequential composition: $\mathbf{Z} = \mathbf{A}^{ch} \odot^{ch} (\mathbf{A}^{sp} \odot^{sp} \mathbf{X})$ introduce different order of non-linearity. Here $\odot^{ch}$ is a channel-wise multiplication operator that repeatedly multiplies $\mathbf{A}^{ch}$ on every channel of $\mathbf{X}$. Here $\odot^{sp}$ is a spatial-wise multiplication operator that repeatedly multiplies $\mathbf{A}^{sp}$ on every spatial element of $\mathbf{X}$.

Existing work reach different conclusions empirically, e.g., in CBAM [26], the sequential composition is better than the parallel one, and in DA [25], the parallel composition is better than the sequential one. This partially indicates that the intended non-linearity of the tasks is not fully modeled within the attention blocks. These issues are typical examples of general attention design that does not target the pixel-wise mapping problem. With the help of Table 1, we re-visit critical design specifications of existing attention blocks and discuss two crucial design aspects of attention blocks for pixel-wise mapping.

(1) Internal Attention Resolution. Recall that the most commonly used backbone networks only produce down-sampled features for robustness and computational efficiency. For instance, a ResNet-based network produces a $1 \times 1 \times 512$ feature tensor for image classification or a $[W/r, H/r]$ feature tensor for bbox regression, where r is the height/width the smallest object bounding box. The pixel-wise mapping problem toward high-resolution outputs cannot afford such loss of internal resolution. This is because the highly non-linearity in object edges and body parts are very difficult to encode in low-resolution features [30,15,14]. However, in Table 1, all attention blocks after NL have their performance saturated at low internal channel or spatial resolutions. The natural question is: *what prevents attention blocks from leveraging higher resolution information?* Our answer is the output non-linearity.

(2) Output Non-linearity. In DCNNs for pixel-wise mapping, outputs are usually encoded as 3D tensors. For instance, the 2D keypoint coordinates are encoded as a stack of 2D Gaussian maps $[\#keypoint\_type \times W \times H]$. The pixel-wise class indices are encoded

Table 1
Attention blocks with the same input–output tensor sizes $[C, W, H]$ ($C < WH$). All parameters/metrics of existing blocks are compared in their top-performance configurations. "SM" denotes SoftMax and "SD" denotes Sigmoid for space saving.

Method	Internal channel resolution	Internal spatial resolution	Non-linear composition	Complexity $O(\cdot)$
NL[19]	C	$[W, H]$	SM	$C^2WH + CW^2H^2$
GC [23]	$C/4$	$[1, 1]$	SM + ReLU	CWH
SE [22]	$C/4$	$[1, 1]$	ReLU + SD	CWH
CBAM [26]	$C/16$	$[W, H]$	SD	CWH
DA [25]	$C/8$	$[W, H]$	SM	$C^2WH + CW^2H^2$
EA [40]	d_k ($\ll C$)	d_v ($\ll min(W, H)$)	SM	CWH
PSA(ours)	$C/2$	$[W, H]$	SM + SD	CWH

as a stack of binary maps $[\#semantic\_classes \times W \times H]$. We consider the 2D dimensional piece-wise step function a cumulative distribution function of binomial variables. From the perspective of DCNN training, the non-linearity that directly fits the distribution upon linear transformations (such as convolution) could potentially alleviate the learning burden of DCNNs. The natural nonlinear functions to fit the above distributions are SoftMax for 2D Gaussian maps, and Sigmoid for 2D Binormial Distribution. No existing attention block Table 1 contains such a combination of nonlinear functions.

3.2. Polarized self-attention (PSA) block

Why polarized self-attention? The analogy comes from "polarized filtering" in photography. Its goal is to remove the noisy lights in the transverse directions that produce glares/reflections, while preserving the light of the scene for better photo quality. The polarized filtering technique cleans the incoming light respectively in two orthogonal oscillation directions, then combines their outcoming lights into the final photo. In this process, the high-order noise that cannot be reconstructed as linear combinations of the two orthogonal directions is removed. Moreover, filtering may reduce the total light intensity, resulting in a small dynamic range. An additional boost, e.g. High Dynamic Range (HDR), is applied to recover the photo contrast.

In image-wise 2D convolution[3], we consider the features of an input tensor $\mathbf{X}$ along the spatial and channel dimensions are orthogonal. Observe that, multiplication is only implemented either between the spatial elements or between the channel elements, while between the spatial and channel elements, only additions are conducted. This means DNNs are not likely to learn useful high-order information across the spatial and channel elements. **Such information are very likely to be high-order noise.**

We borrow the key factors of "polarized filtering" in photography, and propose the Polarized Self-Attention (PSA) mechanism: (1) Filtering: completely collapsing features in one direction while preserving high resolution in its orthogonal direction; (2) HDR: increasing the dynamic range of attention by Softmax normalization at the bottleneck tensor (smallest feature tensor in attention block), followed by tone-mapping with the Sigmoid function. Formally, we instantiate the PSA mechanism as a PSA block below (also see diagram in Fig. 2):

Channel-only branch $A^{ch}(\mathbf{X}) \in \Re^{C\times 1\times 1}$**:**

$$\mathbf{A}^{ch}(\mathbf{X}) = F_{SG}\left[\mathbf{W}_{z|\theta_1}\left((\sigma_1(\mathbf{W}_v(\mathbf{X})) \times F_{SM}(\sigma_2(\mathbf{W}_q(\mathbf{X}))))\right)\right], \tag{2}$$

where $\mathbf{W}_q, \mathbf{W}_v$ and $\mathbf{W}_z$ are 1×1 convolution layers respectively, σ_1 and σ_2 are two tensor reshape operators, and $F_{SM}(\cdot)$ is a SoftMax operator and "$\times$" is the matrix dot-product operation $F_{SM}(\mathbf{X}) = \sum_{j=1}^{N_p} \frac{e^{x_j}}{\sum_{m=1}^{N_p} e^{x_m}} x_j$. The internal number of channels, between $\mathbf{W}_v|\mathbf{W}_q$ and $\mathbf{W}_z$, is $C/2$. The output of channel-only branch is $\mathbf{Z}^{ch} = \mathbf{A}^{ch}(\mathbf{X}) \odot^{ch} \mathbf{X} \in \Re^{C\times H\times W}$.

Spatial-only branch $A^{sp}(\mathbf{X}) \in \Re^{1\times H\times W}$**:**

$$\mathbf{A}^{sp}(\mathbf{X}) = F_{SG}\left[\sigma_3\left(F_{SM}(\sigma_1(F_{GP}(\mathbf{W}_q(\mathbf{X})))) \times \sigma_2(\mathbf{W}_v(\mathbf{X}))\right)\right], \tag{3}$$

where $\mathbf{W}_q$ and $\mathbf{W}_v$ are standard 1×1 convolution layers respectively, θ_2 is an intermediate parameter for these channel convolutions, and σ_1, σ_2 and σ_3 are three tensor reshape operators, and $F_{SM}(\cdot)$ is the SoftMax operator. $F_{GP}(\cdot)$ is a global pooling operator $F_{GP}(\mathbf{X}) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X(:, i, j)$, and $\times$ is the matrix dot-product operation. The output of spatial-only branch is $\mathbf{Z}^{sp} = \mathbf{A}^{sp}(\mathbf{X}) \odot^{sp} \mathbf{X} \in \Re^{C\times H\times W}$.

Composition: The outputs of above two branches are composed either under the parallel layout

$$\begin{aligned} PSA_p(\mathbf{X}) &= \mathbf{Z}^{ch} + \mathbf{Z}^{sp} \\ &= \mathbf{A}^{ch}(\mathbf{X}) \odot^{ch} \mathbf{X} + \mathbf{A}^{sp}(\mathbf{X}) \odot^{sp} \mathbf{X}, \end{aligned} \tag{4}$$

or under the sequential layout

$$\begin{aligned} PSA_s(\mathbf{X}) &= \mathbf{Z}^{sp}(\mathbf{Z}^{ch}) \\ &= \mathbf{A}^{sp}(\mathbf{A}^{ch}(\mathbf{X}) \odot^{ch} \mathbf{X}) \odot^{sp} \mathbf{A}^{ch}(\mathbf{X}) \odot^{ch} \mathbf{X}. \end{aligned} \tag{5}$$

where "+" is the element-wise addition operator.

Remarks: In essence, $PSA_p(\cdot)$ results in a wider model and $PSA_s(\cdot)$ results in a deeper one. We also append PSA to Table 1 and make the following observations:

- *Internal Resolution vs Complexity:* Comparing to existing attention blocks under their top configuration, PSA preserves the highest attention resolution for both the channel ($C/2$)[4] and spatial ($[W, H]$) dimension.
  Moreover, in our channel-only attention, the Softmax-based re-weighting is fused with the Squeeze-Excitation block, which introduces strong nonlinear activation at the bottleneck tensor of size $C/2 \times W \times H$. Note that the same squeeze-excitation pattern of channel numbers (C-$\frac{C}{2}$-C) has benefited GC and SE. Our channel-only block conducts **higher-resolution squeeze-and-excitation** than the GC block while costs comparable computation of the GC block.
  Our spatial-only attention not only keeps the full $[W, H]$ spatial resolution, but also internally keeps $2 \times C \times C/2$ learnable parameters in W_q and W_v for the nonlinear Softmax re-weighting, which is more powerful structure than existing blocks. For instance, the spatial-only attention in CBAM is parameterized by a $7 \times 7 \times 2$ convolution (a linear operator), and EA learns

[3] This is just to avoid confusion with the 3D convolution used for spatial–temporal modeling.

[4] $C/2$ is the smallest channel number when PSA produces the best metrics, and is used throughout our experiments.

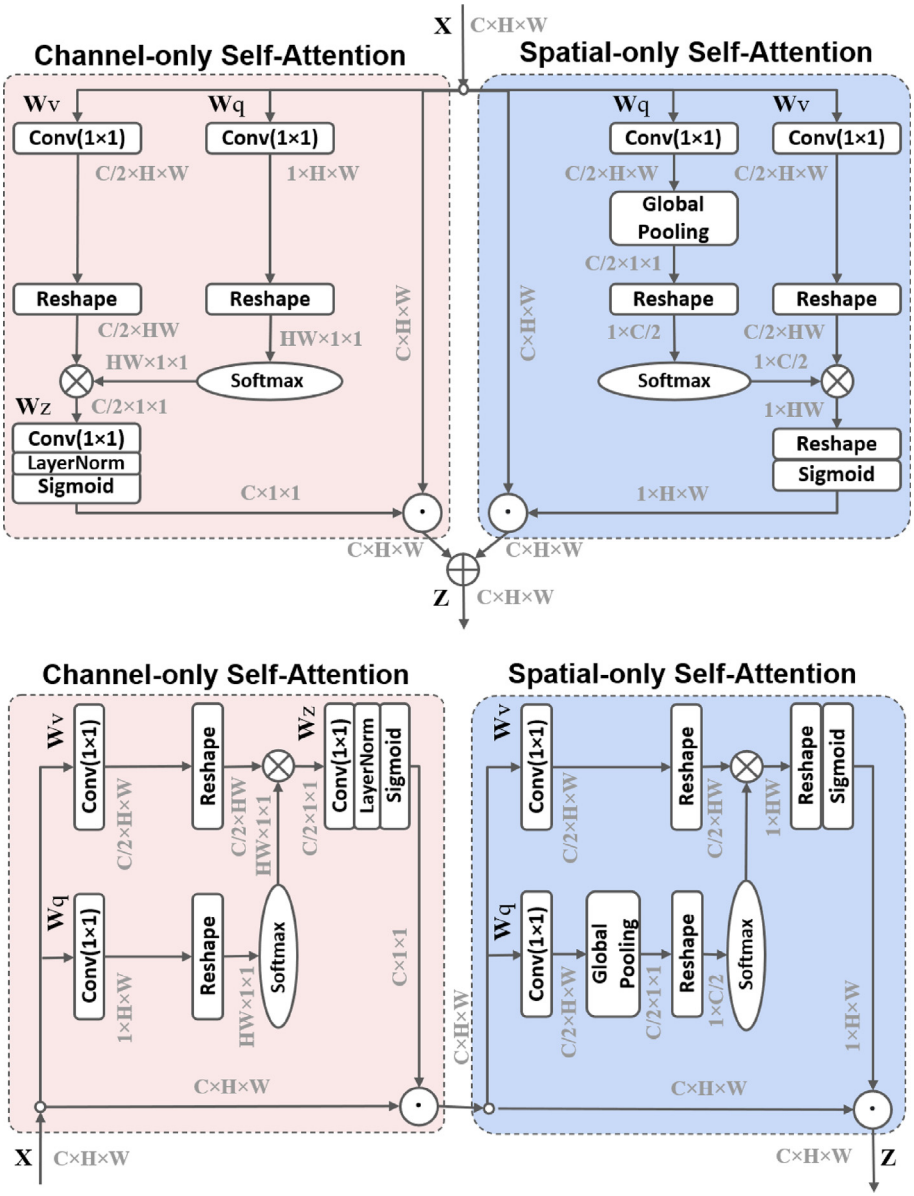


Fig. 2. The Polarized Self-Attention (PSA) block under **(upper)** the parallel layout, and **(lower)** the sequential layout.

$C \times d_k + C \times d_v$ parameters for linear re-weighting ($d_k, d_v \ll C$).

- *Output-specified Non-linearity*. Both the PSA channel-only and spatial-only branches use a Softmax-Sigmoid composition. Considering the Softmax-Sigmoid composition as a probability distribution function, both the multi-mode Gaussian maps (keypoint heatmaps) and the piece-wise Binomial maps (segmentation masks) can be approximated upon linear combination, i.e. 1×1 convolutions in PSA. We therefore expect such non-linearity make DNNS easier to leverage the high resolution information preserved in PSA branches.

4. Experiments

Implementation details. For any baseline networks with the bottleneck or basic residual blocks, such as ResNet and HRnet, we add PSAs after the first 3×3 convolution in every residual blocks, respectively. For 2D pose estimation, we kept the same training strategy and hyper-parameters as the baseline networks. For semantic segmentation, we added a warming-up training phase of 5000 iterations, stretched the total training iteration by 30%, and kept all the rest training strategy and hyper-parameters of the baseline networks. Empirically, these changes allow PSA to train smoothly on semantic segmentation.

4.1. PSA vs. baselines

Top-Down 2D Human Pose Estimation: Among the DCNN approaches for 2D human pose estimation, the top-down approaches generally dominate the top metrics. This top-down pipeline consists of a person bounding box detector and a keypoint heatmap regressor. Specifically, we use the pipelines in [27,14] as our baselines. An input image is first processed by a human detector [27] of 56.4AP (Average Precision) on MS-COCO val2017 dataset [42]. Then all the detected human image patches are cropped from the input image and resized to 384×288. Finally, the 384×288 image patches are used for keypoint heatmap regression by a single person pose estimator. The output heatmap size is 96×72.

We add PSA on Simple-Baseline [27] with the Resnet50/152 backbones and HRnet [14] with the HRnet-w32/w48 backbones.

Table 2
PSA vs. Baselines for top-down human pose estimation on the MS-COCO val2017 dataset. All results were computed with an human detector [27] of 56.4 AP on COCO val2017 dataset. All detected human image patches were resized to 384 × 288.

Method	Backbone	Pretrained	AP ↑	AP_{50}↑	AP_{75}↑	AP_M ↑	AP_L↑	AR ↑	Flops	mPara
Simple-Baseline [27]	Res50	Y	72.2	89.3	78.9	68.1	79.7	77.6	20.0G	34.0 M
+ PSA	Res50	N	**76.5(+4.3)**	**93.6**	**83.6**	**73.2**	**81.0**	**79.0**	20.9G	36.1 M
Simple-Baseline [27]	Res152	Y	74.3	89.6	81.1	70.5	81.6	79.7	35.3G	68.6 M
+ PSA	Res152	N	**78.0(+3.7)**	**93.6**	**84.8**	**75.2**	**82.3**	**80.5**	37.5G	75.2 M
HRNet [14]	HRNet-W32	Y	75.8	90.6	82.5	72.0	82.7	80.9	16.0G	28.5 M
+ PSA	HRNet-W32	Y	**78.7(+2.9)**	**93.6**	**85.9**	**75.6**	**83.5**	**81.1**	17.1G	31.4 M
HRNet [14]	HRNet-W48	Y	76.3	90.8	82.9	72.3	83.4	81.2	32.9G	63.6 M
+ PSA	HRNet-W48	Y	**78.9(+2.6)**	**93.6**	**85.7**	**75.8**	**83.8**	**81.4**	35.2G	70.0 M

Table 3
PSA vs. Baselines for semantic segmentation on the Pascal VOC2012 Aug database.

Method	Backbone	mIoU ↑	Flops	mPara
DeepLabV3Plus [30]	MobileNet	71.1	16.9G	5.22 M
+ PSA	MobileNet	**73.7(+2.6)**	17.1G	5.22 M
DeepLabV3Plus [30]	Res50	77.2	62.5G	39.8 M
+ PSA	Res50	**79.0(+1.8)**	65.2G	42.3 M
DeepLabV3Plus [30]	Res101	78.3	83.2G	58.8 M
+ PSA	Res101	**80.3(+2.0)**	87.7G	63.5 M

Table 4
Comparison with State-of-the-Art top-down 2D pose estimation approaches on the MS-COCO keypoint testdev set. Note that only [29]*Strong Baseline* used extra training data.

Method	Backbone	Input Size	AP	AP_{50}	AP_{75}	AP_M	AP_L	AR	Flops	mPara
8-stage Hourglass [44]	8-stage Hourglass	256 × 192	66.9	-	-	-	-	-	14.3G	25.1 M
CPN [45]	ResNet50	256 × 192	68.6	-	-	-	-	-	6.2G	27.0 M
CPN + OHKM [45]	ResNet50	256 × 192	69.4	-	-	-	-	-	6.2G	27.0 M
SimpleBaseline [27]	ResNet50	256 × 192	70.4	88.6	78.3	67.1	77.2	76.3	8.90G	34.0 M
SimpleBaseline [27]	ResNet101	256 × 192	71.4	89.3	79.3	68.1	78.1	77.1	12.4G	53.0 M
SimpleBaseline [27]	ResNet152	256 × 192	72.0	89.3	79.8	68.7	78.9	77.8	15.7G	72.0 M
HRNet-W32 [14]	HRNet	256 × 192	74.4	90.5	81.9	70.8	81.0	78.9	7.10G	28.9 M
HRNet-W48 [14]	HRNet	256 × 192	75.1	90.6	82.2	71.5	81.8	80.4	14.6G	63.6 M
Dark-Pose [28]	HRNet-W32	256 × 192	75.6	90.5	82.1	71.8	82.8	80.8	7.1G	28.5 M
UDP-Pose [29]	HRNet-W48	256 × 192	77.2	91.8	83.7	73.8	83.7	82.0	14.7G	63.8 M
SimpleBaseline [27]	ResNet152	384 × 288	74.3	89.6	81.1	70.5	79.7	79.7	35.6G	68.6 M
HRNet-W32 [14]	HRNet	384 × 288	75.8	90.6	82.7	71.9	82.8	81.0	16.0G	28.5 M
HRNet-W48 [14]	HRNet	384 × 288	76.3	90.8	82.9	72.3	83.4	81.2	32.9G	63.6 M
Dark-Pose [28]	HRNet-W48	384 × 288	76.2	92.5	83.6	72.5	82.4	81.1	33.0G	63.8 M
UDP-Pose [29]	HRNet-W48	384 × 288	76.8	90.6	83.2	72.8	84.0	81.7	32.9G	63.6 M
UDP-Pose [29]	HRNet-W48	384 × 288	77.8	92.0	84.3	74.2	84.5	82.5	33.0G	63.8 M
Ours										
Dark-Pose-PSA(p)	HRNet-W48	384 × 288	78.2	92.5	83.9	75.4	83.9	81.7	35.4G	69.9 M
Dark-Pose-PSA(s)	HRNet-W48	384 × 288	78.2	92.9	84.0	74.9	84.1	81.8	35.4G	68.8 M
UDP-Pose-PSA(p)	HRNet-W48	256 × 192	78.9	**93.6**	**85.8**	**76.1**	83.6	81.4	15.7G	70.1 M
UDP-Pose-PSA(p)	HRNet-W48	384 × 288	**79.6**	**93.6**	**85.9**	**76.4**	**84.7**	81.9	35.4G	70.1 M
UDP-Pose-PSA(s)	HRNet-W48	384 × 288	**79.7**	**94.5**	**85.8**	**76.3**	**84.3**	82.0	35.4G	69.1 M

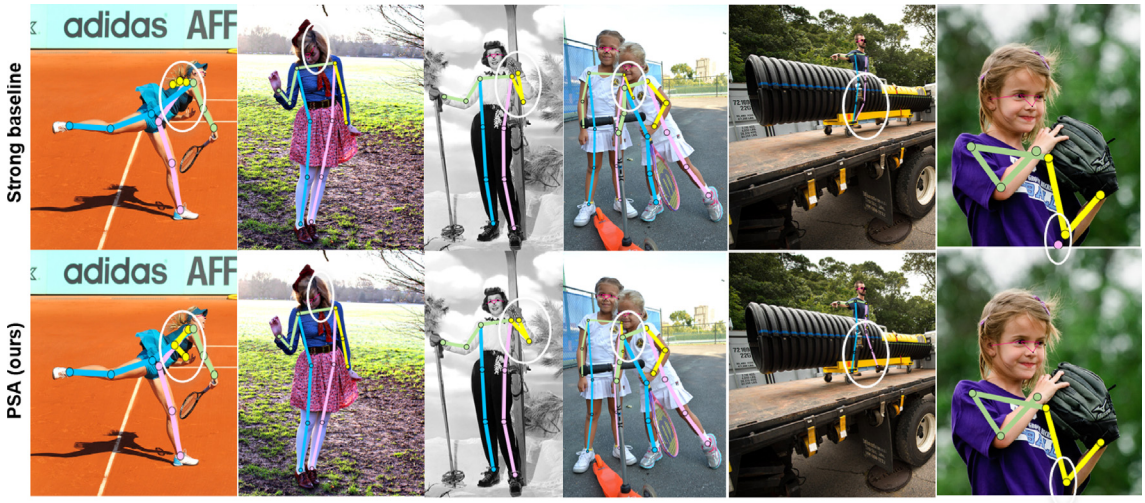
The results on MS-COCO val2017 are shown in Table 2. PSA boosts all the baseline networks by 2.6 to 4.3 AP with minor overheads of computation (Flops) and the number of parameters (mPara). Even without ImageNet pre-training, PSA with "Res50" backbone gets 76.5 AP, which is not only 4.3 better than Simple-Baseline with Resnet50 backbone, but also better than Simple-Baseline even with Resnet152 backbone. A similar benefit is also observed on PSA with HRNet-W32 backbone outperforms the baseline with "HR-w48" backbone. This giant performance gains of PAS and the small overheads make PSA + HRNet-W32 the most cost-effective model among all models in Table 2.

Semantic Segmentation. This task maps an input image to a stack of segmentation masks, one output mask for one semantic class. In Table 3, we compare PSA with the DeepLabV3Plus [30] baseline on the Pascal VOC2012 Aug[43] (21 classes, input image size 513 × 513, output mask size 513 × 513). PSA boosts all the baseline networks by 1.8 to 2.6 mIoU (mean Intersection over Union) with minor overheads of computation (Flops) and the number of parameters (mPara). PSA with "Res50" backbone got 79.0 mIoU, which is not only 1.8 better than the DeepLabV3Plus with the Resnet50 backbone, but also better than DeepLabV3Plus even with Resnet101.

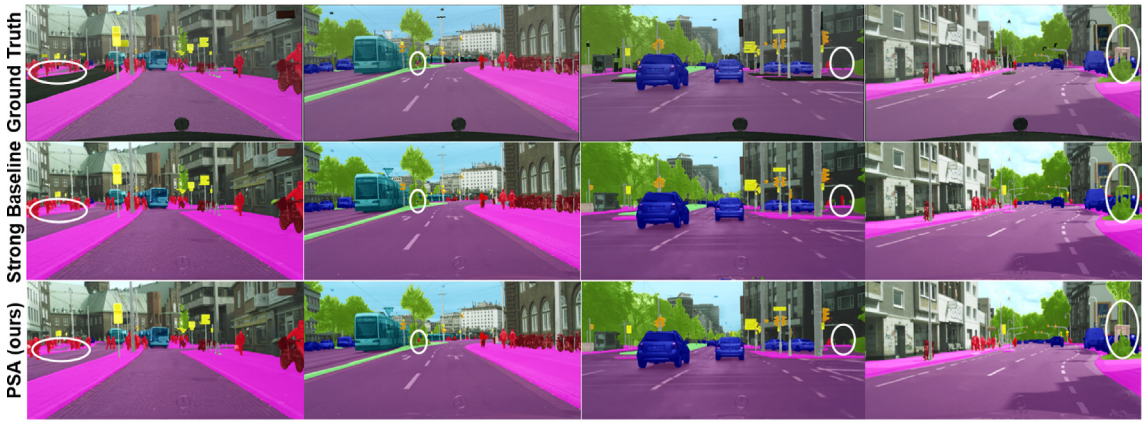
4.2. Comparing with state-of-the arts

We then apply PSA to the current state-of-the-arts of above tasks.

Top-down 2D Human Pose Estimation. To our knowledge, the current state-of-the-art results by single models were achieved by UDP-HRnet with 65.1mAP bbox detector on the MS-COCO keypoint testdev set. In Table 4, we add PSA to the UDP-Pose with HRnet-W48 backbone and achieve a new state-of-the-art AP of 79.7. PSA boosts UDP-Pose (baseline) by 1.9 points (see Fig. 3 (a) for their qualitative comparison). Moreover, we add PSA to DARK-Pose [28] to validate the effectiveness of PSA. As can be seen from Table 4, PSA can also boost DARK-Pose by 2.0 points.



(a) Strong Baseline(UDP-Pose) and PSA(UDP-Pose+PSA) for 2D Pose Estimation



(b) Strong Baseline(HRNetV2+OCR) and PSA(HRNetV2+OCR+PSA) for Semantic Segmentation

Fig. 3. Qualitative comparison of PSA(ours) and *Strong Baselines*: (a) Human Pose Estimation(UDP-Pose, Table 4) and (b) Semantic segmentation(HRNetV2-OCR, Table 5). The white eclipses highlight the fine-grained details that PSAs outperform the s.trong baselines.

Note that there is only a subtle metric difference between the parallel (p) and sequential(s) layout of PSA. We believe this partially validate that our design of the channel-only and spatial-only attention blocks has exhausted the representation power along the channel and spatial dimension.

Semantic Segmentation. To our knowledge, the current state-of-the-art results by single models were produced by HRNet-OCR (MA) [31] on the Cityscapes validation set [67](19 classes, input image size 1024×2048, output mask size 1024×2048). In Table 5, we add PSA to the basic configuration of HRNet-OCR and achieve the new state-of-the-arts mIoU of 86.95. PSA boosts HRNet-OCR (strong baseline) by 2 points(see Fig. 3 (b) for their qualitative comparison). Again that there is only a subtle metric difference between the PSA results under the parallel(p) layout and the sequential(s) layout.

4.3. Ablation study

In Table 6, we conduct an ablation study of PSA configurations on Simple-Baseline(Resnet50) [27] and compare PSA with other related self-attention methods. All the overheads, such as Flops, mPara, inference GPU memory("Mem."), and inference time ("Time")) are inference costs of **one sample**. To reduce the randomness in CUDA and Pytorch scheduling, we ran inference on MS-COCO val2017 using 4 TITAN RTX GPUs, batchsize 128 (batchsize 32/GPU), and averaged over the number of samples.

From the "PSA ablation" cell in Table 6, we observe that **(1)** the channel-only block ($\mathbf{A}^{ch}$) outperform spacial-only attention ($\mathbf{A}^{sp}$), but can be further boosted by their parallel ($[\mathbf{A}^{ch}|\mathbf{A}^{sp}]$) or sequential ($\mathbf{A}^{sp}(\mathbf{A}^{ch})$) compositions; **(2)** The parallel ($[\mathbf{A}^{ch}|\mathbf{A}^{sp}]$) or sequential ($\mathbf{A}^{sp}(\mathbf{A}^{ch})$) compositions has similar AP, Flops, mPara, inference memory(Mem.), and inference (Time).

From the "Related self-attention methods" cell, we observe that **(1)** the NL block costs the most memory while produces the least boost (2.3AP) over the baseline, indicating that NL is highly redundant. **(2)** The channel-only attention block GC is better than SE since it includes SE. GC is even better than the channel + spatial attention block CBAM because the inner-product-based mechanism in GC is more powerful than CBAM. **(3)** PSA $\mathbf{A}^{ch}$ is the best channel-only attention block over GC and SE. We believe PSA benefits from its highest channel resolution ($C/2$) and its output design. **(4)** The channel + spatial attention CBAM with a relatively early design is still better than the channel-only attention SE. **(5)** Under the same sequential layout of spatial and channel attention, PSA is significantly better than CBAM. Finally, **(6)** At similar overheads, both the parallel and sequential PSAs are better than the other compared blocks.

In Table 7, a more detailed study on different PSA configurations is conducted upon the Simple-Baseline (ResNet50) on the MS-

Table 5
Comparison with state-of-the-art semantic segmentation approaches on the Cityscapes validation set.

Method	Backbone	mIoU	iIoU cla.	IoU cat.	iIoU cat.
GridNet [46]	-	69.5	44.1	87.9	71.1
LRR-4x	-	69.7	48.0	88.2	74.7
DeepLab [30]	D-ResNet-101	70.4	42.6	86.4	67.7
LC	-	71.1	-	-	-
Piecewise [47]	VGG-16	71.6	51.7	87.3	74.1
FRRN [48]	-	71.8	45.5	88.9	75.1
RefineNet [13]	ResNet-101	73.6	47.2	87.9	70.6
PEARL [49]	D-ResNet-101	75.4	51.6	89.2	75.1
DSSPN [50]	D-ResNet-101	76.6	56.2	89.6	77.8
LKM [51]	ResNet-152	76.9	-	-	-
DUC–HDC [52]	-	77.6	53.6	90.1	75.2
SAC [53]	D-ResNet-101	78.1	-	-	-
DepthSeg [54]	D-ResNet-101	78.2	-	-	-
ResNet38 [55]	WResNet-38	78.4	59.1	90.9	78.1
BiSeNet [56]	ResNet-101	78.9	-	-	-
DFN [57]	ResNet-101	79.3	-	-	-
PSANet [58]	D-ResNet-101	80.1	-	-	-
PADNet [59]	D-ResNet-101	80.3	58.8	90.8	78.5
CFNet [60]	D-ResNet-101	79.6	-	-	-
Auto-DeepLab [61]	-	80.4	-	-	-
DenseASPP [62]	WDenseNet-161	80.6	59.1	90.9	78.1
SVCNet [63]	ResNet-101	81.0	-	-	-
ANN [64]	D-ResNet-101	81.3	-	-	-
CCNet [65]	D-ResNet-101	81.4	-	-	-
DANet [25]	D-ResNet-101	81.5	-	-	-
HRNetV2 [15]	HRNetV2-W48	81.6	61.8	92.1	82.2
OCR [66]	HRNetV2-W48	84.9	-	-	-
OCR(MA) [31]	HRNetV2-W48	85.4	-	-	-
Ours					
OCR + PSA(p)	HRNet-W48	**86.98**	**71.6**	**92.8**	**85.0**
OCR + PSA(s)	HRNet-W48	**86.76**	**71.3**	92.3	82.8

Table 6
Ablation study of PSA and comparison with related attention blocks(human pose estimation on the MS-COCO val2017 dataset with human detector [27] of 56.4AP, input size 384×288.) $\mathbf{A}^{ch}$ denotes channel-only self-attention. $\mathbf{A}^{ch}$ denotes spatial-only self-attention. $[\mathbf{A}^{ch}|\mathbf{A}^{sp}]$ denotes the parallel layout of the channel-only and spatial-only self-attention. $\mathbf{A}^{ch}(\mathbf{A}^{sp})$ denotes the sequentially layout. "Mem" and "Time" are inference costs of **one sample**, which are averaged over the val2017 set.

Method	AP ↑	AP_{50}↑	AP_{75}↑	AP_M ↑	AP_L↑	AR ↑	Flops ↓	mPara ↓	Mem. ↓	Time↓
Simple-Baseline [27]	72.2	89.3	78.9	68.1	79.7	77.6	**20.0**G	**34.0**M	1.43	**2.56**
PSA ablation										
+$\mathbf{A}^{ch}$	76.3(+4.1)	92.6	83.6	73.0	80.8	78.9	20.4G	35.3 M	1.49	2.58
+$\mathbf{A}^{sp}$	75.0(+2.8)	92.6	81.6	71.5	80.2	77.7	20.7G	35.3 M	1.45	2.63
+$[\mathbf{A}^{ch}\|\mathbf{A}^{sp}]$ (**PSA(p)**)	**76.5(+4.3)**	**93.6**	**83.6**	**73.2**	**81.0**	**79.0**	20.9G	36.5 M	1.54	2.70
+$\mathbf{A}^{sp}(\mathbf{A}^{ch})$ (**PSA(s)**)	**76.6(+4.4)**	**93.6**	**83.6**	**73.2**	**81.2**	**79.1**	20.9G	36.5 M	1.52	2.71
Related methods										
+**A** (NL [19])	74.5(+2.3)	92.6	81.5	70.9	79.9	77.3	21.1G	36.5 M	10.97	2.76
+$\mathbf{A}^{ch}$ (GC [23])	76.1(+3.9)	92.6	82.7	72.9	80.9	78.7	20.2G	34.3 M	1.47	2.69
+$\mathbf{A}^{ch}$ (SE [24])	75.7(+3.5)	93.6	82.6	72.4	80.8	78.3	20.2G	34.2 M	1.29	2.94
+$\mathbf{A}^{ch}$ (ECA [68])	75.9(+3.7)	92.6	82.7	72.7	80.9	78.7	20.2G	34.0 M	1.49	2.96
+$\mathbf{A}^{ch}$ (SK [69])	75.6(+3.4)	92.6	82.6	72.4	80.9	78.5	20.2G	34.0 M	1.47	2.90
+$\mathbf{A}^{ch}$ (SimAM [70])	75.2(+3.0)	92.6	82.6	71.9	80.4	78.1	20.2G	34.0 M	1.09	2.97
+$\mathbf{A}^{sp}(\mathbf{A}^{ch})$ (CBAM [26])	75.9(+3.7)	92.6	82.7	72.9	80.7	78.7	20.2G	34.3 M	1.49	2.96

Table 7
Ablation study of PSA to validate the best configurations of internal resolution and output non-linearity.

Method	AP ↑	AP_{50}↑	AP_L↑	AR ↑	Flops ↓	mPara ↓
Simple-Baseline [27]	72.2	89.3	79.7	77.6	20.0G	34.0 M
+PSA(s☆)	76.0**(+3.8)**	93.5	80.8	78.6	20.9G	36.5 M
+PSA(p☆)	75.9**(+3.7)**	93.4	80.8	78.5	20.9G	36.5 M
+PSA(s†)	75.7**(+3.5)**	93.0	80.5	78.4	20.8G	35.8 M
+PSA(p†)	75.6**(+3.4)**	92.9	80.5	78.3	20.8G	35.8 M
+PSA(s‡)	75.4**(+3.2)**	92.8	80.4	78.0	20.4G	35.2 M
+PSA(p‡)	75.3**(+3.1)**	92.8	80.3	77.9	20.4G	35.2 M
+PSA(s)	**76.6(+4.4)**	**93.6**	**81.2**	**79.1**	20.9G	36.5 M
+PSA(p)	**76.5(+4.3)**	**93.6**	**81.0**	**79.0**	20.9G	36.5 M

COCO dataset. Among them, **PSA(p)** and **PSA(s)** are the default configuration with $C/2$ internal resolution and SM-SD output non-linearity, and **PSA(s☆)** and **PSA(p☆)** are with only SM non-linearity and addition composition. Moreover different internal resolutions, for instance $C/4$ for **PSA(s†)** and **PSA(p†)**, and $C/8$ for **PSA(s‡)** and **PSA(p‡)** are compared under the same SM-SD output non-linearity. **This study validates the superior performance of the high internal resolution and the SM-SD non-linearity over other PSA configurations.**

5. Conclusion and limitation

We presented the Polarized Self-Attention(PSA) block towards high-quality pixel-wise mapping. PSA significantly boosts all compared DCNNs for two critical designs (1) keeping high internal resolution in both polarized channel-only and spatial-only attention branches, and (2) incorporating a nonlinear composition that fully leverages the high-resolution information preserved in the PSA branches. PSA can potentially benefit any computer vision tasks with pixel-wise mapping.

It is still not clear how PSA would best benefit pixel-wise mapping embedded with the classification and displacement regression in complex DCNN heads, such as those in the instance segmentation, anchor-free object detection and panoptic segmentation tasks. To our knowledge, most existing work with self-attention blocks only inserted blocks in the backbone networks. Our future work is to explore the use of PSAs in DCNN heads.

CRediT authorship contribution statement

Huajun Liu: Conceptualization, Methodology, Software, Investigation, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Resources. **Fuqiang Liu:** Software. **Xinyi Fan:** Software. **Dong Huang:** Conceptualization, Methodology, Project administration, Supervision, Resources, Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Dong Huang reports a relationship with Carnegie Mellon University that includes: employment.

Acknowledgement

aaa

References

[1] O. Russakovsky, H.S. Jia Deng, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, Int. J. Comput. Vision 115 (2015) 211–252.
[2] R. Girshick, Fast r-cnn, in: ICCV, 2015.
[3] Z. Luo, Z. Wang, Y. Huang, T. Tan, E. Zhou, Rethinking the heatmap regression for bottom-up human pose estimation, CVPR (2021).
[4] Z. Zhong, Z.Q. Lin, R. Bidart, X. Hu, I.B. Daya, Z. Li, W.-S. Zheng, J. Li, A. Wong, Squeeze-and-attention networks for semantic segmentation, CVPR (2020).
[5] M. Treml, J. Arjona-Medina, T. Unterthiner, R. Durgesh, F. Friedmann, P. Schuberth, A. Mayr, M. Heusel, M. Hofmarcher, M. Widrich, B. Nessler, S. Hochreiter, Speeding up semantic segmentation for autonomous driving, NIPS (2016).
[6] H.-P. Chiu, V. Murali, R. Villamil, G.D. Kessler, S. Samarasekera, R. Kumar, Augmented reality driving using semantic geo-registration, in: IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 2018.
[7] G. Litjens, T. Kooi, B. Ehteshami, B. Arnaud, A. Adiyoso, S. Francesco, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.
[8] Q. Wang, J. Gao, W. Lin, Y. Yuan, Pixel-wise crowd understanding via synthetic data, Int. J. Comput. Vision 129 (2021) 225–245.
[9] J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, NIPS (2014).
[10] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.
[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CVPR, 2016.
[12] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press (2016), URL: http://www.deeplearningbook.org.
[13] G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, CVPR, 2017.
[14] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, CVPR, 2019.
[15] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, B. Xiao, Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 10 (2020) 5686–5696.
[16] A. Vaswani, N. Shazeer, N. Parmar, J.U. abd Llion Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017.
[17] A.V. Peter Shaw, Jakob Uszkoreit, Self-attention with relative position representations, arXiv:1803.02155 (2018).
[18] J.-B. Cordonnier, A. Loukas, M. Jaggi, On the relationship between self-attention and convolutional layers, ICLR, 2020.
[19] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, CVPR, 2018.
[20] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, L. Zhang, Second-order attention network for single image super-resolution, CVPR (2019).
[21] B.N. Xia, Y. Gong, Y. Zhang, C. Poellabauer, Second-order non-local attention networks for person re-identification, in: ICCV, 2019.
[22] J. Hu, L. Shen, S. Albanie, G. Sun, A. Vedaldi, Gather-excite: Exploiting feature context in convolutional neural networks, in: NIPS, 2018.
[23] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, Gcnet: Non-local networks meet squeeze-excitation networks and beyond, in: ICCV, 2019.
[24] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, in: CVPR, 2018.
[25] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, CVPR (2019).
[26] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, ECCV, 2018.
[27] B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, in: ECCV, 2018.
[28] F. Zhang, X. Zhu, H. Dai, M. Ye, C. Zhu, Distribution-aware coordinate representation for human pose estimation, CVPR (2020).
[29] J. Huang, Z. Zhu, F. Guo, G. Huang, The devil is in the details: Delving into unbiased data processing for human pose estimation, in: CVPR, 2020.
[30] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.
[31] A. Tao, K. Sapra, B. Catanzaro, Hierarchical multi-scale attention for semantic segmentation, in: arXiv preprint arXiv:2005.10821, 2020.
[32] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: ICCV, 2017.
[33] X. Zhou, D. Wang, P. Krähenbühl, Objects as points, in: arXiv preprint arXiv:1904.07850, 2019.
[34] B. Cheng, M.D. Collins, Y. Zhu, T. Liu, T.S. Huang, H. Adam, L.-C. Chen, Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation, CVPR (2020).
[35] I. Bello, B. Zoph, A. Vaswani, J. Shlens, Q.V. Le, Attention augmented convolutional networks, in: ICCV, 2019.
[36] J.-M. Andreoli, Convolution, attention and structure embedding, in: NIPS, 2019.
[37] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, Stand-alone self-attention in vision models, in: NIPS, 2019.
[38] Y. Zhang, K. Li, K. Li, B. Zhong, Y. Fu, Residual non-local attention networks for image restoration, ICLR (2019).
[39] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable convnets v2: More deformable, better results, in: CVPR | arXiv:1811.11168, 2019.
[40] Z. Shen, M. Zhang, H. Zhao, S. Yi, H. Li, Efficient attention: Attention with linear complexities, arXiv:1812.01243 (2020).
[41] J. Park, S. Woo, J.-Y. Lee, I. SoKweon, Bam: bottleneck attention module, BMVC (2018).
[42] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollár, Microsoft coco: Common objects in context, arXiv:1405.0312 (2014).
[43] M. Everingham, S.M.A. Eslami, L.V. Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, Int. J. Comput. Vision 111 (1) (2015) 98–136.
[44] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, ECCV, 2016.
[45] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, CVPR (2018).
[46] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Tremeau, C. Wolf, Residual conv-deconv grid network for semantic segmentation, BMCV (2017).
[47] G. Lin, C. Shen, A. van dan Hengel, I. Reid, Efficient piecewise training of deep structured models for semantic segmentation, in: CVPR, 2016.
[48] T. Pohlen, A. Hermans, M. Mathias, B. Leibe, Full-resolution residual networks for semantic segmentation in street scenes, CVPR (2017).

[49] X. Jin, X. Li, H. Xiao, X. Shen, Z. Lin, J. Yang, Y. Chen, J. Dong, L. Liu, Z. Jie, J. Feng, S. Yan, Video scene parsing with predictive feature learning, in: ICCV, 2017.
[50] X. Liang, H. Zhou, E. Xing, Dynamic-structured semantic propagation network, CVPR (2018).
[51] C. Peng, X. Zhang, G. Yu, G. Luo, J. Sun, Large kernel matters – improve semantic segmentation by global convolutional network, CVPR (2017).
[52] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G. Cottrell, Understanding convolution for semantic segmentation, WACV (2018).
[53] R. Zhang, S. Tang, Y. Zhang, J. Li, S. Yan, Scale-adaptive convolutions for scene parsing, in: ICCV, 2017.
[54] S. Kong, C. Fowlkes, Recurrent scene parsing with perspective understanding in the loop, CVPR, 2018.
[55] Z. Wu, C. Shen, A. van den Hengel, Wider or deeper: Revisiting the resnet model for visual recognition, Pattern Recogn. 90 (2019) 119–133.
[56] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Bisenet: Bilateral segmentation network for real-time semantic segmentation, in: ECCV, 2018.
[57] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Learning a discriminative feature network for semantic segmentation, CVPR (2018).
[58] H. Zhao, Y. Zhang, S. Liu, J. Shi, C.C. Loy, D. Lin, J. Jia, Psanet: Point-wise spatial attention network for scene parsing, ECCV (2018).
[59] D. Xu, W. Ouyang, X. Wang, N. Sebe, Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing, CVPR (2018).
[60] H. Zhang, C. Wang, J. Xie, Co-occurrent features in semantic segmentation, CVPR, 2019.
[61] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. Yuille, L. Fei-Fei, Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation, CVPR (2019).
[62] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, CVPR, 2017.
[63] H. Ding, X. Jiang, B. Shuai, A.Q. Liu, G. Wang, Semantic correlation promoted shape-variant context for segmentation, CVPR (2019).
[64] Z. Zhu, M. Xu, S. Bai, T. Huang, X. Bai, Asymmetric non-local neural networks for semantic segmentation, in: ICCV, 2019.
[65] Z. Huang, X. Wang, Y. Wei, L. Huang, H. Shi, W. Liu, T.S. Huang, Ccnet: Criss-cross attention for semantic segmentation, in: ICCV, 2019.
[66] Y. Yuan, X. Chen, J. Wang, Object-contextual representations for semantic segmentation, ECCV, 2020.
[67] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, CVPR (2016).
[68] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Eca-net: Efficient channel attention for deep convolutional neural networks, in: CVPR, 2019.
[69] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, CVPR, 2019.
[70] L. Yang, R.-Y. Zhang, L. Li, X. Xie, Simam: A simple, parameter-free attention module for convolutional neural networks, in: Proceedings of the 38th International Conference on Machine Learning, 2021.

Huajun Liu received his Ph.D. degree at School of Computer Science of Nanjing University of Science and Technology in 2007, and worked as a postdoctoral fellow at Robotics Institute in Carnegie Mellon University from 2018 to 2020.
His research interests include autonomous vehicle, computer vision, information fusion and deep learning. He has published more than 30 papers on ICCV, Information Fusion, Industrial Robots, and Sensors etc.

Fuqiang Liu received his PhD degree in Automation from Harbin Engineering University of China and became as a postdoctoral research associate at Carnegie Mellon University, in 2019.
His research focuses on deep learning perception on autonomous driving and computer vision.

Xinyi Fan is pursuing his master's degree in Nanjing University of Science and Technology. His research focuses on computer vision, especially on contextual modeling and representation learning for images.

Dong Huang is a senior project scientist at the Robotics Institute at Carnegie Mellon University, USA. Since 2018, he directs the DeLight Lab at Carnegie Mellon University (https://www.ri.cmu.edu/robotics-groups/delight/).
His research focuses on deep learning perception on multi-modality systems and embedded platforms. He received his M.Sc. in Automation and PhD degrees in Computer Science from University of Electronic Science and Technology of China, respectively, in 2005 and 2009, Chengdu, China. He became as a postdoctoral research associate, project scientist and senior project scientist in 2009, 2012 and 2018, respectively, at Carnegie Mellon University. He also worked as a Research Scientist (Part time) at Facebook Inc. between 2017 and 2019.