J. Vis. Commun. Image R. 89 (2022) 103620

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Short communication

Improving small objects detection using transformer☆

Shikha Dubey [a], Farrukh Olimov [b], Muhammad Aasim Rafique [a], Moongu Jeon [a,*]

[a] Gwangju Institute of Science and Technology (GIST), School of Electrical Engineering and Computer Science, Gwangju, South Korea
[b] Threat Intelligence Team, Monitorapp, Seoul, South Korea

ARTICLE INFO

Keywords:
Normalized inductive bias
Features fusion
Small-objects detection
Transformer
Self-attention

ABSTRACT

General artificial intelligence counteracts the inductive bias of an algorithm and tunes the algorithm for out-of-distribution generalization. A conspicuous impact of the inductive bias is an unceasing trend in improving deep learning performance. Although a quintessential attention-based object detection technique, DETR, shows better accuracy than its predecessors, its accuracy deteriorates for detecting small-sized (in-perspective) objects. This study examines the inductive bias of DETR and proposes a normalized inductive bias for object detection using data fusion, SOF-DETR. A technique of lazy-fusion of features is introduced in SOF-DETR, which sustains deep contextual information of objects present in an image. The features from multiple subsequent deep layers are fused for object queries that learn long and short-distance spatial association in an image using the attention mechanism. Experimental results on the MS COCO and Udacity Self Driving Car datasets assert the effectiveness of the added normalized inductive bias and feature fusion techniques, showing increased COCO mAP scores on small-sized objects.

1. Introduction

Object detection is one of the long-standing research topics in computer vision (CV) and more challenges in the detection surface with every technology update and artificial intelligence (AI) venture. Most futuristic applications demand rigorous, accurate, and efficient detection of all objects of interest in an image. AI scooped by advancements in deep learning (DL) generally benefits from a hard inductive bias of convolutional neural networks (CNNs) with the region proposal network (RPN) in two-stage object detection [1–4]. Later, single-shot object detection using anchors efficiently produced competitive results [5–7]. Despite outstanding results, RPN generates highly overlapping region proposals for the same set of objects and needs handcrafted post-processing like non-maximum suppression (NMS) [4–6]. Recently a new paradigm of inductive bias, transformer [8], is introduced, which has shown better performance in the natural language processing (NLP) than conventional DL network compositions. Subsequently, transformers are adopted in a wide range of machine learning (ML) problems like image captioning [9], speech processing [10], biomedical imaging [11], and the most perceptible CV [12–14] domain. Detection using the transformer (DETR) [13] proposes a new DL architecture for object detection in images that efficiently generates bounding boxes and classes in parallel using the self-attention mechanism proposed in the transformer. DETR uses bipartite matching to remove redundant detection and performs better than conventional DL-based object detection networks. Nevertheless, detecting small-sized objects is challenging due to inadequate resolution of the objects and relatively fewer training images containing more small-sized objects in the training dataset. This particular study approaches the challenges of detecting small-sized objects with inadequate resolutions in a transformer-based model without affecting its capability to detect medium and large-sized objects. We indicate detected objects at a different scale in Fig. 1.

Evidently, major performance breakthroughs in AI using artificial neural networks are exhibited using inductive bias exposited through a network composition: convolutional for CV [12,15] or sequential for NLP [8,16]. Generally, the hard inductive biases are referred as a hard-encoded form of the possible architectural limitations of the CNNs for efficient sample training. In contrast, soft inductive biases are introduced to preserve the local spacial features in the self-attention module of the transformer for sample efficiency [17]. Recently, these inductive biases were explored to improve the performance of transformer-based models, ViT [12] and ConViT [17], in the image classification task. These studies encouraged our current model to integrate the strengths of inductive biases for unraveling the present detection challenge. Consequently, this work examines the impact of inductive biases on the transformer-based object detection model.

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail addresses: shikha.d@gm.gist.ac.kr (S. Dubey), olimov.farrukh@gm.gist.ac.kr (F. Olimov), aasimrafique@gist.ac.kr (M.A. Rafique), mgjeon@gist.ac.kr (M. Jeon).

https://doi.org/10.1016/j.jvcir.2022.103620
Received 10 March 2022; Received in revised form 3 July 2022; Accepted 13 August 2022
Available online 6 September 2022

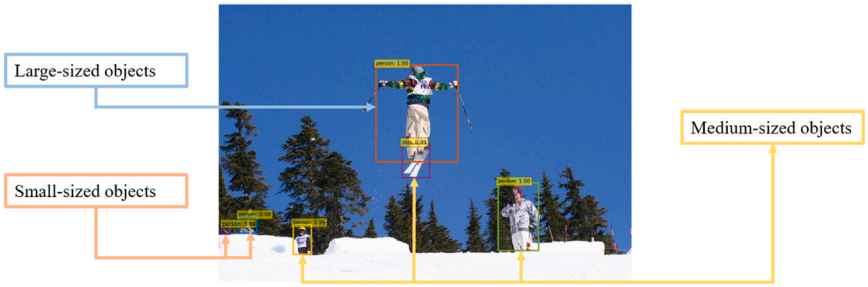

Fig. 1. Object detection at different sizes: small-, medium- and large-sized objects.

CNN explores local regions in spatial data using confined location connections in the network layers and provides performance improvements, whereas the transformer pays attention to learned features from spatially related regions. However, passing a direct encoding of high-level CNN extracted features to the transformer losses some information from small local regions, particularly for small-sized objects in an image. Therefore, in this work, we propose a model, small object favoring detection using the transformer (SOF-DETR), where a normalized inductive bias is included, which pronounces features for small-sized objects without losing other objects essential features before passing them through the transformer's encoder layer. Normalized inductive bias applies group normalization (GN) [18] on features from successive spatial filtering layers in a CNN and fuses them before passing to the attention module. Fusion techniques are commonly used to combine multi-sensor or multi-source data [19]. In our study, multiple layers of the normalized convolutional layers are fused together to give some hard-encoded inductive bias to favor the small-sized objects. Then, these advanced deep features are progressed through the transformer network as object queries, and thus the network learns associations among present objects in the image by using the attention mechanism. SOF-DETR predicts unique bounding boxes associated with a single object by utilizing a global set-based cost function for the bi-partite matching technique similar to studies [13,20].

Contributions: This study contributes to the object detection field as follows. First, we introduce a normalized inductive bias for detection using a transformer to get distinct features from different filtering layers of a CNN. Second, the normalized filters are fused to generate diverse and focused self-attention maps. Our extensive experimentation demonstrates the effectiveness of particular choices of hard inductive bias, normalization techniques for filtered features, and fusion of features. Detailed explanations are given in the following sections.

2. Method

The architecture of the proposed model, SOF-DETR, is depicted in Fig. 2. SOF-DETR introduces a new composition of the neural networks in the transformer-based detection model, DETR. It channelizes the hard inductive bias of a CNN to the soft inductive bias of a transformer by interleaving a normalized inductive bias on fused feature maps from different convolutional blocks. The model consists of three modules: convolutional backbone, normalized inductive bias, and transformer detection. For unique detection, this study follows set-based object detection [20] to associate a unique bounding box with an object. The composition of each module is given in the following sections.[1]

[1] The code is publicly available on https://github.com/shikha-gist/SOF-DETR/

2.1. Convolutional backbone

In this study, we have utilized the standard convolutional model, ResNet [21], as a backbone of SOF-DETR for extracting features in an image. We have experimented our model on the ResNet-50. ResNet consists of five distinct building blocks of stacked convolutional layers with layer names conv1, conv2_x, conv3_x, conv4_x, and conv5_x, followed by one fully connected layer (a detailed structure is given in [21]), where x is the number of stacked blocks of that particular building block and varies in different architectures of ResNet. For example, in ResNet-50, there are 3, 4, 6, and 3 building blocks of conv2, conv3, conv4, and conv5, respectively. Moreover, each building blocks conv2, conv3, conv4, and conv5 consists of 3 convolutional layers. In SOF-DETR, we have utilized a pre-trained ResNet-50 model. Generally, we can obtain four intermediate layers, 1, 2, 3, and 4, from the pre-trained ResNet model. In SOF-DETR, intermediate layers 3 and 4 are chosen after experimentation for further processing. Besides, intermediate layers 3 and 4 are achieved from the outputs of blocks conv4_6 and conv5_3 with feature dimensions 1024 and 2048, respectively.

2.2. Normalized inductive bias

DETR practices the CNN hard inductive bias for features and soft inductive bias to learn the relationship. The hard inductive bias of a CNN helps to explore the spatially connected features in spatial data using its local receptive fields. However, the benefits become a limitation with the abundance of available data. The soft inductive bias realized by an attention mechanism in transformers tweaks and refines the distant spatial features while holding the learned relations of local features intact during training and inference [17]. However, a challenge persists in detecting small-sized objects due to their infinitesimal presence, presumably caused by missing indispensable features of the objects and thus generating poor attention maps for such objects. Consequently, to overcome this challenge in transformer-based detection, SOF-DETR channelizes hard inductive bias and normalizes it before passing it to the soft inductive bias of the transformer encoder. SOF-DETR combines features from successive convolution layers from a backbone network. The features are spatially localized and downscaled in the successive layers, and a normalization of the inductive bias in the feature space helped to pronounce the spatially reduced effect of small-object features. The normalized hard inductive bias fuses the features extracted at different convolution blocks so that the essential features of the small-sized are intact for further processing. Particularly, features from the last two layers (intermediate layers 3 and 4) of the backbone model are used by down-projecting the features from the high-level (layer 4) using a de-convolutional layer of 1024-size and transferring it through a convolutional layer of 1024-size, then fused with the lower-level (layer 3) features. A detailed ablation study in Section 3.3.2 shows that fusing more layers decreases the overall performance. These features are fused after GN using element-wise-summation as given in the following equations.

$$GN_L = GN\left(Conv\left(DC\left(F(I,L)\right)\right)\right), \tag{1}$$

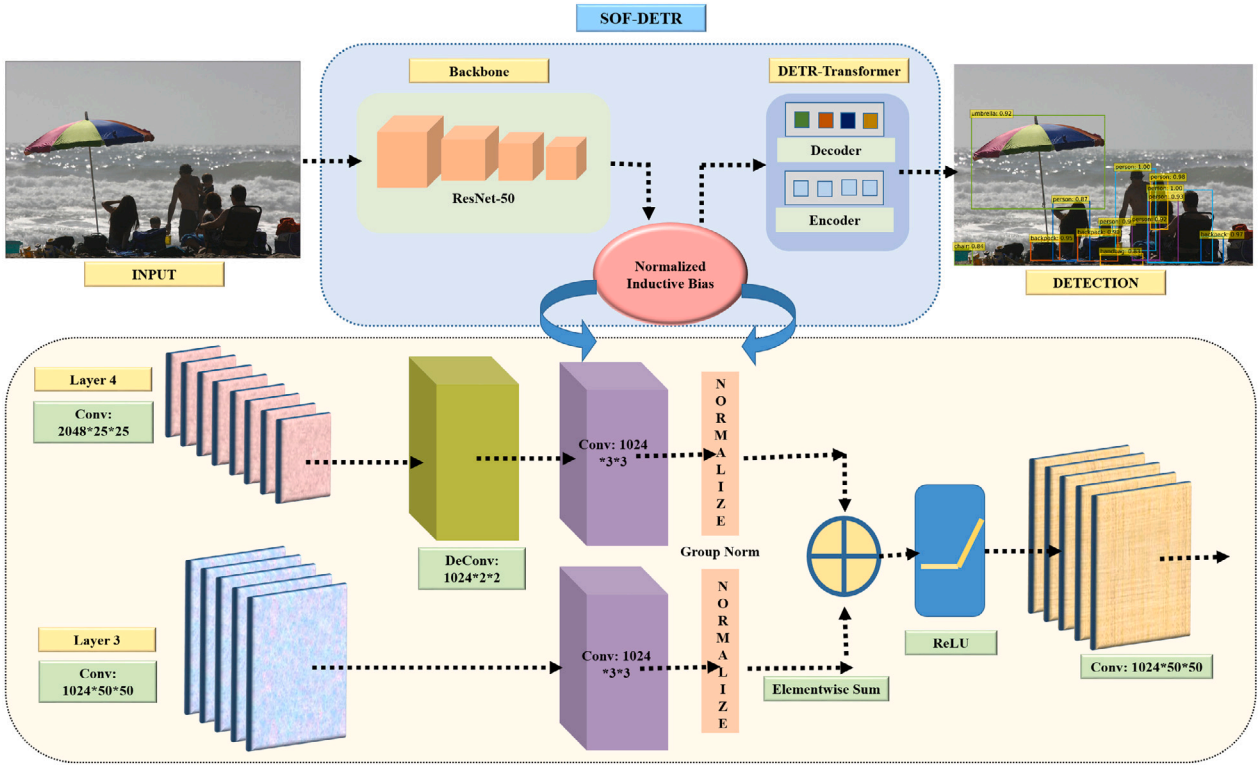

Fig. 2. SOF-DETR architecture.

$$GN_{L-1} = GN\left(Conv\left(F\left(I, L-1\right)\right)\right), \tag{2}$$

$$F_c = GN_L \oplus GN_{L-1}, \tag{3}$$

where $DC(\cdot)$ and $Conv(\cdot)$ represent 2D-deconvolution and 2D-convolution layers, respectively. In Eqs. (1) and (2), $F(\cdot)$ gives extracted features from an image, I of layers L or $L-1$. Whereas $GN(\cdot)$ stands for GN layer [18], GN_L and GN_{L-1} represent normalized features from layers L and $L-1$, respectively. In Eq. (3), the symbol $\oplus$ represents the element-wise-summation operation of the fusion technique, and F_c denotes combined features. These combined features are then passed through the relu layer, followed by a convolutional layer of size 1024 for extracting deep features from these combined high-level and low-level features. These extensive features from small-sized objects channelize normalized inductive bias to support the transformer in generating more focused self-attention maps. Concatenating these features has not improved the current model; therefore, we have experimented with the element-wise summation of the low-level and high-level features (demonstrated in Section 3.3.2). The proposed model, SOF-DETR, employs GN over other normalization techniques since it is more suitable while training the model with a small batch size. Another inspiration for using GN is that the features after normalization are passed through channel fusion; therefore, GN also helps to normalize the inductive bias in the channel dimension and shows superior performance compared to other normalization techniques and discussed in detail in Section 3.3.2. Now, we transfer these contextual features, including inductive bias, to the transformer module for further processing.

2.3. Transformer detection

The transformer architecture [13] for object detection consists of a simple encoder–decoder architecture and a feed-forward network (FFN) for bounding box and class predictions. **The encoder layer** consists of standard multi-head self-attention layers and fully connected layers [8]. Positional encoding is added to image data [28], and the encoded features of objects (N objects) each of d-dimension are passed through the decoder layers. The queried objects (N) are supposedly larger than the number of objects present in an image in a particular dataset. **The decoder layer** also follows the standard transformer-based architecture where the embedded features of N objects generate attention maps using multi-head attention with the encoder output and itself. SOF-DETR follows DETR, where all the embeddings of N objects are passed together as queries and generate the output for each object simultaneously as decoder embedding. The embedded outputs are individually decoded into bounding box coordinates and class labels by the FFN. The bounding boxes are normalized center coordinates, height, and width of a predicted class.

Following studies [13,20], for a set of N-sized embedded outputs generated from the decoder, we utilize a set-based loss function to train the network and assign each object a unique bounding box. The lowest cost for finding bipartite matching among sets of ground-truths G and detections $\hat{D}$ elements, we calculate the best permutation, $\hat{\vartheta}$, as below:

$$\hat{\vartheta} = arg \min_{\vartheta \in \alpha_N} \sum_{i}^{N} \mathbb{C}_{bm}(G_i, \hat{D}_{\vartheta(i)}), \tag{4}$$

where $\mathbb{C}_{bm}$ is element-wise bipartite matching cost, calculated using the Hungarian algorithm similar to [20], both sets G and $\hat{D}$ have N

Table 1
Performance analysis using mAP metrics on MS COCO val set, where s, m, and l stand for small-, medium-, and large-sized objects, respectively. AP stands for average precision.

Models	AP^{all}	AP^{50}	AP^{75}	AP^{s}	AP^{m}	AP^{l}
Pre-trained Faster-RCNN [4]	36.7	–	–	14.1	32.9	47.1
DETR (ResNet-50)[13]	42.0	**62.4**	44.2	20.5	45.8	61.1
SOF-DETR (ResNet-50) (Ours)	**42.7**	61.8	**45.4**	**21.7**	**45.9**	**61.5**

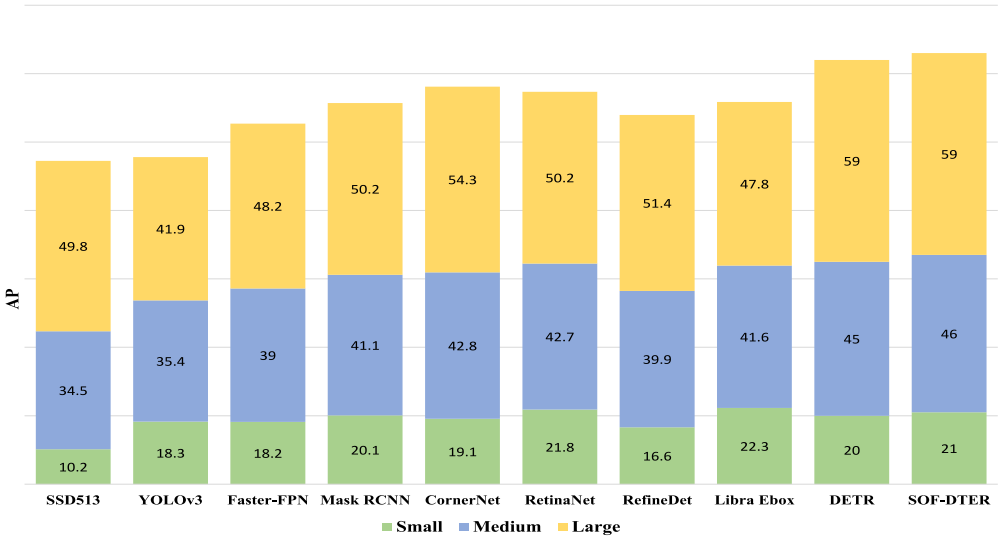

Fig. 3. AP metrics comparison among recent algorithms and SOF-DETR on MS COCO test-dev set for small-sized, medium-sized, and large-sized objects. SSD513 [5], YOLOv3 [22], Faster-FPN [3], Mask RCNN [23], CornerNet [24], RetinaNet [25], RefineDet [26], Libra EBox [27], DETR [13].

Table 2
Online performance analysis using mAP metrics on MS COCO test-dev set.

Models	Backbone	AP^{all}	AP^{50}	AP^{75}	AP^{s}	AP^{m}	AP^{l}
YOLOv2 [29]	DarkNet-19	21.6	44.0	19.2	5.0	22.4	35.5
Faster-RCNN [4]	VGG16	23.5	43.9	22.6	8.1	25.1	34.7
SSD513 [5]	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
YOLOv3 [22]	DarkNet-53	33.0	57.9	34.4	18.3	35.4	41.9
DSSD513 [30]	ResNet-101	33.2	53.3	35.2	13.0	35.4	51.1
Faster-FPN [3]	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
RefineDet [26]	ResNet-101	36.4	57.5	39.5	16.6	39.9	51.4
Mask RCNN [23]	ResNet-101	38.2	60.3	41.7	20.1	41.1	50.2
ExtremeNet [31]	Hourglass-104	40.2	55.5	43.2	20.4	43.2	53.1
CornerNet [24]	Hourglass-104	40.6	56.4	43.2	19.1	42.8	54.3
RetinaNet[a] [25]	ResNet-50	36.1	55.9	38.7	20.1	38.9	45.2
FoveaBox[a] [32]	ResNet-50	36.8	56.7	39.3	20.3	40.1	45.2
FSAF[a] [33]	ResNet-50	37.0	56.1	9.2	20.4	39.2	46.0
FCOS[a] [34]	ResNet-50	37.1	56.6	39.6	20.6	39.8	46.9
Libra RCNN[a] [35]	ResNet-50	37.6	57.0	40.0	21.3	40.4	47.0
Libra EBox [27]	ResNet-50	38.4	59.4	41.1	22.3	41.6	47.8
DETR[b] [13]	ResNet-50	42.2	62.0	44.8	20.2	45.0	59.0
SOF-DETR (Ours)	*ResNet-50*	*43.0*	*62.1*	*45.9*	*21.1*	*45.9*	*59.0*

[a]Models' result are taken from the study [27].
[b]Model is evaluated online by us.

number of objects, where G is padded if the number of objects in G is less than N. Ground-truths G has pair elements of the class and a bounding box for each object. Similarly, predictions $\hat{D}$ has predicted class, probability, and bounding box for each object. This bipartite matching takes place for both: class and bounding box predictions. The loss function used is Hungarian loss for all matched pairs, and it is calculated similarly to works [13,20]. Class imbalances are handled likewise to [4]. The bounding box loss, $\mathbb{C}_{bbox}$, is calculated utilizing l_1 loss and generalized intersection-over-union (IoU) loss, L_{IoU} similar to work [36]. The loss function and shared layer-norm are added to each decoder layer. SOF-DETR is trained using this loss function and predicts a unique bounding box and class for each object present in an image.

3. Experimentation and discussion

3.1. Dataset

SOF-DETR is evaluated on two public object detection datasets: MS COCO (2017) [37] and Udacity Self Driving Car [38]. Each dataset contains a fair number of small-sized objects, and the denotation of the size is appraised in accordance with the MS COCO annotations. A labeled object with a bounding box having an area less than 32^2(pixels) is a small-sized object, the area between 32^2 and 96^2 is a medium-sized object, and the objects with an area greater than 96^2 are large-sized. Fig. 1 provides a visual reference of different object sizes. In the case of occluded objects, the area is computed on the size of a partially labeled bounded box. MS COCO detection dataset consists of 41% of the small-sized objects, whereas the Udacity Self Driving Car dataset consists of 57% of the same. Additionally, data-augmentation techniques like scaling, random cropping with a probability of 0.5, and resizing to $800 * 800$ are used for training.

MS COCO dataset labeled 80 object classes in 115k images for the training set and 5k images for the validation dataset (val set). Additional 20k images without annotations are available for online evaluations only (test-dev set). There are at least 7 objects and at most 63 objects present (labeled) in a single image of the training set. Original Udacity Self Driving Car dataset consists of 4 object classes with 15k images [38], but it missed objects in annotations. This study uses a version annotated by Roboflow [39], which has additional annotations and classes. MS COCO evaluation metrics are used for this dataset, and the object classes are mapped to "person, car, truck and traffic light" label classes of MS COCO.

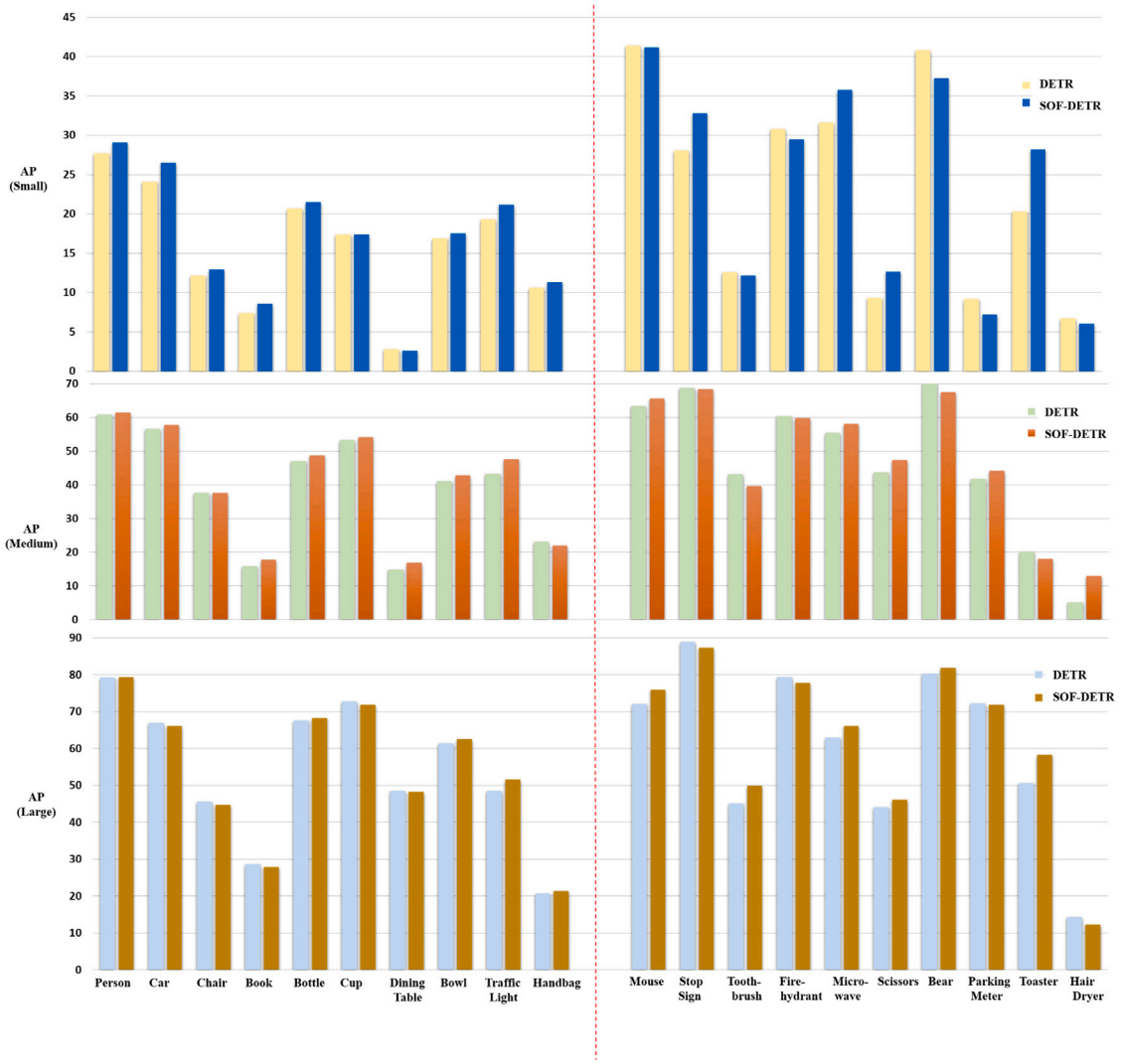

Fig. 4. AP metrics comparison on selected object classes between SOF-DETR and DETR on the MS COCO test-dev set for small-sized, medium-sized, and large-sized objects. The object classes on the left-side and right-side of the red-dotted line are top-10 and last-10 object classes of the training dataset, respectively (based on the number of objects of that particular class in the training dataset).

3.1.1. Evaluation metrics

This study presents quantitative and qualitative evaluations of SOF-DETR. The quantitative results are generated using a standard MS COCO evaluation metric: mAP (mean average precision) of bounding boxes averaged on thresholds $\in [0.5 : 0.05 : 0.95]$ for all detection. Moreover, we have also compared the proposed model's performance using conventional precision, recall, and F-1 scores @0.5 for small-sized, medium-sized, and large-sized objects. Furthermore, the PR curve (Precision-Recall curve) is also utilized for performance analysis. The qualitative results are depicted with detection in selected images and the attention maps of transformers.

3.2. Implementation details

SOF-DETR is trained similar to DETR for a fair evaluation of the proposed normalized inductive bias module. It uses an AdamW optimizer with an initial learning rate of $10^{(-4)}$ and weight decay of $10^{(-4)}$ with beta values, $(\beta_1, \beta_2) = (0.9, 0.999)$. The learning rate for the backbone (ResNet-50) is $10^{(-5)}$. SOF-DETR uses 0.1 dropout and Xavier initialization [40] for starting weights. We have experimented SOF-DETR without any additional dilation layer [41], and future work can extend this work. We have reported the results after 500 epochs of training for an overall evaluation, with a learning rate dropping by 10 after 300 epochs. SOF-DETR is trained on 8–V100 Nvidia-GPUs on a single node with a batch size of 3 on each GPU (performs better than 2, 4, and 6 batch sizes) and tested on a single GPU and single node with 2 batch sizes.

Table 3
Performance analysis on Udacity Self Driving Car Dataset, where s, m, and l stand for small-, medium-, and large-sized objects, respectively.

Models	Backbone	AP^{all}	AP^{50}	AP^{75}	AP^{s}	AP^{m}	AP^{l}
DETR	ResNet-50	15.4	38.9	9.6	8.0	28.0	53.7
SOF-DETR		**18.3**	**43.0**	**12.9**	**10.3**	**31.0**	**55.3**

3.3. Performance analysis of SOF-DETR

3.3.1. Quantitative analysis

We have compared the performance of SOF-DETR with transformer-based algorithm DETR [13] and other recent state-of-the-art algorithms [5,24,25,27,30,32–35]. Tables 1 and 2 show quantitative analysis of SOF-DETR on val set and test-dev set, respectively, for MS COCO dataset, while the performance analysis for the Udacity Self Driving Car dataset [38] is demonstrated in Table 3. In Table 1, SOF-DETR is compared with the pre-trained faster-RCNN [4] and DETR. SOF-DETR

Table 4
Ablation study: Performance analysis using mAP metrics on MS COCO val set for a different number of fusion layers, fusion techniques, various backbones, and normalization effects. All models are trained for 150 epochs.

Backbone	Norm-type	Fusion-type	Fusion-layers	AP^{all}	AP^{50}	AP^{75}	AP^{s}	AP^{m}	AP^{l}
ResNet-18	Group	⊕	4 & 3	36.7	56.0	38.5	15.6	39.7	55.1
ResNet-34				38.8	58.4	41.0	16.9	42.3	57.6
ResNet-50	Group	⊕	4 & 3 & 2	39.6	60.2	41.3	17.8	43.0	59.2
ResNet-50	–	–	–	39.8	60.4	41.5	18.1	43.0	59.1
	Group	||		38.6	58.7	40.7	17.4	42.2	57.7
	–	⊕	4 & 3	35.6	55.2	37.2	14.7	38.2	54.2
	Batch			40.4	60.3	42.7	19.0	44.1	59.7
	Instance			40.4	60.2	42.6	19.1	43.6	58.2
	Layer			40.7	60.5	43.0	**19.2**	44.2	59.8
	Group-8			40.5	60.4	43.0	18.8	43.9	60.0
	Group-32			40.6	60.6	42.7	18.9	44.2	59.7
ResNet-50	*Group-16*	⊕	4 & 3	***40.9***	***60.7***	***43.2***	***19.2***	***44.3***	***60.1***

⊕ and || stand for the element-wise summation and concatenation fusion techniques, respectively.

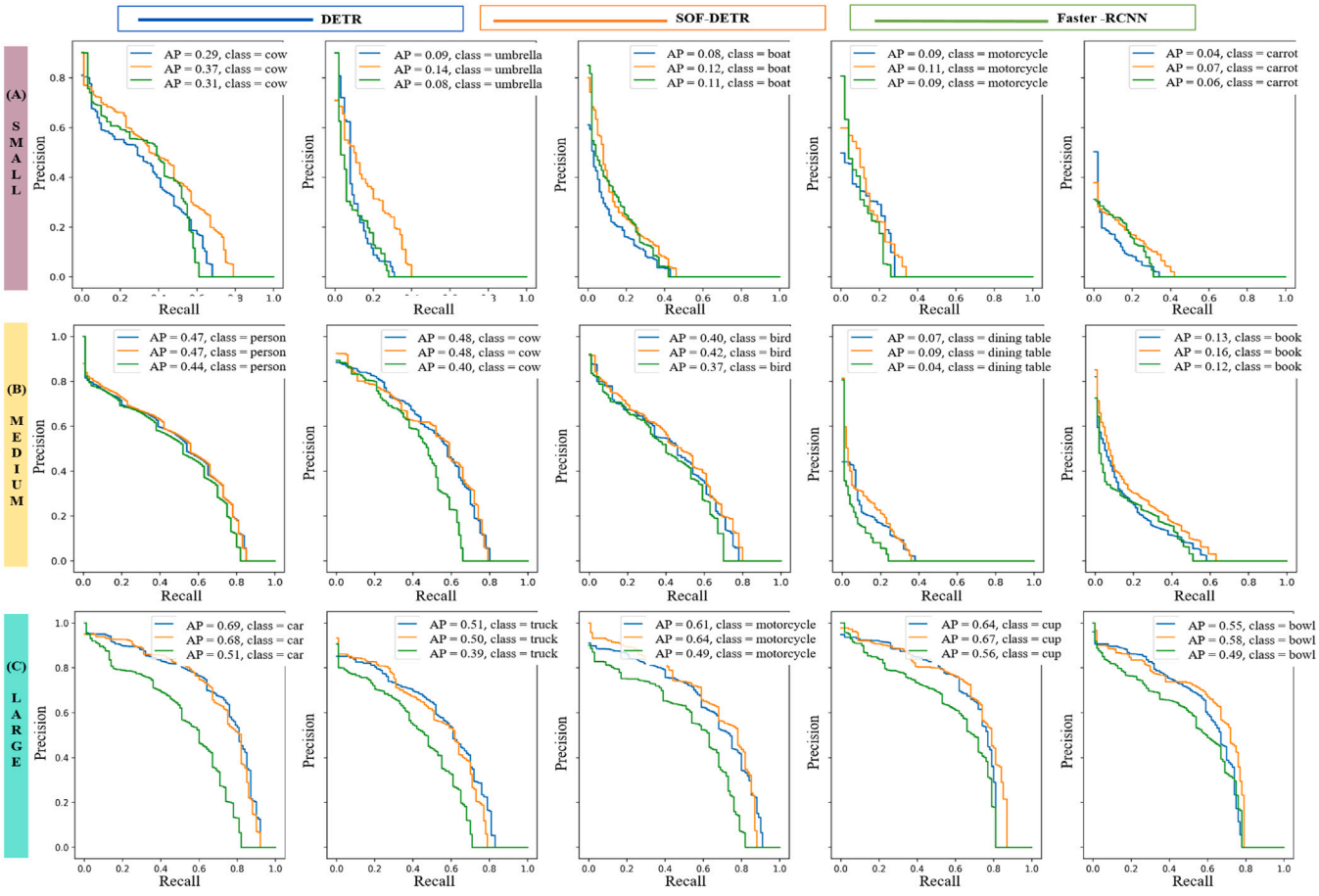

Fig. 5. PR-Curves comparison among DETR, Faster-RCNN, and SOF-DETR on top-5 object classes for small-sized, medium-sized, and large-sized MS COCO val set objects. The curve surpassing other curves shows the best performance in all. The average precision of each object class is also given in the top-right corner.

outperforms DETR with a 0.7% improvement in mAP score overall, and 1.2%, 0.1% and 0.4% improvements in detecting small-sized, medium-sized, and large-sized objects, respectively. Moreover, Table 2 depicts the online evaluation of SOF-DETR with other state-of-the-art algorithms. SOF-DETR sustains the performance trade-off among small-sized, medium-sized, and large-sized object detection compared to recent detection algorithms [27,32–35], where the detection performance deteriorates for medium-sized and large-sized objects while improving the detection of small-sized objects. Moreover, for a fair comparison, SOF-DETR uses ResNet-50 as the backbone. However, detection results with various other backbones are discussed in the ablation study (Section 3.3.2). Further, average precision (AP) metrics for small-sized, medium-sized, and large-sized objects are depicted in Fig. 3, and SOF-DETR shows an overall better performance than others.

Fig. 4 depicts a performance comparison between DETR and SOF-DETR on top-10 and last-10 object classes. The number of objects present in the training set for the particular class provides top-10 and last-10 object classes. This classwise-metrics comparison illustrates that even though our model is trained with a significant difference among the number of objects for different object classes, it performs adequately for top object classes as well as for low object classes. SOF-DETR illustrates better AP metrics for these selected object classes

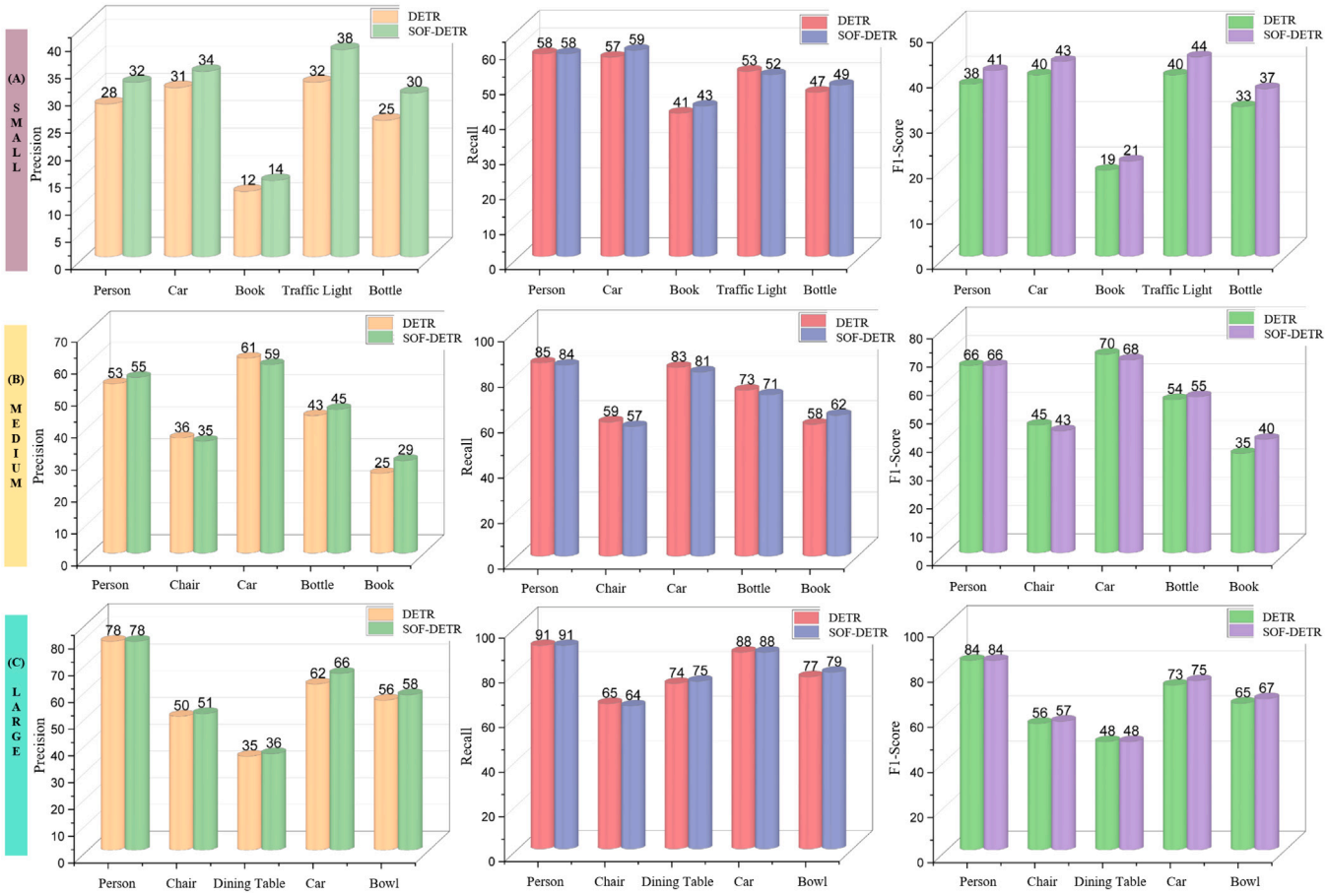

Fig. 6. Precision, Recall, and F1-Values comparison between DETR and SOF-DETR on top-5 object classes for small-sized, medium-sized, and large-sized MS COCO val set objects. All values are given on a scale of 0%–100%.

than DETR. For example, in the case of top-1 object class "Person", SOF-DETR gives AP metrics improvements of 1.4%, 0.7%, and 0.4% for small-sized, medium-sized, and large-sized objects, respectively. Similarly, for classes with fewer examples, like "Microwave", SOF-DETR shows AP metrics improvements of 4.2%, 2.6%, and 3.2% for small-sized, medium-sized, and large-sized objects, respectively.

The results indicate that normalized inductive bias in SOF-DETR improves the detection of small-sized objects without affecting the medium-sized and large-sized objects detection. To emphasize the efficacy of our proposed model in small-sized object detection, we have also presented the quantitative analysis of SOF-DETR using AP@0.5 metric and PR curve, as shown in Figs. 5 and 6, respectively. In the case of class imbalance, PR-Curve presents a more reliable evaluation comparison. These curves display the correlation between precision (positive predictive value) and recall (sensitivity) values for each attainable cut-off value. A curve higher than the other curve exhibits better performance, and Fig. 5 indicates that SOF-DETR outperforms other algorithms in most cases. Also, Fig. 6 exhibits model cogency using precision, recall, and F-1 values on top-5 object classes of small-, medium- and large-sized objects of MS COCO val set.

Table 3 presents mAP scores of SOF-DTER and DETR when tested on Udacity Self Driving Car dataset. Both models use weights trained on MS COCO dataset. SOF-DETR outperforms DETR with 2.3%, 2.0% and 1.6% AP improvements in detection results of small-sized, medium-sized, and large-sized objects, respectively, and an overall 2.9% improvement in the score. It indicates that training our model on this dataset can certainly improve these results.

3.3.2. Ablation study

The quantitative results of an ablation study of the proposed technique are shown in Table 4. First two rows depict mAP scores using different backbones, ResNet-18 and ResNet-34, for the same model, SOF-DETR. The scores advocate the choice of ResNet-50 for the proposed model. Moreover, fusing the features from the last three layers (layers 4, 3, and 2) degrades the performance, suggesting that the fusion of earlier layers features with the high-level features does not positively impact the proposed model. Table 4 also demonstrates that the model performance with normalized inductive biases surpasses the model performance without utilizing any fusion technique in the detector. Additionally, the element-wise fusion technique compared to the concatenated-fusion technique shows more promising results. Later, we have also tested the importance of using GN for the hard-inductive biases. SOF-DETR without normalization and with different normalization techniques such as batch, instance, layer, and group (8 and 32) normalization techniques, along with element-wise fusing of the last two layers features, show performance degradation when compared with the proposed GN (16 groups) technique. Overall, in the last row, the proposed combination of SOF-DETR demonstrates the significance of using GN, element-wise fusion, and fusion layers with better quantitative results on evaluation metrics.

3.3.3. Qualitative analysis

We have shown qualitative results of the proposed technique alongside DETR results in Fig. 7, and Fig. 8 for a comparative study. It is evident in Fig. 7 that SOF-DETR detects small-sized objects that DETR misses. Furthermore, SOF-DETR produces a higher confidence score for other small-sized objects than the confidence score of DETR. Fig. 7 also

Fig. 7. Qualitative analysis of SOF-DETR, where we have several pairs of images with the first and second images illustrating detections from DETR and SOF-DETR, respectively. Additional detections through SOF-DETR are represented with a pink color class box. Detections only with more than 90% confidence value are displayed.

signifies SOF-DETR's higher confidence in predicting the small-sized objects without affecting the performance of medium-sized and large-sized objects. Another potential trace of the perceptible normalized inductive bias introduced in this study is depicted in Fig. 8, which compares self-attention activation of SOF-DETR with self-attention maps of DETR. It is noticeable that SOF-DETR attention maps are more vibrant for small-sized objects, including medium-sized and large-sized objects.

4. Conclusion

This study improves hard inductive bias of DETR for small-sized object detection without affecting the performance of medium-sized and large-sized objects. A normalized inductive bias is introduced using a lazy fusion of feature maps before passing it to the transformer layers of our proposed technique, SOF-DETR. The proposed technique shows higher confidence scores for detected small-sized objects and overall better performance than DETR. Future studies can explore another direction of improving inductive biases for small objects, like introducing a penalty in loss function for object sizes. This can be introduced in parallel to the normalized inductive biases. Future work can also include the study of such inductive biases in segmentation, panoptic segmentation, and instance segmentation.

CRediT authorship contribution statement

Shikha Dubey: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Visualization. **Farrukh Olimov:** Methodology, Data curation, Writing – original draft, Validation, Visualization. **Muhammad Aasim Rafique:** Visualization, Validation, Investigation, Data curation, Writing – reviewing and editing. **Moongu Jeon:** Resources, Supervision, Project administration, Funding acquisition, Writing – reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data associated with this research are available in publicly accessible repositories. These datasets can be found in the locations: (1) MS COCO Dataset: https://cocodataset.org/ (2) Udacity Self Driving Car Dataset: https://github.com/udacity/self-driving-car/tree/master/annotations , https://public.roboflow.com/object-detection/self-driving-car/3.

The code is publicly available on https://github.com/shikha-gist/SOF-DETR/.

Acknowledgments

This work was financially supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) of Ministry of Science and ICT (MSIT) (No. 2014-3-00077, AI National Strategy Project), and by the Korea Creative Content Agency (KOCCA) of Ministry of Culture, Sports, and Tourism (MCST) (No. R2020060002 and R2022060001, the Culture Technology Research Development Program).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jvcir.2022.103620.

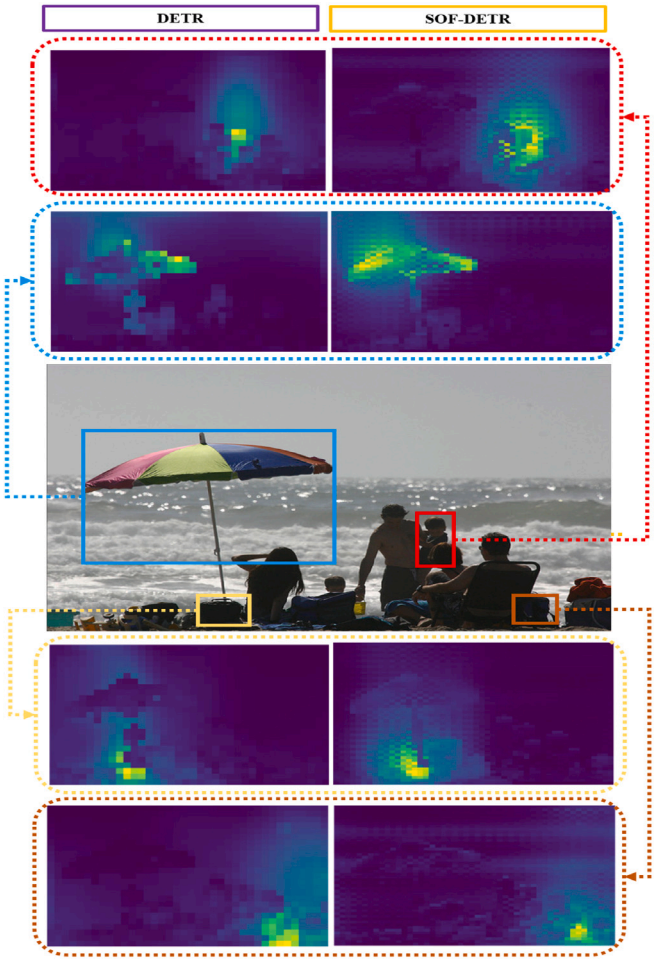

Fig. 8. Effectiveness of normalized inductive bias on objects' self-attention maps of the small-, medium-, and large-sized objects.

References

[1] R. Girshick, Fast R-CNN, in: IEEE International Conference on Computer Vision, ICCV, 2015, pp. 1440–1448.

[2] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) (2015) 1904–1916.

[3] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017.

[4] S. Ren, K. He, R.B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) (2015) 1137–1149.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: single shot MultiBox detector, in: European Conference on Computer Vision, ECCV, 2016, pp. 21–37.

[6] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 779–788.

[7] B. Wu, F. Iandola, P. Jin, K. Keutzer, SqueezeDet: unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving, in: IEEE Conference on Computer Vision and Pattern Recognition Workshop, CVPRW, 2017, pp. 446–454.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Vol. 30, NIPS, 2017.

[9] S. Dubey, F. Olimov, M.A. Rafique, J. Kim, M. Jeon, Label-attention transformer with geometrically coherent objects for image captioning, 2021, http://arxiv.org/abs/2109.07799, ArXiv.

[10] N. Chen, S. Watanabe, J. Villalba, P. Żelasko, N. Dehak, Non-autoregressive transformer for speech recognition, IEEE Signal Process. Lett. 28 (2021) 121–125.

[11] A. Khan, B. Lee, Gene transformer: transformers for the gene expression-based classification of lung cancer subtypes, 2021, http://arxiv.org/abs/2108.11833, ArXiv.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, in: International Conference on Learning Representations, ICLR, 2021.

[13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, ECCV, 2020, pp. 213–229.

[14] G. Zhang, H.-C. Wong, S.-L. Lo, Multi-attention network for unsupervised video object segmentation, IEEE Signal Process. Lett. 28 (2021) 71–75.

[15] B. Neyshabur, Towards learning convolutions from scratch, in: Advances in Neural Information Processing Systems, Vol. 33, NIPS, 2020.

[16] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2015, http://arxiv.org/abs/1409.0473, ArXiv.

[17] S. D'Ascoli, H. Touvron, M.L. Leavitt, A.S. Morcos, G. Biroli, L. Sagun, ConViT: improving vision transformers with soft convolutional inductive biases, in: Internation Conference on Machine Learning, ICML, 2021, pp. 2286–2296.

[18] Y. Wu, K. He, Group normalization, in: European Conference on Computer Vision, ECCV, 2018.

[19] S. Fang, X. Pan, S. Xiang, C. Pan, Meta-MSNet: meta-learning based multi-source data fusion for traffic flow prediction, IEEE Signal Process. Lett. 28 (2021) 6–10.

[20] R. Stewart, M. Andriluka, End-to-end people detection in crowded scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015.

[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015.

[22] J. Redmon, A. Farhadi, YOLOv3: an incremental improvement, 2018, http://arxiv.org/abs/1804.02767, ArXiv.

[23] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: IEEE International Conference on Computer Vision, ICCV, 2017, pp. 2980–2988.

[24] H. Law, J. Deng, CornerNet: detecting objects as paired keypoints, Int. J. Comput. Vis. 128 (2020) 642–656.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: IEEE International Conference on Computer Vision, ICCV, 2017, pp. 2999–3007.

[26] S. Zhang, L. Wen, X. Bian, Z. Lei, S. Li, Single-shot refinement neural network for object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 4203–4212.

[27] S. Huang, Q. Liu, Addressing scale imbalance for small object detection with dense detector, Neurocomputing 473 (2022) 68–78.

[28] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, Image transformer, in: International Conference on Machine Learning, ICML, 2018.

[29] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 6517–6525.

[30] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, A.C. Berg, DSSD: deconvolutional single shot detector, 2017, http://arxiv.org/abs/1701.06659, ArXiv.

[31] X. Zhou, J. Zhuo, P. Krähenbühl, Bottom-up object detection by grouping extreme and center points, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 850–859.

[32] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, J. Shi, FoveaBox: beyound anchor-based object detection, IEEE Trans. Image Process. 29 (2020) 7389–7398.

[33] C. Zhu, Y. He, M. Savvides, Feature selective anchor-free module for single-shot object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019.

[34] Z. Tian, C. Shen, H. Chen, T. He, FCOS: fully convolutional one-stage object detection, in: IEEE International Conference on Computer Vision, ICCV, 2019.

[35] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, D. Lin, Libra r-cnn: Towards balanced learning for object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 821–830.

[36] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: a metric and a loss for bounding box regression, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019.

[37] T. Lin, M. Maire, S.J. Belongie, L.D. Bourdev, R.B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: European Conference on Computer Vision, ECCV, 2014.

[38] Udacity self-driving car driving data, 2017, https://github.com/udacity/self-driving-car/tree/master/annotations.

[39] Roboflow: udacity self-driving car driving data, 2020, https://public.roboflow.com/object-detection/self-driving-car/3.

[40] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: AISTATS, 2010.

[41] Y. Li, H. Qi, J. Dai, X. Ji, Y. Wei, Fully convolutional instance-aware semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017.