SpatialFormer: Towards Generalizable Vision Transformers with Explicit Spatial Understanding

Han Xiao[1,2*], Wenzhao Zheng[1,3*],
Sicheng Zuo[1], Peng Gao[2], Jie Zhou[1], and Jiwen Lu[1†]

[1]Tsinghua University [2]Shanghai AI Laboratory [3]UC Berkeley
https://wzzheng.net/SpatialFormer
xiaohan@pjlab.org.cn; wenzhao.zheng@outlook.com;
zsc23@mails.tsinghua.edu.cn;
gaopeng@pjlab.org.cn; {jzhou,lujiwen}@tsinghua.edu.cn

Abstract. Vision transformers have demonstrated promising results and become core components in many tasks. Most existing works focus on context feature extraction and incorporate spatial information through additional positional embedding. However, they only consider the local positional information within each image token and cannot effectively model the global spatial relations of the underlying scene. To address this challenge, we propose an efficient vision transformer architecture, SpatialFormer, with explicit spatial understanding for generalizable image representation learning. Specifically, we accompany the image tokens with adaptive spatial tokens to represent the context and spatial information respectively. We initialize the spatial tokens with positional encoding to introduce general spatial priors and augment them with learnable embeddings to model adaptive spatial information. For better generalization, we employ a decoder-only overall architecture and propose a bilateral cross-attention block for efficient interactions between context and spatial tokens. SpatialFormer learns transferable image representations with explicit scene understanding, where the output spatial tokens can further serve as enhanced initial queries for task-specific decoders for better adaptations to downstream tasks. Extensive experiments on image classification, semantic segmentation, and 2D/3D object detection tasks demonstrate the efficiency and transferability of the proposed SpatialFormer architecture. Code is available at https://github.com/Euphoria16/SpatialFormer.

1 Introduction

Feature extraction is at the core of computer vision and has been dominated by convolutional neural networks (CNNs) since the deep learning era. The monopoly has been broken by vision transformers (ViTs), which patchify each image into a sequence of tokens and then process them using alternating self-attention and

* Equal contributions.
† Corresponding author.

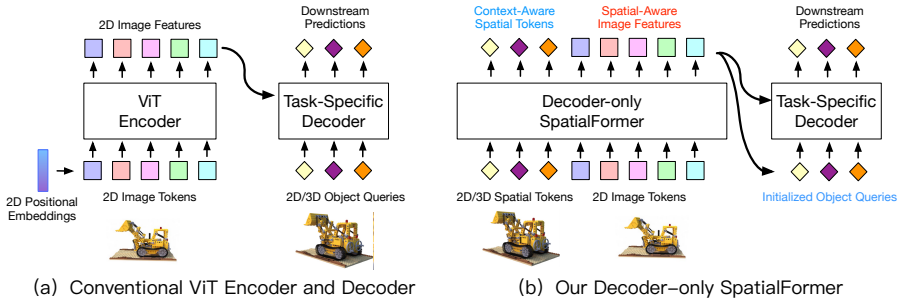

Fig. 1: Motivation of our SpatialFormer. (a) Conventional ViT encoder has limitations in effectively modeling spatial scenes, relying on task-specific decoders to introduce spatial priors for downstream tasks. (b) Our SpatialFormer learns spatial-aware image representations through the incorporation of adaptive spatial tokens, which can serve as enhanced initial queries for task-specific decoders.

multilayer perception (MLP) operations. ViTs are capable of modeling long-range global dependencies among tokens with great flexibility and thus empower the state-of-the-art performance for various tasks, such as image classification [30,42,43,51], object detection [4,66], and semantic segmentation [35,47,60].

Different from CNNs which implicitly encourage translation invariance by using local receptive fields, ViTs introduce less inductive bias by leveraging the self-attention mechanism to incorporate long-range relations between patch tokens. Recent methods improve the original ViT by designing various token interaction modules with better efficiency [9, 13, 30, 53], yet most of them still focus on modeling the interactions between the image tokens, with little efforts on the spatial information incorporation. While existing methods employ the positional embeddings added directly to each token or incorporated into attention layers [15, 30, 43], these additional positional embeddings are designed primarily to inject the knowledge of local position to each token. This design constrains the modeling of the broader spatial scene, posing challenges in generalization to downstream tasks. For instance, on tasks demanding fine-tuning on high-resolution images, most ViTs require the upsampling of positional embeddings through bicubic interpolation, resulting in significant information loss. Furthermore, when adapting to 3D tasks, the learned features only capture 2D context information from the images, lacking the perceptual capacity for 3D scenes [62]. Although introducing spatial queries in task-specific decoders [5, 25, 46] can partially alleviate this issue, the image encoder still falls short in learning meaningful image features. This hampers task performance due to the restricted perception of the spatial scene. In contrast, we primarily focus on designing an efficient image backbone capable of directly learning spatial-aware image features that facilitate transferability to downstream tasks.

In this paper, we propose SpatialFormer, a spatial-aware vision transformer model as an effective solution, as shown in Figure 1. In addition to the image tokens, we incorporate a set of spatial tokens to capture the spatial scene

information. We initialize each spatial token with positional encoding to introduce general spatial prior and further incorporate learnable embeddings to encode adaptive spatial information. These spatial tokens provide essential spatial prior to guide the spatial-aware context feature learning from the images. Using the spatial tokens as queries, we perform cross-attention to transfer context information to obtain enhanced spatial representation. We employ a decoder-only architecture for more generalizable image backbones and propose bilateral cross-attention to enable information flow between context and spatial tokens. Notably, the adaptable representation of spatial tokens enables seamless transfer to images of varying sizes or those augmented with 3D information. The updated spatial tokens can also serve as enhanced initial queries for task-specific adaptations, facilitating the transfer of knowledge acquired from image pre-training to downstream tasks. Extensive experiments on ImageNet-1K [12] for image classification, MS-COCO [26] for 2D object detection, ADE20K [64] for 2D semantic segmentation, and nuScenes [2] for 3D object detection verify the effectiveness, efficiency, and generalization ability of the proposed SpatialFormer architecture.

2 Related Work

Vision Transformer. Since Vaswani et al. [41] first proposed the transformer architecture for machine translation, it has conquered almost all NLP tasks and become the new state-of-the-art model in the language domain. Inspired by its great success, Dosovitskiy et al. [15] introduced the transformer to computer vision by partitioning an image into a sequence of non-overlapping patches as input. Further works then designed more efficient self-attention modules and adopted a multi-stage pyramid architecture to gradually process tokens of lower resolutions [9, 13, 30, 32, 53]. Despite many efforts being expended on processing inputs of different formulations, there has been limited exploration into the integration of spatial information. Most existing ViTs resort to additional positional embedding to inject the knowledge of local position into each image token. Absolute positional embeddings are first introduced to be added into each token after the image patchfication step [15, 41]. Relative positional embeddings are further proposed to directly incorporate relative positional information into the attention matrix [16, 28, 30]. However, the acquired positional bias proves insufficient in modeling the broader spatial scene, thereby posing challenges in generalizing to downstream tasks with varying image resolutions and 3D spatial information. Differently, the proposed SpatialFormer adopts a decoder-only architecture to efficiently perform interactions between image tokens and spatial tokens to obtain spatial-aware image representations.

Learning Spatial Information from Images. Learning spatial information from images is a long-standing problem in computer vision, particularly crucial for 3D perception tasks. Equipped with large pretrained image models, vision-based methods have achieved comparable performance to LiDAR-based methods in various 3D perception tasks such as 3D object detection [25,29,55,57] and 3D occupancy prediction [20,22,38,48,59,61,68]. With cameras being much

cheaper than 3D sensors like LiDAR, vision-based scene perception has attracted intense attention and is increasingly important in real-world applications including autonomous driving [25, 57, 58, 63]. Existing works follow a conventional paradigm, which first uses image backbones to extract 2D image features and then incorporate the spatial information to lift the 2D feature into the 3D space. DETR3D [46] employs 3D object queries and performs cross-attention to iteratively refine them to decode 3D object information. BEVFormer [25] and TPVFormer [21] further improved it using deformable cross-attention to obtain bird's eye view and tri-perspective view representations, respectively, to model the 3D scene. Different from vision-based scene perception, our objective is to design an efficient image backbone that directly learns spatial-aware image features encoding the 3D scene information when adapting to 3D perception tasks.

3 Proposed Approach

3.1 Learning Spatial-aware Image Representation

Vision transformers (ViTs) process image patches as input, capturing their long-range relationships within the context feature space. While most ViTs incorporate additional positional embeddings to impart local positional awareness to each image token, this approach only captures limited positional information of the input 2D pixels during pre-training. As depicted in Fig. 1, it falls short of effectively modeling the spatial information inherent in the underlying scene, relying on task-specific decoders to introduce spatial priors for downstream tasks.

To address this, we propose SpatialFormer, an efficient architecture with adaptive spatial tokens to directly model the underlying 2D/3D scene from images. We leverage a set of adaptive spatial tokens to represent the spatial scene and decode the spatial information through interactions between spatial tokens and image tokens. To facilitate efficient interactions, we modify conventional vision transformer decoder blocks into bilateral cross-attention (BCA) blocks, as depicted in Fig. 2. By introducing this novel design of adaptive spatial tokens and a decoder-only architecture with BCA blocks, we enhance the ability to learn generalizable representations of the 2D/3D underlying scene from images.

Adaptive Spatial Tokens. To incorporate the 2D/3D spatial scene information, we introduce a set of adaptive spatial tokens capable of generalizing to various types of input images. Specifically, for a given input image, we initiate the position of spatial tokens by sampling a grid of $N \times N$ spatial points, denoted as $P = \{p_i \in \mathcal{R}^m, i = 1, 2, ..., N^2\}$.

For a common single-view image, we simply employ the pixel coordinate:

$$p_i = (u, v)^T \in \mathcal{R}^2. \tag{1}$$

For multi-view images, we generate 3D coordinates $p_i = (x, y, z)^T \in \mathcal{R}^3$ within the real-world scene. The process involves lifting the 2D pixel to a 3D spatial point by introducing a depth dimension orthogonal to the image plane. By discretizing the camera frustum space and sampling candidate depth values along

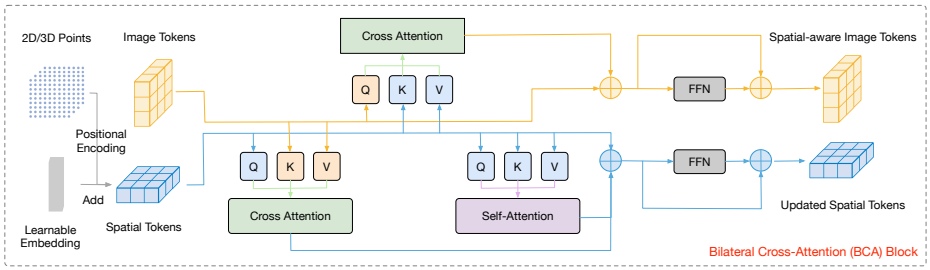

Fig. 2: Illustration of the bilateral cross-attention (BCA) block. We simultaneously perform image-to-spatial and spatial-to-image cross-attention, along with self-attention among spatial tokens in parallel to achieve low latency.

the axis, we obtain corresponding 3D points as $\{(u \times d_j, v \times d_j, d_j)^T\}_{j=1}^{D}$, where $\{d_j\}_{j=1}^{D}$ represents the set of candidate depth values. Subsequently, these pixel coordinates are transformed into 3D world coordinates using the corresponding camera intrinsic matrix K and extrinsic matrix E:

$$p_i = (x, y, z)^T = EK\{(u \times d_j, v \times d_j, d_j)^T\}_{j=1}^{D}. \tag{2}$$

We then employ the mapping function γ to transform the points coordinates into high-dimensional fourier features $\gamma(p) \in \mathcal{R}^{N^2 \times C}$ following [37]:

$$\gamma(P) = \{[\cos(2\pi \mathcal{S} p_i'), \sin(2\pi \mathcal{S} p_i')]^T, i = 1, 2, ..., N^2\}, \tag{3}$$

where $p_i' \in \mathcal{R}^m$ denotes the normalized coordinate in the $[-1, 1]$ of the spatial points (with $m = 2$ for 2D pixels and $m = 3$ for 3D points, respectively). Each entry in $\mathcal{S} \in \mathcal{R}^{C/2 \times m}$ is sampled from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$.

Subsequently, we utilize the obtained positional encodings, summed with learnable embeddings, to initialize the corresponding spatial tokens:

$$\mathbf{Z_S^0} = \gamma(P) + \mathbf{Q_p}, \tag{4}$$

where $\mathbf{Q_p}$ is the learnable embedding designed to enhance the representation capacity of the spatial tokens. By adopting the positional encoding format, our spatial tokens can effectively represent both 2D and 3D spatial scenes. This facilitates a versatile interaction with the image tokens in the subsequent blocks.

Bilateral Cross-Attention Block. To enable the efficient interaction between the spatial tokens and image tokens, we customize the conventional vision transformer decoder layers into Bilateral Cross-Attention (BCA) blocks. In the proposed BCA block, spatial tokens are first enhanced through cross-attention with the context feature from the image tokens. This mechanism enables the guidance of spatial tokens by context features, facilitating the acquisition of meaningful representations. We then perform self-attention among the spatial tokens to capture the spatial clues within the underlying spatial scene, followed by LayerNorm (LN) and feed-forward networks to generate the updated spatial

tokens. For the l-th decoder block, we update the spatial scene representation by the spatial tokens $\mathbf{Z_S}$ using both cross-attention and self-attention:

$$\begin{aligned}\hat{\mathbf{Z}}_{\mathbf{S}}^{\mathbf{l}} &= \mathbf{CA}(\mathbf{Q}_{\mathbf{S}}^{\mathbf{l}}, \mathbf{K}_{\mathbf{I}}^{\mathbf{l}}, \mathbf{V}_{\mathbf{I}}^{\mathbf{l}}) + \mathbf{Z}_{\mathbf{S}}^{\mathbf{l}}, \\ \hat{\mathbf{Z}}_{\mathbf{S}}^{\mathbf{l}\,\prime} &= \mathbf{SA}(\mathbf{Q}_{\mathbf{S}}^{\mathbf{l}}, \mathbf{K}_{\mathbf{S}}^{\mathbf{l}}, \mathbf{V}_{\mathbf{S}}^{\mathbf{l}}) + \hat{\mathbf{Z}}_{\mathbf{S}}^{\mathbf{l}}, \\ \mathbf{Z}_{\mathbf{S}}^{\mathbf{l+1}} &= \mathbf{FFN}(\mathbf{LN}(\hat{\mathbf{Z}}_{\mathbf{S}}^{\mathbf{l}\,\prime})) + \hat{\mathbf{Z}}_{\mathbf{S}}^{\mathbf{l}\,\prime},\end{aligned} \tag{5}$$

where $\mathbf{K}_{\mathbf{I}}^{\mathbf{l}}, \mathbf{V}_{\mathbf{I}}^{\mathbf{l}}, \mathbf{K}_{\mathbf{S}}^{\mathbf{l}}, \mathbf{V}_{\mathbf{S}}^{\mathbf{l}}$ denote the keys and values obtained from the image tokens $\mathbf{Z}_I^l$ and input spatial tokens $\mathbf{Z}_S^l$.

The updated spatial tokens encode the spatial prior essential for the context feature extraction. Therefore, we introduce another cross-attention that utilizes the spatial prior provided by the spatial tokens to update the image tokens. Specifically, given the input image tokens $\mathbf{Z}_I^l$ in the l-th decoder block, we update them through cross-attention with the spatial tokens:

$$\hat{\mathbf{Z}}_{\mathbf{I}}^{\mathbf{l}} = \mathbf{CA}(\mathbf{Q}_{\mathbf{I}}^{\mathbf{l}}, \mathbf{K}_{\mathbf{S}}^{\mathbf{l}}, \mathbf{V}_{\mathbf{S}}^{\mathbf{l}}) + \mathbf{Z}_{\mathbf{I}}^{\mathbf{l}}, \quad \mathbf{Z}_{\mathbf{I}}^{\mathbf{l+1}} = \mathbf{FFN}(\mathbf{LN}(\hat{\mathbf{Z}}_{\mathbf{I}}^{\mathbf{l}})) + \hat{\mathbf{Z}}_{\mathbf{I}}^{\mathbf{l}}. \tag{6}$$

By introducing the learnable spatial tokens, our approach can produce spatial-aware features from images in an efficient manner. This approach enables us to obtain comprehensive scene representations directly from images, facilitating the downstream 2D and 3D visual perception task. The updated spatial tokens can function as initial queries for task-specific decoders, transferring the spatial perception capacity acquired from image pre-training to downstream tasks. Further details about task-specific adaptations are elaborated in Section 3.3.

3.2 SpatialFormer

We propose a decoder-only vision backbone, SpatialFormer, which leverages the Bilateral Cross-Attention (BCA) block as the fundamental building block for learning spatial-aware image representation efficiently, as shown in Fig. 3. Inspired by the recent state-of-the-art architectures [9, 30], we design our models with a four-stage pyramid structure, ensuring that the tiny, small, and base models have comparable parameters and FLOPs to existing ones. We provide the detailed architectures of our SpatialFormer in the supplementary.

In the initial stage, we employ non-overlapping convolutional layers to extract patch embeddings from the input image. In the subsequent stages, we utilize the patch merging module to reduce spatial resolution while simultaneously increasing the dimensions of image features. As for the spatial tokens, the number can be adjusted flexibly according to the target tasks and datasets regardless of the input resolutions. Within the first two stages, we stack the proposed BCA block, which facilitates both self-attention and cross-attention to update the spatial tokens, while only applying cross-attention to the image tokens. In the third and fourth stages, we introduce additional self-attention to the image tokens. We implement the Self-Attention (SA) block, where we concatenate the spatial tokens and image tokens and apply self-attention to them.

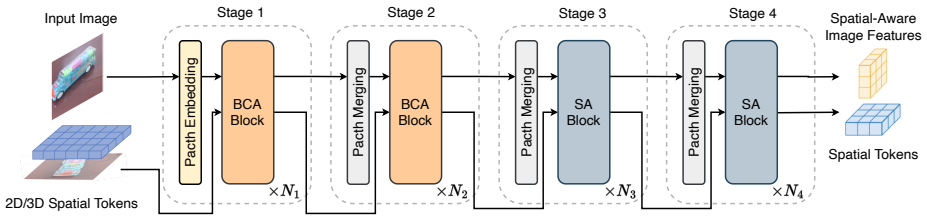

Fig. 3: Illustration of our SpatialFormer architecture. We adopt a four-stage pyramid structure and employ the bilateral attention block in the first two stages. With fewer image tokens in the latter stages, we add self-attention between image tokens to improve the performance without introducing too much computation load.

Our SpatialFormer achieves a favorable trade-off between accuracy and computational complexity with the BCA block and SA block. This allows for effective information exchange and integration between the 3D scene tokens and 2D image representations, enhancing the overall understanding of the 3D scene.

3.3 Adaptation to Downstream Tasks

In contrast to traditional vision transformer backbones, our SpatialFormer demonstrates remarkable adaptability for downstream tasks by incorporating spatial tokens to integrate spatial scene information. For instance, in the context of 3D perception tasks, we seamlessly integrate 3D scene knowledge into the backbone model by utilizing 3D positional encodings. Furthermore, the spatial tokens updated by our decoder-only architecture can be directly utilized by task-specific heads as initial queries to acquire dense predictions, enhancing spatial understanding capacities for tasks such as detection and segmentation.

Image Classification. We employ a classification head consisting of a fully connected layer to process the output tokens. The default token number is set to 8×8 to balance accuracy and computational complexity. Adjusting the number of spatial tokens allows for tailoring to specific needs, enabling improvements in accuracy for higher-resolution fine-tuning or reducing computational complexity.

2D Detection and Segmentation. Our SpatialFormer is a general vision backbone compatible with widely used frameworks like Mask-RCNN and also generalizes to multi-scale training by reconfiguring the initial positions of spatial tokens. Moreover, the generated spatial tokens can serve as initial object queries, feeding into multiple transformer decoder layers, thereby directly contributing to the acquisition of dense prediction results.

3D Perception. Traditional vision transformers are typically pretrained on 2D single-view images, resulting in feature extraction that lacks 3D knowledge. In contrast, our SpatialFormer processes multi-view images and encodes 3D scene representation directly by utilizing camera parameters to derive 3D positional encodings, initializing the spatial tokens for enhanced 3D perception capabilities.

4 Experiments

In this section, we conduct extensive experiments to evaluate our SpatialFormer. We provide more dataset and implementation details in the supplementary.

4.1 ImageNet Classification

Experimental Settings. We first evaluate our SpatialFormer on ImageNet [34], which is a widely used benchmark for image classification. ImageNet consists of around 1.2M training images and 50K validation images from 1K different categories. For a fair comparison, we follow the same training recipe as the DeiT model [39]. Specifically, we train the models from scratch for 300 epochs with an input size of 224×224. We use the default data augmentation and regularization strategy. Additionally, we finetune the SpatialFormer backbones at a resolution of 384×384 for 30 epochs following previous settings [31].

Results. Table 1 shows that our SpatialFormer outperforms state-of-the-art methods in terms of accuracy and computation efficiency across different model sizes. Specifically, our SpatialFormer-T achieves a top-1 accuracy of 81.5%, outperforming PVTv2-b1 [44] and Shunted-T [33] by 2.8% and 1.7%, respectively. Our small model SpatialFormer-S achieves a competitive result of 83.8% with only 4.8G FLOPs, demonstrating the best trade-offs between computational cost and accuracy among models of similar size. Our largest model, SpatialFormer-B, achieves 84.5% top-1 accuracy with only 50M parameters and 9.8G FLOPs, surpassing models with much higher FLOPs, such as Swin-B [30], ConvNeXt-B [31], and CSWin-B [13]. Furthermore, our SpatialFormer shows promising performance potential when fine-tuned at higher resolution. The performance of SpatialFormer-S and SpatialFormer-B can be further boosted to 84.7% and 85.3%, respectively, outperforming existing architectures with fewer FLOPs.

4.2 Object Detection and Instance Segmentation

Experimental Settings. To evaluate SpatialFormer on downstream dense prediction tasks, we conduct experiments on object detection and instance segmentation using the COCO 2017 dataset [26]. We use the SpatialFormer-S and SpatialFormer-B models pretrained on ImageNet as the backbones using Mask-RCNN with a 1x training schedule and Cascade Mask-RCNN [3] with a 3x training schedule and multi-scale training for further validation. Moreover, our SpatialFormer can utilize spatial tokens as initial object queries to build a transformer decoder-based detection framework. We duplicate the 8×8 spatial tokens, select the top 150 queries, and stack 6 deformable DETR decoder layers to obtain detection results. The model is trained using a 1x training schedule.

Results. We report the performance on COCO object detection and instance segmentation in Table 2. Our SpatialFormer outperforms all the other vision backbones under different model sizes. For Mask-RCNN, SpatialFormer-S outperforms UniFormer-S by 2.2 while SpatialFormer-B outperforms UniFormer-B

Table 1: Comparisons on ImageNet classification. We compare the parameters, FLOPs, and accuracy of our SpatialFormer with other state-of-the-art architectures.

Model	Image Size	Params	FLOPs	Top-1 Acc (%)	Top-5 Acc(%)
ResNet-18 [19]	224^2	11.7M	1.8G	69.8	89.1
DeiT-T [39]	224^2	5.7M	1.6G	72.2	91.3
PVT-T [43]	224^2	13.2M	1.9G	75.1	92.4
PVTv2-b1 [44]	224^2	13.1M	3.1G	78.7	-
Shunted-T [33]	224^2	11.5M	2.1G	79.8	-
SpatialFormer-T	224^2	11.6M	2.4G	**81.5**	**95.8**
ResNet-50 [19]	224^2	26M	4.1G	78.3	94.3
DeiT-S [39]	224^2	22M	4.6G	79.9	95.0
Swin-T [30]	224^2	29M	4.5G	81.2	95.5
PVT-S [43]	224^2	25M	3.8G	79.8	95.0
CrossFormer-S [45]	224^2	31M	4.9G	82.5	-
RegionViT-S [6]	224^2	31M	5.3G	82.6	96.1
CSWin-T [13]	224^2	23M	4.3G	82.7	-
UniFormer-S [23]	224^2	24M	4.2G	82.9	96.2
PaCa-Small [17]	224^2	21M	5.4G	83.2	-
CMT-S [18]	224^2	25M	4.0G	83.5	96.6
SpatialFormer-S	224^2	25M	4.8G	**83.8**	96.4
CvT-13 [49]	384^2	20M	16.3G	83.0	-
CaiT-XS24 [40]	384^2	27M	19.3G	84.1	96.9
CSWin-T [13]	384^2	23M	14.0G	84.3	-
SpatialFormer-S	384^2	25M	12.8G	**84.7**	**97.1**
ResNet-152 [19]	224^2	60M	11.6G	81.3	95.5
DeiT-B [39]	224^2	86 M	17.5G	81.8	95.6
Swin-B [30]	224^2	88M	15.4G	83.5	96.5
ConvNeXt-B [31]	224^2	89M	15.4G	83.8	-
CSWin-B [14]	224^2	78M	15.0G	84.2	-
DAT-B [50]	224^2	88M	15.8G	84.0	96.7
UniFormer-B [23]	224^2	50M	8.3G	83.9	96.7
RegionViT-B [6]	224^2	74M	13.6G	83.8	96.1
QFormer-B [56]	224^2	90M	15.7G	84.1	96.8
SpatialFormer-B	224^2	50M	9.8G	**84.5**	**96.8**
Swin-B [30]	384^2	88M	47.0G	84.5	97.0
DAT-B [50]	384^2	88M	49.8G	84.8	97.0
CaiT-S24 [40]	384^2	47M	32.2G	85.1	97.4
SpatialFormer-B	384^2	50M	23.5G	**85.3**	**97.5**

by 1.8 in terms of AP^b, verifying the advantages of spatial-aware image representation learning. For Cascade Mask-RCNN, SpatialFormer surpasses models with much higher parameters and FLOPs such as Swin-S and DAT-S. Moreover, we observe a significant performance improvement (1.2 of AP^b) for SpatialFormer-S compared to the most competitive transformer decoder-based frameworks. No-

Table 2: Results on COCO for object detection and instance segmentation. The FLOPs are measured on input sizes of 1280×800.

(a) Mask-RCNN Object Detection and Instance Segmentation

Backbone	Params	FLOPs	Schedule	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
PVT-S [43]	44M	245G	1x	40.4	62.9	43.8	37.8	60.1	40.3
Swin-T [30]	48M	267G	1x	42.2	64.6	46.2	39.1	61.6	42.0
CrossFormer-S [45]	50M	301G	1x	45.4	68.0	49.7	41.4	64.8	44.6
UniFormer-S [23]	41M	269G	1x	45.6	68.1	49.7	41.6	64.8	45.0
SpatialFormer-S	42M	255G	1x	**47.8**	**70.1**	**52.5**	**42.5**	**66.6**	**45.8**
Swin-S [30]	69M	354G	1x	44.8	66.6	48.9	40.9	63.4	44.2
DAT-S [50]	69M	378G	1x	47.1	69.9	51.5	42.5	66.7	45.4
CrossFormer-B [45]	72M	408G	1x	47.2	69.9	51.8	42.7	66.6	46.2
UniFormer-B [23]	69M	399G	1x	47.4	69.7	52.1	43.1	66.0	46.5
SpatialFormer-B	59M	316G	1x	**49.2**	**71.1**	**54.4**	**43.7**	**67.7**	**47.1**

(b) Cascade Mask-RCNN Object Detection and Instance Segmentation

Backbone	Params	FLOPs	Schedule	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
Swin-T [30]	86 M	745G	3x	48.1	67.1	52.2	41.7	64.4	45.0
DAT-T [50]	86 M	750G	3x	49.1	68.2	52.9	42.5	65.4	45.8
Swin-S [30]	107 M	838G	3x	50.4	69.2	54.7	43.7	66.6	47.3
DAT-S [50]	107 M	857G	3x	51.3	70.1	55.8	44.5	67.5	48.1
SpatialFormer-S	82 M	734G	3x	**52.2**	**70.8**	**56.6**	**45.2**	**68.3**	**49.0**
Swin-S [30]	86 M	745G	3x	48.1	67.1	52.2	41.7	64.4	45.0
DAT-S [50]	86 M	750G	3x	49.1	68.2	52.9	42.5	65.4	45.8
Swin-S [30]	107 M	838G	3x	50.4	69.2	54.7	43.7	66.6	47.3
DAT-S [50]	107 M	857G	3x	51.3	70.1	55.8	44.5	67.5	48.1
SpatialFormer-B	98M	812G	3x	**52.9**	**71.357.4**		**45.9**	**68.9**	**49.9**

(c) Transformer Decoder-Based Object Detection

Framework	Backbone	Two-Stage	Epoch	AP^b	AP^b_{50}	AP^b_{75}	AP^b_S	AP^b_M	AP^b_L
DETR [5]	ResNet-50	×	12	15.5	29.4	14.5	4.3	15.1	26.7
DAB-DETR [27]	ResNet-50	×	12	38.0	60.3	39.8	19.2	40.9	55.4
Dynamic DETR [11]	ResNet-50	×	12	42.9	61.0	46.3	24.6	44.9	54.4
Deformable DETR [67]	ResNet-50	×	12	37.2	55.5	40.5	21.1	40.7	50.5
Deformable DETR [67]	ResNet-50	✓	12	43.7	62.9	47.2	26.7	46.9	57.2
Deformable DETR [67]	Swin-T	×	12	41.9	59.0	44.5	26.4	41.3	45.9
Deformable DETR [67]	Swin-T	✓	12	45.3	64.8	49.0	27.8	48.5	60.6
SpatialFormer	Ours-Small	×	12	43.5	61.8	47.0	27.3	46.3	58.2
SpatialFormer	Ours-Small	✓	12	**46.5**	**64.9**	**50.5**	**29.3**	**49.5**	**61.0**

tably, SpatialFormer surpasses Deformable DETR with ResNet50 and Swin-T in the one-stage setting (43.5% vs 37.2% vs 41.9% in terms of mAP), even without spatial tokens as initialization. This results from the ability of SpatialFormer to capture spatial scene information, providing additional information about the object details. By encoding such spatial information, our model can effectively handle challenging scenarios in object detection.

Table 3: Results on ADE20K for semantic segmentation. The FLOPs are measured at the input resolution of 512 × 2048.

(a) UpperNet Semantic Segmentation

Backbone	Params	FLOPs	mIoU	mIoU (MS)	mAcc
Swin-T [30]	60M	945G	44.4	45.8	55.6
DAT-T [50]	60M	957G	45.5	46.4	58.0
CrossFormer-S [45]	62M	980G	47.6	48.4	-
SpatialFormer-S	52M	935G	**48.1**	**49.2**	**59.4**
Swin-B [30]	121M	1188G	48.1	49.7	59.1
DAT-B [50]	121M	1212G	49.4	50.5	61.8
CrossFormer-B [45]	84M	1079G	49.2	50.1	-
SpatialFormer-B	73M	1028G	**50.3**	**51.2**	61.8

(b) Transformer Decoder-Based Semantic Segmentation

Framework	Backbone	Iterations	mIoU	mIoU (MS)	mAcc
MaskFormer [8]	Swin-T	160k	46.7	48.8	-
Mask2Former [7]	Swin-T	160k	47.7	49.6	-
kMaX-DeepLab [54]	ConvNeXt-T	100k	48.3	-	-
SpatialFormer	Ours-Small	160k	**50.0**	**52.8**	**63.1**

4.3 Semantic Segmentation

Experimental Settings. We further evaluate the performance of our SpatialFormer on the ADE20K dataset [65] for semantic segmentation. We first adopt the UpperNet [52] framework with the integration of our SpatialFormer backbone. Additionally, we employ the generated spatial tokens as initial object queries to derive segmentation masks using the Mask2Former [7] decoder layers. For all the models, we train them using the AdamW optimizer for 160k iterations, with a batch size of 16 and input image cropped to 512 × 512. We follow the settings of Swin Transformer [30] and Mask2Former [7] for fair comparison.

Results. We present results using the UpperNet semantic segmentation framework on the ADE20K [65] dataset in Table 3. Our SpatialFormer-S outperforms CrossFormer-S by 0.5 mIoU and our SpatialFormer-B can attain 2.2 higher mIoU than Swin-B, which has much higher FLOPs. Furthermore, SpatialFormer surpasses other decoder-based segmentation frameworks by incorporating spatial tokens. Unlike conventional backbones relying on context feature extraction, SpatialFormer excels in capturing fine-grained spatial information and exhibits strong transferability to transformer-based segmentation architectures.

4.4 3D Object Detection on nuScenes

Experimental Settings. To evaluate the generalization of SpatialFormer for 3D perception tasks, we applied SpatialFormer with to BEVFormer [25] to perform 3D object detection tasks on nuScenes. To verify the effectiveness of 3D scene information incorporation, we experiment with both 2D and 3D types of

Table 4: 3D object detection results on nuScenes.

Framework	Backbone	Epochs	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVDepth [24]	ResNet-50	24	0.367	0.315	0.702	0.271	0.621	1.042	0.315
BEVDepth [24]	ResNet-101	24	0.381	0.320	0.682	0.272	0.562	0.997	0.284
BEVFormer [25]	ResNet-50	24	0.354	0.252	0.900	0.294	0.655	0.657	0.216
BEVFormer [25]	Swin-T	24	0.369	0.265	0.898	0.284	0.631	0.594	0.224
SpatialFormer(2D)	Ours-Small	24	0.389	0.290	0.856	0.287	0.582	0.611	0.221
SpatialFormer(3D)	Ours-Small	24	**0.392**	0.297	0.848	0.285	0.610	0.608	0.217

Table 5: Ablation on the number of spatial tokens under various input resolutions of SpatialFormer-S.

#Token	Size	Params	FLOPs	Acc (%)
None	224^2	24M	4.0G	81.3
16	224^2	24M	4.3G	83.0
64	224^2	25M	4.8G	83.8
64	384^2	25M	12.8G	84.7
256	224^2	25M	7.0G	84.1
256	384^2	25M	15.0G	**85.0**

Table 6: Ablation on different architecture designs of patialFormer-S. BCA-BCA-SA-SA represents using the Bilateral Cross-Attention block at the first two stages and the Self-Attention block at the following two stages.

Stages	Params	FLOPs	Acc(%)
BCA-BCA-SA-SA	25M	4.8G	83.8
BCA-BCA-BCA-SA	31M	5.2G	84.0
BCA-BCA-BCA-BCA	37M	5.4G	**84.1**

positional encoding to initialize the spatial tokens. We train the models for 24 epochs, following the other hyperparameter settings of BEVFormer [25].

Results. Table 4 summarizes the detailed comparison results. Our SpatialFormer notably improves the NDS scores of baseline BEVFormer with Swin-T backbone from 0.369 to 0.389 on nuScenes val. This improvement underscores the effectiveness of the proposed spatial tokens in refining the localization accuracy of 3D object detection predictions. Furthermore, we observe our SpatialFormer can further boost the performance to 0.392 when adopting the 3D type of spatial tokens. These experimental results demonstrate the superior capabilities of our SpatialFormer for 3D scene understanding based on multi-view images.

4.5 Quantitive Analysis

Incorporation of Spatial Tokens. To investigate the impact of spatial tokens, we remove the positional encoding from the bilateral decoder layer and explore various numbers of spatial tokens across different input resolutions, as shown in Table 5. Without positional encoding, the spatial tokens devolve into random-initialized queries, lacking any spatial priors and resulting in performance drops. Decreasing the number of spatial tokens to 4×4 yields lower accuracy but reduces FLOPs. Conversely, increasing to 16×16 improves accuracy by incorporating more fine-grained spatial information, albeit at the cost of increased computational complexity. To achieve the best trade-off between performance and computation, we use 8×8 spatial tokens in our experiments as default.

Table 7: Comparisons with conventional positional embeddings/encodings.

Method	Params	FLOPs	Acc (%)
Baseline	24M	4.0G	81.3
Absolute PE [15]	24M	4.0G	81.7 (+**0.4**)
Relative PE [30]	25M	4.2G	82.0 (+**0.7**)
Conditional PE [10]	25M	4.7G	82.1 (+**0.8**)
Rotary PE [36]	24M	4.5G	81.8 (+**0.5**)
Dynamic Conv. [1]	25M	4.3G	81.7 (+**0.4**)
SpatialFormer-S	25M	4.8G	**83.8** (+**2.5**)

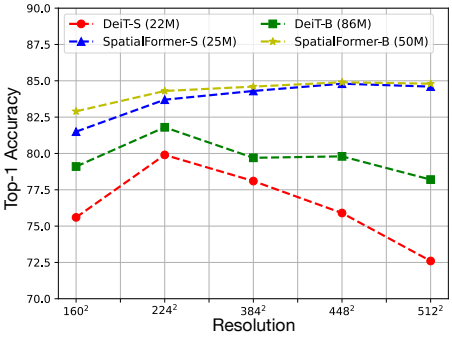

Fig. 4: Effect of different embedding sizes.

Architecture Design. We further investigate the impact of different architectural designs of our SpatialFormer. We replace the Self-Attention block with our proposed Bilateral Cross-Attention (BCA) block in the third and fourth stages. Results in Table 6 demonstrate that increasing the number of BCA blocks improves performance. This highlights the effectiveness of incorporating more BCA blocks in learning the underlying 3D scene from images. However, it is less efficient compared to the SA block in the last two stages. Hence, we choose to use BCA blocks in the first two stages and SA blocks for the remaining stages.

4.6 Qualitative Analysis

Attention maps for a certain spatial token. To better illustrate the interactions between image tokens and spatial tokens, we provide visualizations of attention maps generated from a spatial token in Fig. 5. The Cross-Attention maps showcase interactions between spatial and image tokens, emphasizing the focus on context features. Meanwhile, the Self-Attention maps illustrate dependencies among spatial tokens, revealing a concentration on location-related information. These visualizations demonstrate the ability of learned spatial tokens to discern context and location details through cross and self-attention computations.

Attention maps for multi-view images. To better understand the spatial token, we further provide visualizations of attention maps for multi-view images from nuScenes [2]. As depicted in Fig. 6, the spatial token tends to pay attention to the same object, even in different views. This shows that spatial tokens encoding 3D information establish position correlations between different views and can enhance spatial reasoning capabilities.

Comparisons with other positional embeddings. We provide comparisons with existing positional embedding/encoding (PE) methods in Table 7, including absolute positional embedding [15], relative positional embedding [30], conditional position encoding [10] , rotary position embedding [36], and position encoding generated by dynamic convolution layers [1]. We establish a baseline model without spatial tokens, using only learnable embeddings to perform bilateral cross attentions with image tokens. In this configuration, the original

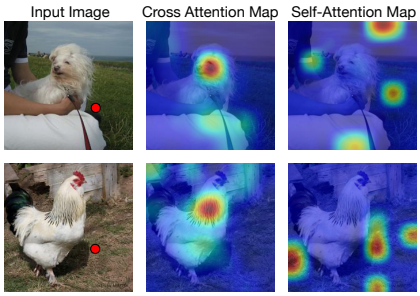

Fig. 5: Visualization of attention maps generated from a fixed spatial token.

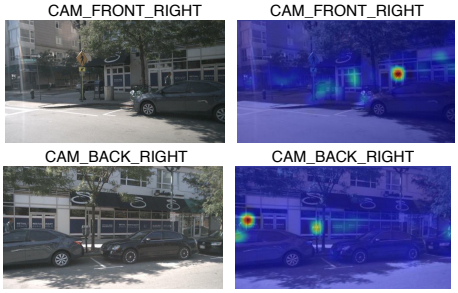

Fig. 6: Visualization of attention maps on multi-view images.

spatial tokens degenerate into image feature queries, while maintaining the primary model architectures. We see that our method outperforms existing PEs by a large margin. While some methods explore input-relevant positional information, they only add positional embedding to individual image patches at early stages. SpatialFormer enables a more spatial-aware representation by updating adaptive spatial tokens through bilateral cross-attention with image tokens.

Generalization ability without fine-tuning. To further verify the generalization ability of our model, we evaluate the transferring ability without fine-tuning. Specifically, we transfer models trained on 224^2 images directly to images of different resolutions without fine-tuning. We see that DeiT suffers significant performance drops on other resolutions while our method can directly generalize to larger image sizes without additional fine-tuning, as demonstrated in Fig. 4. This is because conventional positional embeddings defined on local image patches are inherently task-agnostic and may not effectively leverage spatial information during fine-tuning, resulting in inferior generalization. This motivates the use of adaptive spatial tokens for enhanced generalizability across various downstream tasks, as also demonstrated in Sections 4.2, 4.3, and 4.4.

5 Conclusion

In this paper, we have introduced a SpatialFormer model to enhance spatial-aware image feature learning and improve generalization to downstream tasks. By incorporating adaptive spatial tokens alongside image tokens, SpatialFormer efficiently captures spatial scene information, effectively addressing limitations in conventional transformers that rely solely on local positional embeddings. Our approach, employing a decoder-only architecture, facilitates efficient interaction between image and spatial tokens, resulting in enhanced spatial representation. The proposed model not only advances standard image classification tasks but also exhibits promising performance across diverse downstream applications, including 2D dense prediction and 3D perception tasks. We hope our work can inspire future research to develop more spatial-aware image backbones.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFB280690, and in part by the National Natural Science Foundation of China under Grant 62321005, Grant 62336004, and Grant 62125603.

References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. NeurIPS **33**, 12449–12460 (2020)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
3. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: CVPR. pp. 6154–6162 (2018)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229 (2020)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020)
6. Chen, C.F., Panda, R., Fan, Q.: Regionvit: Regional-to-local attention for vision transformers. arXiv preprint arXiv:2106.02689 (2021)
7. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR. pp. 1290–1299 (2022)
8. Cheng, B., Schwing, A.G., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: NeurIPS (2021)
9. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C.: Twins: Revisiting the design of spatial attention in vision transformers. In: NeurIPS (2021)
10. Chu, X., Tian, Z., Zhang, B., Wang, X., Shen, C.: Conditional positional encodings for vision transformers. In: ICLR (2022)
11. Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L., Zhang, L.: Dynamic detr: End-to-end object detection with dynamic attention. In: ICCV. pp. 2988–2997 (2021)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009)
13. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. arXiv preprint arXiv:2107.00652 (2021)
14. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: CVPR. pp. 12124–12134 (2022)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020)
16. Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M.: Levit: a vision transformer in convnet's clothing for faster inference. In: ICCV. pp. 12259–12269 (2021)
17. Grainger, R., Paniagua, T., Song, X., Cuntoor, N., Lee, M.W., Wu, T.: Paca-vit: learning patch-to-cluster attention in vision transformers. In: CVPR. pp. 18568–18578 (2023)

18. Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., Xu, C.: Cmt: Convolutional neural networks meet vision transformers. In: CVPR. pp. 12175–12185 (2022)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
20. Huang, Y., Zheng, W., Zhang, B., Zhou, J., Lu, J.: Selfocc: Self-supervised vision-based 3d occupancy prediction. In: CVPR (2024)
21. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Tri-perspective view for vision-based 3d semantic occupancy prediction. arXiv preprint arXiv:2302.07817 (2023)
22. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In: ECCV (2024)
23. Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unifying convolution and self-attention for visual recognition. arXiv preprint arXiv:2201.09450 (2022)
24. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. arXiv preprint arXiv:2206.10092 (2022)
25. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: ECCV (2022)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
27. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329 (2022)
28. Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B., Nadai, M.: Efficient training of visual transformers with small datasets. NeurIPS **34**, 23818–23830 (2021)
29. Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection. arXiv preprint arXiv:2203.05625 (2022)
30. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
31. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. arXiv preprint arXiv:2201.03545 (2022)
32. Lu, J., Yao, J., Zhang, J., Zhu, X., Xu, H., Gao, W., Xu, C., Xiang, T., Zhang, L.: Soft: Softmax-free transformer with linear complexity. In: NeurIPS (2021)
33. Ren, S., Zhou, D., He, S., Feng, J., Wang, X.: Shunted self-attention via multi-scale token aggregation. In: CVPR. pp. 10853–10862 (2022)
34. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
35. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: ICCV (2021)
36. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
37. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. NeurIPS **33**, 7537–7547 (2020)
38. Tong, W., Sima, C., Wang, T., Chen, L., Wu, S., Deng, H., Gu, Y., Lu, L., Luo, P., Lin, D., et al.: Scene as occupancy. In: ICCV. pp. 8406–8415 (2023)

39. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML. pp. 10347–10357 (2021)
40. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. In: ICCV. pp. 32–42 (2021)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)
42. Wang, C., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Opera: omni-supervised representation learning with hierarchical supervisions. In: ICCV. pp. 5559–5570 (2023)
43. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: ICCV (2021)
44. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022)
45. Wang, W., Yao, L., Chen, L., Lin, B., Cai, D., He, X., Liu, W.: Crossformer: A versatile vision transformer hinging on cross-scale attention. In: ICLR (2023)
46. Wang, Y., Guizilini, V., Zhang, T., Wang, Y., Zhao, H., Solomon, J.M.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: CoRL (2021)
47. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: CVPR. pp. 8741–8750 (2021)
48. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In: ICCV. pp. 21729–21740 (2023)
49. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: CVPR. pp. 22–31 (2021)
50. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Vision transformer with deformable attention. In: CVPR. pp. 4794–4803 (2022)
51. Xiao, H., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Token-label alignment for vision transformers. In: ICCV. pp. 5495–5504 (2023)
52. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: ECCV. pp. 418–434 (2018)
53. Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., Gao, J.: Focal self-attention for local-global interactions in vision transformers. arXiv preprint arXiv:2107.00641 (2021)
54. Yu, Q., Wang, H., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: k-means mask transformer. In: ECCV. pp. 288–307. Springer (2022)
55. Zeng, S., Zheng, W., Lu, J., Yan, H.: Hardness-aware scene synthesis for semi-supervised 3d object detection. TMM (2024)
56. Zhang, Q., Zhang, J., Xu, Y., Tao, D.: Vision transformer with quadrangle attention. TPAMI (2024)
57. Zhang, Y., Zheng, W., Zhu, Z., Huang, G., Zhou, J., Lu, J.: A simple baseline for multi-camera 3d object detection. arXiv preprint arXiv:2208.10035 (2022)
58. Zhang, Y., Zhu, Z., Zheng, W., Huang, J., Huang, G., Zhou, J., Lu, J.: Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. arXiv preprint arXiv:2205.09743 (2022)
59. Zhao, L., Xu, X., Wang, Z., Zhang, Y., Zhang, B., Zheng, W., Du, D., Zhou, J., Lu, J.: Lowrankocc: Tensor decomposition and low-rank recovery for vision-based 3d semantic occupancy prediction. In: CVPR. pp. 9806–9815 (2024)

60. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR. pp. 6881–6890 (2021)
61. Zheng, W., Chen, W., Huang, Y., Zhang, B., Duan, Y., Lu, J.: Occworld: Learning a 3d occupancy world model for autonomous driving. In: ECCV (2024)
62. Zheng, W., Lu, J., Jie, Z.: Structural deep metric learning for room layout estimation. In: ECCV (2020)
63. Zheng, W., Song, R., Guo, X., Chen, L.: Genad: Generative end-to-end autonomous driving. In: ECCV (2024)
64. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)
65. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. IJCV **127**, 302–321 (2019)
66. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR (2020)
67. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR (2021)
68. Zuo, S., Zheng, W., Huang, Y., Zhou, J., Lu, J.: Pointocc: Cylindrical tri-perspective view for point-based 3d semantic occupancy prediction. arXiv preprint arXiv:2308.16896 (2023)