# BAEFormer: Bi-directional and Early Interaction Transformers for Bird's Eye View Semantic Segmentation

Cong Pan[1,2,*] Yonghao He[3] Junran Peng[4] Qian Zhang[3] Wei Sui[3] Zhaoxiang Zhang[1,2,5,✉]
[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]School of Future Technology, University of Chinese Academy of Sciences
[3]Horizon Robotics [4]Huawei Inc. [5]Center for Artificial Intelligence and Robotics, HKISI_CAS
{pancong2018, zhaoxiang.zhang}@ia.ac.cn {yonghao.he}@aliyun.com
{jrpeng4ever}@126.com {qian01.zhang, wei.sui}@horizon.ai

## Abstract

*Bird's Eye View (BEV) semantic segmentation is a critical task in autonomous driving. However, existing Transformer-based methods confront difficulties in transforming Perspective View (PV) to BEV due to their unidirectional and posterior interaction mechanisms. To address this issue, we propose a novel Bi-directional and Early Interaction Transformers framework named BAEFormer, consisting of (i) an early-interaction PV-BEV pipeline and (ii) a bi-directional cross-attention mechanism. Moreover, we find that the image feature maps' resolution in the cross-attention module has a limited effect on the final performance. Under this critical observation, we propose to enlarge the size of input images and downsample the multi-view image features for cross-interaction, further improving the accuracy while keeping the amount of computation controllable. Our proposed method for BEV semantic segmentation achieves state-of-the-art performance in real-time inference speed on the nuScenes dataset, i.e., 38.9 mIoU at 45 FPS on a single A100 GPU.*

## 1. Introduction

Recently, pure vision-based perception methods [16, 18, 25, 38] have occupied a significant position in autonomous driving due to their higher signal-to-noise ratio and lower cost compared to LIDAR-based methods [6, 14, 35, 39, 42]. Among them, Bird's-Eye-View (BEV) perception has become the mainstream method. The BEV representation learning in vision-centric autonomous driving is to take consecutive frames from multiple surrounding cameras as input and transform the pixel panel view to Bird's-Eye-View to conduct perception tasks such as 3D object detection, map-view semantic segmentation, and motion prediction.

The BEV representation offers several inherent advantages for visual perception. Firstly, it facilitates the integration of pure vision-based results with those from other modalities. Secondly, it provides a natural means to unify and express different perspectives under BEV to simplify the subsequent module development and deployment, such as planning and control. Finally, the object representation in BEV circumvents the common scale and occlusion difficulties that arise in 2D tasks.

The enhancement of BEV perception performance hinges on the rapid and graceful acquisition of road and object feature representations. Figure 1 illustrates that there are two categories of BEV perception pipelines based on distinct interaction mechanisms: (a) Late-interaction and (b) Middle-interaction. The late-interaction pipeline [24] employs independent perception on each camera view, followed by temporal and spatial fusion of the results into a unified BEV space. Recently, the most widely used pipeline is the middle-interaction [16, 18, 25, 38, 41]. It concatenates all camera inputs as a whole into the network, transforms them into the BEV space, and then outputs the result directly. The middle-interaction pipeline comprises a well-defined workflow for feature extraction, space transformation, and BEV space learning. Nevertheless, the transformation of PV to BEV using these two interaction strategies remains arduous. To tackle this challenge, we propose a novel paradigm: (c) Early-interaction method, which deserves further attention from the research community.

Our proposed early-interaction method offers distinct advantages when compared to the two existing strategies. Firstly, the image-space backbone only extracts image features with different resolutions sequentially without any information integration across resolutions. In contrast, our proposed method advocates for the integration of global

∗ This work was done during Cong Pan's internship at Horizon Robotics. ✉ Zhaoxiang Zhang is the Corresponding author.

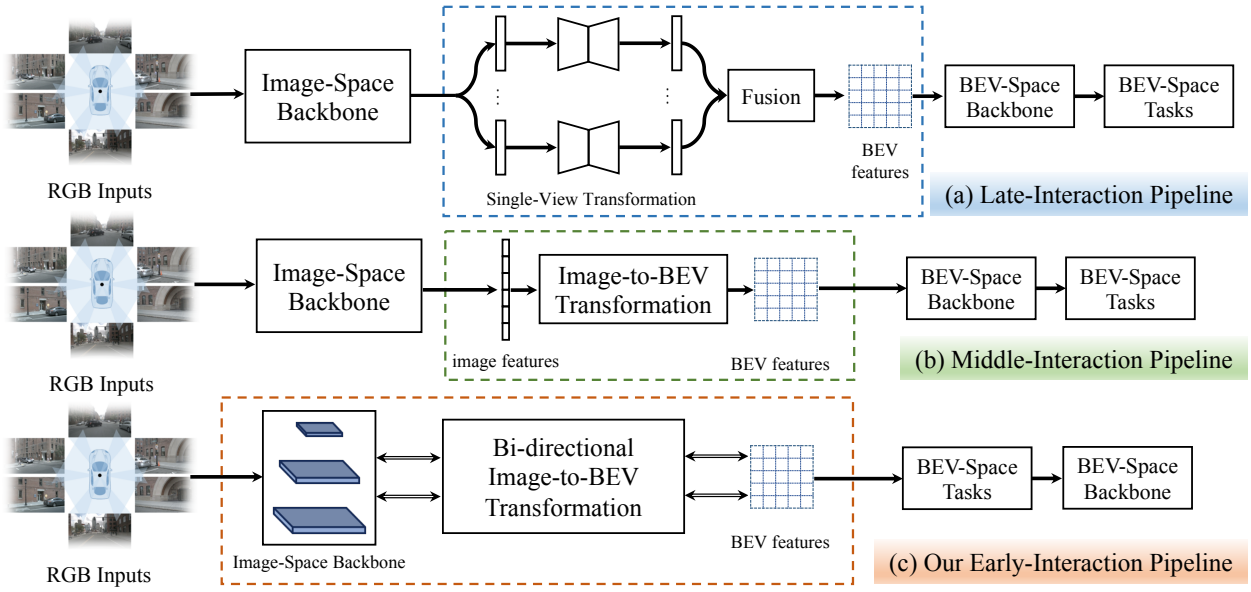



Figure 1. View-transformation taxonomy. The illustrations of the previous (a) late-interaction pipeline, (b) middle-interaction pipeline, and (c) our proposed early-interaction pipeline.

contextual information and local details, which enables the delivery of richer semantic information to the BEV space. Secondly, the computation of the core module in the existing strategies is predominantly occupied by the image-space backbone, which does not incorporate any BEV space information. Moreover, the information flow in the forward processing of late-interaction and middle-interaction strategies is unidirectional, with information flowing from the image space to the BEV space, while the information in the BEV space does not effectively affect the features in the image space. To address these issues, we suggest that the view transformation of image features to BEV features should occur not only after the image feature extraction but can be converted gradually during the extraction process. This way, the information flow can implicitly interact bilaterally, aligning features in the PV and BEV. Ultimately, the core challenge of vision-based BEV perception is the transformation from image space to BEV space. We propose a solution by distributing the learning of cross-space alignment throughout the entire structure, with the image network learning not only good feature representation but also cross-space alignment.

To this end, we propose a novel framework named Bi-directional and Early Interaction Transformers (BAEFormer) to effectively aggregate multi-scale image features into a better BEV feature representation and perform map-view semantic segmentation. Firstly, we initialize grid-shape BEV Queries and encode the camera parameters into positional embeddings, following the CVT [41] method. Subsequently, we utilize the cross-attention mechanism in Transformer [8, 37] to interact BEV features with multi-scale image features in both directions. This bi-directional interaction involves the use of an unshared attention map to update BEV and image features simultaneously, with image features from the previous scale influencing the extraction of the following scale's features. Furthermore, we observe that the resolution of the multi-scale feature maps during the interaction has a negligible effect on the final accuracy. As such, we can maintain the full amount of interaction while managing the number of parameters and computational costs by increasing the input image resolution and downsampling the image features at each scale before interaction.

Our contributions can be summarized as follows:

- We propose a novel framework, Bi-directional and Early Interaction Transformers (BAEFormer), to achieve a better view transformation from image feature space to BEV feature space.
- We find that the image features' resolution during the interaction does not significantly influence performance, so we can increase the input image resolution and downsample the image features at each scale before an interaction. This will lead to superior performance while simultaneously controlling parameters and computational costs.
- We conduct extensive experiments on nuScenes [3] and Lyft [13] datasets with comprehensive ablation studies, demonstrating the efficiency and effectiveness of our proposed BAEFormer method. We achieve state-of-the-art performance at real-time inference speed on BEV semantic segmentation.

## 2. Related Work

Most BEV Semantic Segmentation works follow late-interaction or middle-interaction pipelines to extract features from monocular or multi-view images and then convert features from PV to BEV. Pure vision-based view transformation strategies are different and can be divided into geometry-based methods and network-based methods.

### 2.1. Geometry-based BEV Semantic Segmentation

Geometry-based BEV Semantic Segmentation methods leverage geometric projection to transform PV-BEV. IPM [22] is the earliest work that utilizes a homography matrix to define the mapping between the camera and the ground planes. Some methods [1, 32] follow IPM to warp the front-view image onto the ground plane via a homography. Since IPM is based on the strong assumption of a flat ground plane, which usually leads to distortion for objects above the ground. To alleviate this shortcoming, BridgeGAN [43] takes the homography view as an intermediate view and proposes a multi-GAN-based model to learn the cross-view translation. Cam2BEV [27] trains a BEV decoder with a synthetic dataset to refine the IPM projection.

Another line of works [12, 23, 26, 29, 33] explicitly estimates depth to lift 2D features to 3D space and then splat them in BEV space. OFT [29] maps image-based features into an orthographic 3D space. It assumes that the distribution is uniform, i.e., all features along the ray are identical. Instead of predicting a uniform distribution along the depth for each image pixel, LSS [26] learns a categorical distribution and the context vector to better approximate the proper depth distribution. Recently, FIERY [12] has extended LSS [26] from single timestep to multi-timestep observation to do occupancy forecasting.

### 2.2. Network-based Semantic Segmentation

An alternative to explicit geometric projection is to model the view transformation implicitly in a data-driven way. To this end, the neural network serves as a mapping function from PV to BEV. Some methods [4, 7, 11, 15, 21, 24, 28, 30, 40, 44] use multilayer perceptron (MLP) to learn implicit representations of camera calibrations. VED [21] predicts a semantic occupancy grid with a variational encoder-decoder network directly from an image. VPN [24] uses MLP-based modules to aggregate the information from multiple first-view observations with different angles and modalities and outputs the top-down-view semantic map. PON [28] uses a feature pyramid to augment the high-resolution features with spatial context from lower pyramid layers and uses a stack of dense transformer layers to map the image feature into BEV. HDMapNet [15] proposes a feature projection module from PV to BEV that consists of both neural feature transformation to model a 3D environment implicitly and geometric projection to consider the camera extrinsic explicitly. PYVA [40] exploits the cycle consistency between views to fully use their correlation to strengthen the cross-view transformation module. HFT [44] designs a hybrid feature transformation consisting of IPM-based and MLP-based branches to make full use of geometry information and capture global context.

Other works [2, 9, 16, 19, 25, 38, 41] query corresponding image features through the attention mechanism in Transformers, which allows learning of long-range dependencies, to conduct a top-down framework. Inspired by the pioneering 2D detection framework DETR [5], DETR3D [38] uses a geometry-based reference points projection and feature sampling operation to refine the learnable sparse queries iteratively. Following [38], BEVFormer [16] designs a dense grid-shape learnable BEV queries along with a spatial deformable cross-attention layer and a temporal self-attention layer to lookup spatial features from cross cameras and temporal features from history BEV, respectively. Also, BEVSegFormer [25] utilizes deformable attention to transform multi-view image features to BEV representations for semantic map construction. LaRa [2] proposes latent representations of multi-view images, which are processed by a series of self-attention blocks, and then achieves BEV features through querying the latent space with a cross-attention module. CVT [41] uses a camera-aware cross-view attention mechanism that equips each camera-view feature with positional embeddings that depend on its intrinsic and extrinsic calibration. Though similar to our work, a direct cross-attention between multi-scale multi-camera image features and BEV grid in CVT [41] is computationally expensive, which limits the extensibility of the module. Meanwhile, the transformer-based methods mentioned above rely on a fixed structure that involves extracting features from PV images before transforming them to BEV space. This approach may lead to insufficient interaction between the image space and the BEV space, resulting in a homogeneous flow of information.

## 3. Method

### 3.1. Framework overview

As shown in Figure 2, we design a simple yet effective architecture that transforms PV to BEV for BEV semantic segmentation. Given images from multiple camera views along with their corresponding camera intrinsic and extrinsic parameters, the approach is designed to estimate a binary semantic segmentation mask in the map-view embedding coordinates. The framework consists of two key components: (i) a bi-directional early interaction encoder that simultaneously extracts image features and transforms them from PV to BEV; and (ii) a decoder that upsamples low-resolution BEV features to high-resolution BEV features,

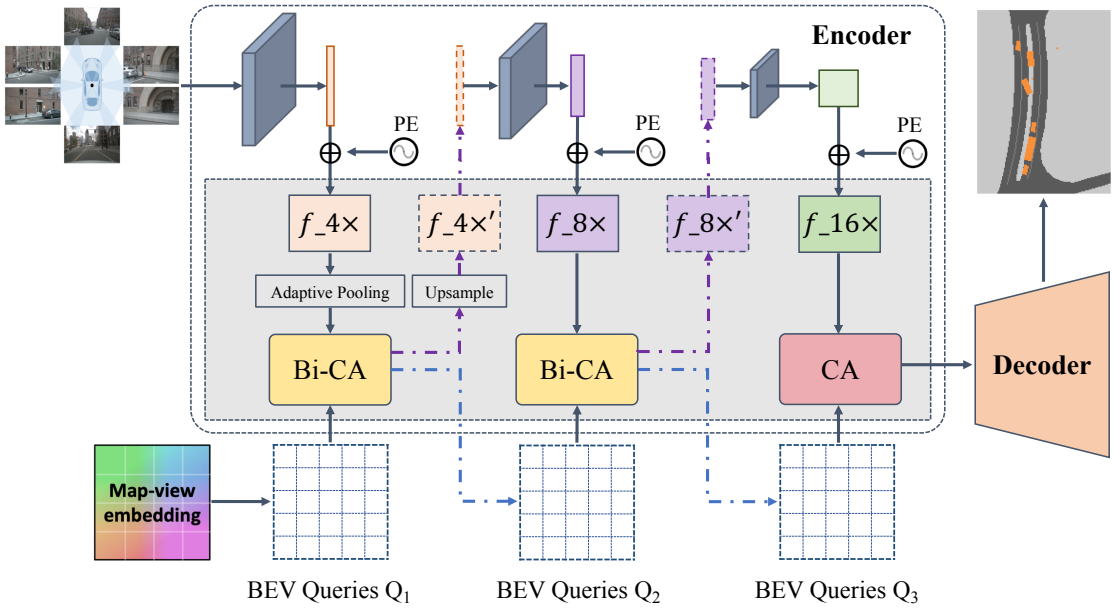



Figure 2. Overall framework of our proposed EBAFormer approach. We extract multi-view image features across multiple scales by EfficientNet-B4. Using known camera pose and intrinsics, we construct a camera-aware positional embedding. The encoder layer contains Bi-CA (Bi-directional Cross-view Attention) and CA (Cross-view Attention). Bi-CA is used to refine image features and grid-shaped BEV queries simultaneously, and CA is only used to refine BEV queries. Finally, a lightweight convolutional decoder upsamples the refined map-view embedding and produces the final segmentation output.

which is pertinent for the downstream task.

Specifically, we first extract high-resolution (4×) multi-view image features by early stage in EfficientNet-B4 [34] and downsample the 4× features to 16× through adaptive pooling module. Using known camera pose and intrinsics, we construct a camera-aware positional embedding both for image embeddings and BEV embeddings following CVT [41]. Secondly, we use a bi-directional cross-attention module to update multi-view image features and BEV features at the same time. Then we recover the scale of the refined multi-view image features through an upsample module and use it as the input of the middle stage in EfficientNet-B4 [34] to extract lower resolution (8×) multi-view image features. By analogy, the BEV features are interacted with a series of multi-view image features at different scales to obtain the refined BEV features. Finally, a lightweight convolutional decoder upsamples the advanced BEV features to generate the final segmentation results. The entire network is end-to-end differentiable.

| Module | # Parameters(K) |
|---|---|
| Encoder.image-backbone | 340 |
| Encoder.view-transformation | 306 |
| Encoder.BEVEmbedding | 80 |
| Encoder.others | 71.2 |
| Decoder | 244 |

Table 1. Comparison of parameters of different modules in our baseline model [41].

### 3.2. Input and output modeling

Our model starts with a set of images from $N$ views $(I_n, K_n, R_n, t_n)_{n=1}^{N}$, with $I_n \in \mathbb{R}^{H\times W\times 3}$ the image produced by camera $n$, $K_n \in \mathbb{R}^{3\times 3}$ the intrinsics, $R_n \in \mathbb{R}^{3\times 3}$ and $t_n \in \mathbb{R}^3$ the extrinsic rotation and translation, respectively. $H$ and $W$ mean the height and width of the input images.

Following [12, 26, 41], we use a pretrained EfficientNet-B4 [34] as a shared image extractor $E$ to obtain multi-view image features $F_n = E(I_n) \in \mathbb{R}^{h\times w\times c}$. These spatial feature maps in $\mathbb{R}^{N\times h\times w\times c}$ are then rearranged as a sequence of feature vectors in $\mathbb{R}^{N\times(hw)\times c}$, where $c$ means the number of channels. Also, we predefine a group of grid-shaped learnable parameters $Q_{bev} \in \mathbb{R}^{h_{bev}\times w_{bev}\times c_{bev}}$ as the queries of BEV features. Following [41], the respective camera-aware position embeddings are added to the image feature vectors and BEV embeddings to retain positional information. The final step is to upsample the BEV features $Q_{bev}$ to a specified resolution with a decoder and predict the binary bird's-eye-view semantic map $\hat{y} \in \{0,1\}^{h_{\text{bev}}\times w_{\text{bev}}\times C}$.

### 3.3. Early Interaction

As shown in Table 1, the core (encoder) part of the network is composed of the image backbone and the view transformation modules, with the former accounting for nearly half of the total parameters, despite only being used

to extract image features devoid of any BEV spatial information. To address this, we merge the image backbone and view transformation modules using a bi-directional cross-attention mechanism, forming the proposed early interaction module, which consists of three layers obtaining $4\times$, $8\times$ and $16\times$ resolutions of image feature maps ($f_{4\times}$, $f_{8\times}$ and $f_{16\times}$) from the pretrained model, respectively.

After bi-directional and early interaction, the resulting refined multi-view image feature maps, $f'_{4\times}$ and $f'_{8\times}$, replace the original features and serve as the input to the subsequent stage. Our multi-scale early interaction method leverages the hierarchical structure of pretrained models to integrate multi-scale image features. Additionally, the BEV spatial information is incorporated into the backbone network, enabling the early interactional backbone to partially assume the function of heterogeneous-space alignment.

### 3.4. Bi-directional Cross-attention

As shown in Figure 3, our proposed transformer block with a bi-directional cross-attention module contains two branches that refine multi-view image features and BEV features, respectively. Each branch follows the simple and standard Transformer Encoder structure in ViT [8].

Specifically, the multi-view spatial feature maps are first transformed into queries $Q_f \in \mathbb{R}^{N\times(hw)\times c}$, keys $K_f \in \mathbb{R}^{N\times(hw)\times c}$ and values $V_f \in \mathbb{R}^{N\times(hw)\times c}$, where $c$ indicates their dimensions. Similarly, the initialized BEV embeddings are transformed into queries $Q_{bev} \in \mathbb{R}^{(h_{bev}w_{bev})\times c_{bev}}$, keys $K_{bev} \in \mathbb{R}^{(h_{bev}w_{bev})\times c_{bev}}$ and values $V_{bev} \in \mathbb{R}^{(h_{bev}w_{bev})\times c_{bev}}$, where $c_{bev}$ indicates its dimensions. The bi-directional cross-attention for image features $Z_f$ and BEV features $Z_{bev}$ can be computed as

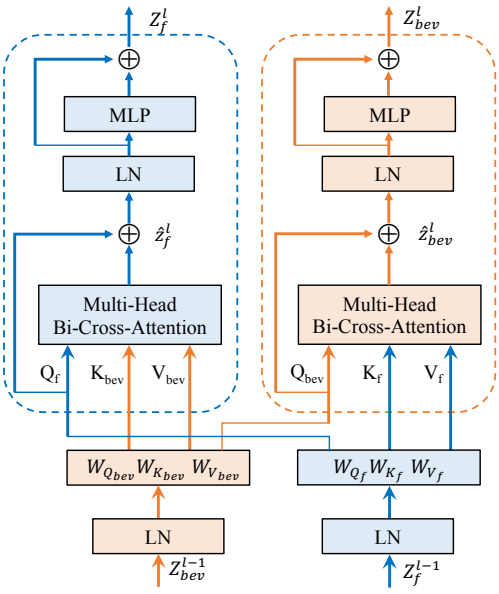



Figure 3. Overview of transformer block with bi-directional cross-attention. LN denotes layer normalization and MLP denotes multi-layer perceptron. $W_Q$, $W_K$, and $W_V$ denote transformation parameters to generate queries, keys and values, respectively.

$$\mathrm{BiCA}_{Z_f}(Z_f, Z_{bev}) = \mathrm{softmax}\left(\frac{Q_f K_{bev}^T}{\sqrt{c}}\right) V_{bev}, \quad (1)$$

$$\mathrm{BiCA}_{Z_{bev}}(Z_{bev}, Z_f) = \mathrm{softmax}\left(\frac{Q_{bev} K_f^T}{\sqrt{c_{bev}}}\right) V_f. \quad (2)$$

The entire transformer block can be formulated as

$$\hat{Z}_f^l = \mathrm{MHBiCA}\left(\mathrm{LN}\left(Z_f^{l-1}\right), \mathrm{LN}\left(Z_{bev}^{l-1}\right)\right) + Z_f^{l-1}, \quad (3)$$

$$Z_f^l = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{Z}_f^l\right)\right) + \mathrm{LN}\left(\hat{Z}_f^l\right), \quad (4)$$

$$\hat{Z}_{bev}^l = \mathrm{MHBiCA}\left(\mathrm{LN}\left(Z_{bev}^{l-1}\right), \mathrm{LN}\left(Z_f^{l-1}\right)\right) + Z_{bev}^{l-1}, \quad (5)$$

$$Z_{bev}^l = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{Z}_{bev}^l\right)\right) + \mathrm{LN}\left(\hat{Z}_{bev}^l\right), \quad (6)$$

where $Z_f^{l-1}$ and $Z_{bev}^{l-1}$ denote inputs for the $l$th transformer block, $Z_f^l$ and $Z_{bev}^l$ denote the corresponding outputs for the $l$th transformer block. $\mathrm{LN}(\cdot)$ denotes layer normalization [36], $\mathrm{MLP}(\cdot)$ denotes 2-layer fully connected neural network with a GELU non-linearity [10], and $\mathrm{MHBiCA}(\cdot)$ is $\mathrm{BiCA}(\cdot)$ defined in Equation 1 and 2 with multi-head cross attention.

Since it is challenging to transform PV to BEV, the exchange of information between image space and BEV space is crucial. Our proposed bi-directional cross-attention mechanism offers an implicit means of constraining this information flow. Not only do image features impact BEV features, but in turn, progressively refined BEV features promote a tailored image feature extraction process.

## 4. Experiments

### 4.1. Dataset and Evaluation Metrics

**Dataset.** We evaluate our proposed framework on the challenging nuScenes [3] and Lyft [13] datasets. NuScenes contains 1000 scenes that are captured from four locations in Boston and Singapore under various weather conditions and at different times of the day. Each scene lasts roughly 20 seconds and the key samples are annotated 2 Hz. The dataset includes RGB images from 6 cameras with a 360° horizontal field of view, and there is a slight overlap between the cameras' fields of view. Camera intrinsics and extrinsics are provided for each camera in every scene. Following [41], we generate the ground-truth for our binary semantic segmentation task by rendering the raw annotations, the 3D bounding boxes, into the discretized BEV of scenes. In addition, following [26], 48 of the Lyft scenes are separated to obtain 6048 samples for validation.

| Method | visibility > 0% | | visibility > 40% | | # Parameters(M) | FPS |
|---|---|---|---|---|---|---|
| | Setting 1 (IoU%) | Setting 2 (IoU%) | Setting 1 (IoU%) | Setting 2 (IoU%) | | |
| VPN [24] | 25.5 | - | - | - | 18 | - |
| STA [31] | 36.0 | - | - | - | - | - |
| FIERY Static [12] | 37.7 | 35.8 | - | - | 7.4 | 8 |
| BEVFormer [16] | - | **43.2** | - | - | 68.1 | 2 † |
| PON [28] | 24.7 | - | - | - | 38 | 30 |
| LSS [26] | - | 32.1 | - | - | 14 | 25 |
| CVT [41] | - | - | 37.5 | 36.0 | 5 | 48 † |
| BAEFormer (224×480) | **39.5** | **36.0** | **42.0** | **38.9** | 5.6 | 45 |
| BAEFormer (504×1056) | **41.2** | **37.8** | **44.2** | **41.0** | 5.6 | 24 |

Table 2. Vehicle map-view segmentation on nuScenes. Setting 1 refers to the 100m×50m at 25cm resolution setting and Setting 2 refers to the 100m×100m at 50cm resolution setting. Different visibility levels of vehicles (0% or 40%) are considered for training and validation. For a fair comparison, † denotes our evaluated inference speed on a single A100 GPU, which is faster than the ones reported in the original paper. The top and bottom rows correspond to the non-real-time and real-time models.

**Evaluation.** We utilize Intersection-over-Union (IoU) score between the predicted results and the ground-truth BEV labels as the main performance measure. Setting 1 refers to the 100m×50m at 25cm resolution setting proposed by [28] and Setting 2 refers to the 100m×100m at 50cm resolution setting proposed by [26]. All ablation studies are conducted under Setting 2 with 40% visibility level.

## 4.2. Implementation Details

**Architecture.** The input images are scaled to $224 \times 480$ unless specified otherwise. We feed the input multi-view images to EfficientNet-B4 [34] and obtain multi-scale image feature maps - (56, 120), (28, 60) and (14, 30) with stride 4, 8, and 16, respectively. The initial map-view embedding is a tensor of learnable parameters $w_{bev} \times h_{bev} \times c_{bev}$, where $w_{bev} = h_{bev} = 25$. All the dimensions of image features and BEV features are set to $c = c_{bev} = 128$. We use multi-head attention with four heads and an embedding size $d_{head} = 64$. The decoder consists of three bilinear upsample and convolutional layers to upsample the last BEV features to the final output size. Each decoder layer enlarges the BEV feature size by a factor of 2 up to a final output resolution of $200 \times 200$, corresponding to a $100 \times 100$ meters area centered around the ego-car.

**Training.** We use the AdamW [20] optimizer with a constant learning rate of 4e-3 and a weight decay of 1e-7. We train our model on 4 Nvidia A100 GPUs with a total batch size of 16. All inference speeds are given on a single Nvidia A100 GPU. The model is optimized by a Focal Loss [17] with our predicted soft segmentation maps and the binary ground-truth.

## 4.3. Comparison with previous works

In Table 2, we compare IoU performances, parameters, and inference speed of our proposed BAEFormer under two settings with the previous competitive works on vehicle BEV segmentation. Though BEVFormer [16] achieves high performance, it is time-consuming with a huge model size of up to 68.1M parameters. Ours with large input image resolution ($504 \times 1056$) runs $12\times$ faster than BEVFormer [16] and is approximately $\frac{1}{12}$ of its size.

| Method | nuScenes | | Lyft | |
|---|---|---|---|---|
| | Veh. | Driv. | Veh. | Driv. |
| OFT [29] | 30.1 | 71.7 | 40.43 | - |
| LSS [26] | 32.1 | 72.9 | 44.6 | - |
| CVT [41] | 36.0 | 74.3 | 45.4† | 80.2† |
| BAEFormer (224×480) | **38.9** | **76.0** | **46.6** | **82.8** |

Table 3. IoU for vehicles and driveable area on nuScenes and Lyft datasets under Setting 2. † denotes our implementation.

Our proposed BAEFormer can run at 45 FPS on A100 GPU with 42.0 mIoU for Setting 1 and 38.9 mIoU for Setting 2, which is **4.5/2.9** points higher for Setting 1/2 than the real-time model CVT [41], respectively. Under larger input image resolution ($504 \times 1056$), our BAEFormer is **6.7/5.0** points higher for Setting 1/2 than CVT [41].

Also, Table 3 shows that our BAEFormer outperforms all alternative approaches for both vehicles and driveable area on nuScenes [3] and Lyft [13] datasets. Specifically, BAEFormer is **1.7** points higher with the same input resolution ($224 \times 480$) for driveable area than CVT [41] on nuScenes. And our BAEFormer is **1.2/2.6** points higher for vehicle/driveable area than CVT [41] on Lyft.

## 4.4. Ablation Study

### 4.4.1 Ablations on different interaction methods

Table 4 illustrates different interaction methods for vehicles on nuScenes [3] dataset. (a) We reimplement CVT [41] and use $4\times$, $8\times$, and $16\times$ features for interaction as our baseline, which is a unidirectional and posterior interaction method. (b) Only refining different scales of multi-view image features in cross-attention modules and using the same query for each module. The updated multi-scale features are taken and interacted with the BEV queries layer by layer. (c) After extracting multi-scale image features through a

| | Early Interaction | BiCA | mIoU(%) |
|---|---|---|---|
| (a) Baseline | - | - | 37.3 |
| (b) Baseline + EI | ✓ | - | 37.6 |
| (c) Baseline + BiCA | - | ✓ | 37.4 |
| (d) Our BAEFormer | ✓ | ✓ | **38.9** |

Table 4. Ablation studies for different interaction methods for vehicles on nuScenes.

| | 4× | 8× | 16× | Baseline | BAEFormer |
|---|---|---|---|---|---|
| (a) | ✓ | - | - | 22.9 | 22.0 |
| (b) | - | ✓ | - | 32.2 | 32.2 |
| (c) | - | - | ✓ | 35.8 | 36.0 |
| (d) | ✓ | ✓ | - | 33.7 | 36.4 |
| (e) | ✓ | - | ✓ | 37.0 | 37.8 |
| (f) | - | ✓ | ✓ | 36.8 | 38.4 |
| (g) | ✓ | ✓ | ✓ | 37.3 | **38.9** |

Table 5. Ablation studies for the combinations of different image feature scales for vehicles on nuScenes.

| | 4× | 8× | 16× | input resolution | mIoU(%) | Memory(GB) |
|---|---|---|---|---|---|---|
| (a) | CA | - | - | 224×480 | 22.9 | 11.5 |
| (b) | - | CA | - | 224×480 | 32.2 | 6.3 |
| (c) | - | - | CA | 224×480 | 35.8 | 6.4 |
| (d) | CA | CA | - | 224×480 | 33.7 | 13.4 |
| (e) | CA | - | CA | 224×480 | 37.0 | 11.4 |
| (f) | - | CA | CA | 224×480 | 37.0 | 5.7 |
| (g) | CA | CA | CA | 224×480 | 37.3 | 13.8 |
| (h) | CA | CA | CA | 296×640 | 37.7 | 15.6 |
| (i) | CA | CA | CA | 372×800 | 38.5 | 24.2 |
| | 4×→16× | 8× | 16× | input resolution | mIoU(%) | Memory(GB) |
| (j) | CA | CA | CA | 504×1056 | 38.8 | 23.8 |
| (k) | BiCA | BiCA | CA | 224×480 | 38.9 | 8.0 |
| (l) | BiCA | BiCA | CA | 296×640 | 39.2 | 12.2 |
| (m) | BiCA | BiCA | CA | 372×800 | 39.8 | 18.9 |
| (n) | BiCA | BiCA | CA | 504×1056 | **41.0** | 24.0 |

Table 6. Combinations of different input resolutions and different image feature scales for vehicles on nuScenes. (a)-(j) is based on the **baseline model** (interaction through CA) and (k)-(n) is based on **BAEFormer model** (interaction through BiCA). Memory is measured on A100 GPU with batch size 4. 4× image features in (j)-(n) are all downsampled to 16×.

pretrained image backbone, both multi-view image features and BEV features are updated by bi-directional cross-attention modules. Therefore, the image backbone is fine-tuned with BEV-space information and without inter-scale information. (d) Our proposed BAEFormer method with Bi-directional cross-attention and early interaction modules. The mIoU performance denotes that our proposed BAEFormer method not only fully integrates inter-scale information to obtain a better semantic representation but also achieves better spatial alignment through bi-directional constraints on the flow of information in heterogeneous spaces.

#### 4.4.2 Ablations on different scales and resolutions

**Combination of different scales.** Table 5 shows the combinations of different resolutions. The results show that 16× resolution image features are the most important in the baseline and our proposed models. Moreover, the more scales of multi-view image features we have, the better the performance of our model compared to the baseline model.

**Combination of different resolutions and scales.** Table 6 illustrates mIoU performance and memory usage of models with different input resolutions and image feature scales. However, the A100 GPU with 40G available memory cannot cover the entire model when the input resolution of the baseline model is at its maximum ($504 \times 1056$). Therefore, as shown in (j), we can get better performance while keeping the computation manageable through downsampling the 4× features to 16×. As shown in (j)-(n), the resolution of the input images during the interaction does not have much impact on the final accuracy. Also, as the input image resolution continuously increases, the baseline models and our proposed models can both obtain higher accuracy. So we can keep this full amount of interaction within a manageable number of parameters and computational costs by downsampling the image features at each scale.

| | CA | Bi-CA | mIoU(%) |
|---|---|---|---|
| (a) | 8×,16× | 4× | 37.5 |
| (b) | 4×,16× | 8× | 38.3 |
| (c) | 16× | 4×,8× | 37.8 |
| (d) | 8×→16×,16× | 4× | 37.2 |
| (e) | 4×→16×,16× | 8× | 38.6 |
| (f) | 8×,16× | 4×→16× | 37.6 |
| (g) | 4×,16× | 8×→16× | 37.0 |
| (h) | 16× | 4×→16×, 8×→16× | 37.9 |
| (i) | 16× | 4×→16×, 8× | **38.9** |
| (j) | 8×→16×,16× | 4×→16× | 36.8 |
| (k) | 4×→16×,16× | 8×→16× | 38.3 |

Table 7. Ablation studies for the combinations of different scales for standard cross-attention (CA) and our proposed bi-directional cross-attention (Bi-CA) mechanisms for vehicles on nuScenes.

#### 4.4.3 Ablations on combinations of CA and Bi-CA

Table 7 illustrates the combination of different scales for standard cross-attention (CA) and our proposed bi-directional cross-attention (Bi-CA) mechanism. From the results in (b), (e), (i), and (k) we can see that applying Bi-

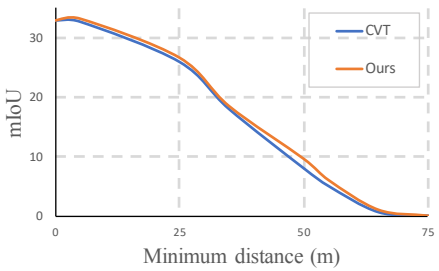
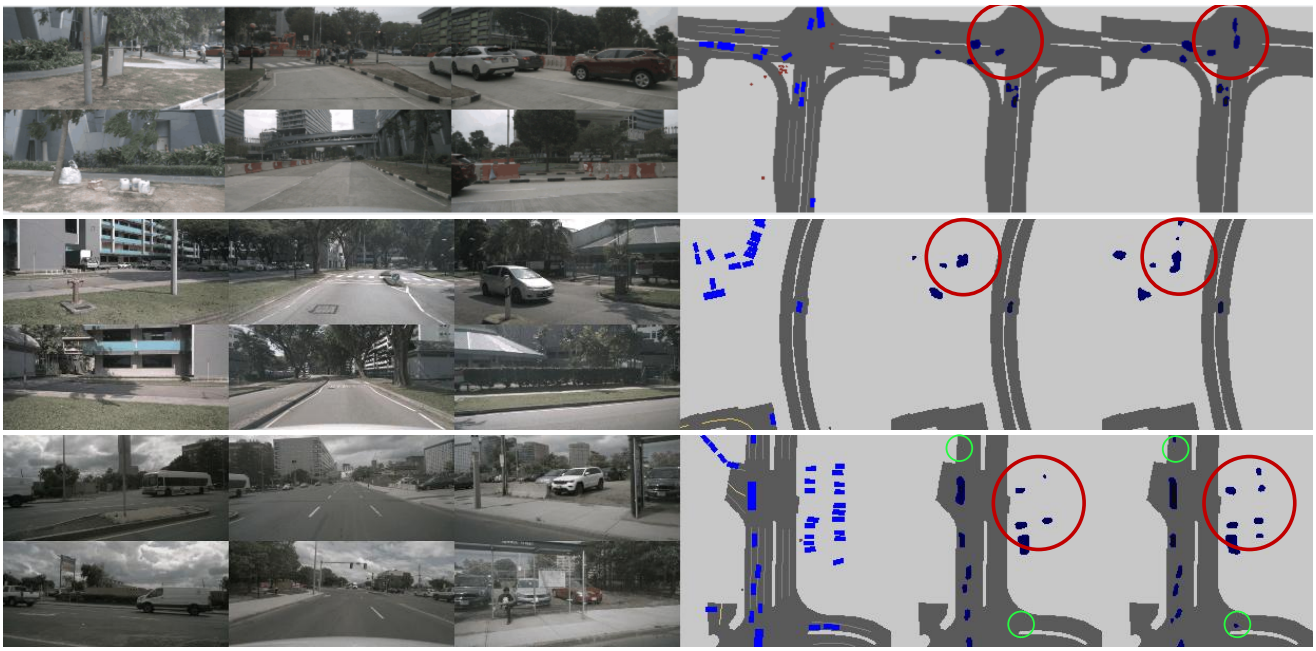



Figure 4. A comparison of model performance v.s. distance to the ego-car between CVT and ours.



(a) multi-view images (b) BEV_gt (c) BEV_pred baseline model (d) BEV_pred ours

Figure 5. Qualitative results of different models on nuScenes. (a) shows the 6 camera views surrounding the vehicle. The top 3 views are front-facing and the bottom 3 views are back-facing. (b) shows the ground-truths map-view semantic segmentation for vehicles and driveable areas. (c) shows the BEV semantic segmentation prediction results of the baseline model for vehicles. (d) shows the BEV semantic segmentation prediction results of our proposed model for vehicles. To clarify, the driveable areas in (c) and (d) are ground-truths.

CA method on $8\times$ multi-view image features will achieve better performance. As shown in (d)-(e) and (f)-(k), the input resolution of image features of Bi-CA module is more important than that of the standard CA module.

### 4.5. Accuracy vs distance

As shown in Figure 4, we evaluate how well our model performs as the distance to the ego-car increases. In this experiment, all predictions that are closer than a certain distance to the ego-car are ignored. The results show that our method not only outperforms CVT at close-by distances but also outperforms CVT at longer ranges.

### 4.6. Qualitative Results

We show the BEV visualization of our proposed model for BEV semantic segmentation in Figure 5. Our proposed bi-directional cross-attention and early interaction model is effective in reducing the number of missed near objects of ego-car (red circles) compared to the baseline model, and can also perceive distant objects of ego-car (green circles), further illustrating the perceptual ability of our approach.

## 5. Conclusion

In this paper, we propose a novel framework for BEV semantic segmentation called BAEFormer, which utilizes a Bi-directional and Early Interaction Transformers approach. Firstly, we employ a bi-directional cross-attention mechanism to establish improved cross-space alignment by imposing bi-directional constraints on information flow in the heterogeneous space of image feature and BEV feature spaces. Secondly, we utilize the early interaction method to incorporate inter-scale information and achieve a more refined semantic representation. Additionally, we make a crucial observation that augmenting the image resolution in the cross-attention module has a limited impact on final performance but significantly increases computational overhead. Thus, we propose the enlargement of input images, followed by downsampling for cross-interaction, to enhance accuracy while maintaining computational efficiency. Empirical evaluations conducted on nuScenes and Lyft datasets reveal that our approach significantly surpasses real-time competitors.

## 6. Acknowledgements

This work was supported in part by the Major Project for New Generation of AI (No.2018AAA0100400), the National Natural Science Foundation of China (No.61836014, No.U21B2042, No.62072457, No.62006231) and the InnoHK program.

## References


[1] Syed Ammar Abbas and Andrew Zisserman. A geometric approach to obtain a bird's eye view from an image. In *ICCV Workshops*, pages 0–0, 2019. 3

[2] Florent Bartoccioni, Éloi Zablocki, Andrei Bursuc, Patrick Pérez, Matthieu Cord, and Karteek Alahari. Lara: Latents and rays for multi-camera bird's-eye-view semantic segmentation. *arXiv preprint arXiv:2206.13294*, 2022. 3

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2, 5, 6

[4] Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, and Luc Van Gool. Structured bird's-eye-view traffic scene understanding from onboard images. In *ICCV*, pages 15661–15670, 2021. 3

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 3

[6] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *CVPR*, pages 1907–1915, 2017. 1

[7] Kashyap Chitta, Aditya Prakash, and Andreas Geiger. Neat: Neural attention fields for end-to-end autonomous driving. In *ICCV*, pages 15793–15803, 2021. 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5

[9] Shi Gong, Xiaoqing Ye, Xiao Tan, Jingdong Wang, Errui Ding, Yu Zhou, and Xiang Bai. Gitnet: Geometric prior-based transformation for birds-eye-view segmentation. *arXiv preprint arXiv:2204.07733*, 2022. 3

[10] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[11] Noureldin Hendy, Cooper Sloan, Feng Tian, Pengfei Duan, Nick Charchut, Yuesong Xie, Chuang Wang, and James Philbin. Fishing net: Future inference of semantic heatmaps in grids. *arXiv preprint arXiv:2006.09917*, 2020. 3

[12] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *ICCV*, pages 15273–15282, 2021. 3, 4, 6

[13] R Kesten, M Usman, J Houston, T Pandya, K Nadhamuni, A Ferreira, M Yuan, B Low, A Jain, P Ondruska, et al. Lyft level 5 av dataset 2019. *urlhttps://level5. lyft. com/dataset*, 1:3, 2019. 2, 5, 6

[14] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, pages 12697–12705, 2019. 1

[15] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *2022 International Conference on Robotics and Automation(ICRA)*, pages 4628–4634. IEEE, 2022. 3

[16] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 1, 3, 6

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 6

[18] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. *arXiv preprint arXiv:2203.05625*, 2022. 1

[19] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022. 3

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[21] Chenyang Lu, Marinus Jacobus Gerardus van de Molengraft, and Gijs Dubbelman. Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks. *IEEE Robotics and Automation Letters(RAL)*, 4(2):445–452, 2019. 3

[22] Hanspeter A Mallot, Heinrich H Bülthoff, JJ Little, and Stefan Bohrer. Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biological cybernetics*, 64(3):177–185, 1991. 3

[23] Mong H Ng, Kaahan Radia, Jianfei Chen, Dequan Wang, Ionel Gog, and Joseph E Gonzalez. Bev-seg: Bird's eye view semantic segmentation using geometry and semantic point cloud. *arXiv preprint arXiv:2006.11436*, 2020. 3

[24] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters(RAL)*, 5(3):4867–4873, 2020. 1, 3, 6

[25] Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, and Erkang Cheng. Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. *arXiv preprint arXiv:2203.04050*, 2022. 1, 3

[26] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pages 194–210. Springer, 2020. 3, 4, 5, 6

[27] Lennart Reiher, Bastian Lampe, and Lutz Eckstein. A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view. In *2020 IEEE 23rd International Conference on Intelligent Transportation System(ITSC)*, pages 1–7. IEEE, 2020. 3

[28] Thomas Roddick and Roberto Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. In *CVPR*, pages 11138–11147, 2020. 3, 6

[29] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. *arXiv preprint arXiv:1811.08188*, 2018. 3, 6

[30] Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. Enabling spatio-temporal aggregation in birds-eye-view vehicle estimation. In *2021 IEEE International Conference on Robotics and Automation(ICRA)*, pages 5133–5139. IEEE, 2021. 3

[31] Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. Enabling spatio-temporal aggregation in birds-eye-view vehicle estimation. In *2021 IEEE International Conference on Robotics and Automation(ICRA)*, pages 5133–5139. IEEE, 2021. 6

[32] Sunando Sengupta, Paul Sturgess, L'ubor Ladickỳ, and Philip HS Torr. Automatic dense visual semantic mapping from street-level imagery. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS)*, pages 857–862. IEEE, 2012. 3

[33] Shashank Srikanth, Junaid Ahmed Ansari, R Karnik Ram, Sarthak Sharma, J Krishna Murthy, and K Madhava Krishna. Infer: Intermediate representations for future prediction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 942–949. IEEE, 2019. 3

[34] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019. 4, 6

[35] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *CVPR*, pages 4604–4612, 2020. 1

[36] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*, 2019. 5

[37] Xiyu Wang, Pengxin Guo, and Yu Zhang. Domain adaptation via bidirectional cross-attention transformer. *arXiv preprint arXiv:2201.05887*, 2022. 2

[38] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 1, 3

[39] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1

[40] Weixiang Yang, Qi Li, Wenxi Liu, Yuanlong Yu, Yuexin Ma, Shengfeng He, and Jia Pan. Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation. In *CVPR*, pages 15536–15545, 2021. 3

[41] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *CVPR*, pages 13760–13769, 2022. 1, 2, 3, 4, 5, 6

[42] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, pages 4490–4499, 2018. 1

[43] Xinge Zhu, Zhichao Yin, Jianping Shi, Hongsheng Li, and Dahua Lin. Generative adversarial frontal view to bird view synthesis. In *2018 International conference on 3D Vision(3DV)*, pages 454–463. IEEE, 2018. 3

[44] Jiayu Zou, Junrui Xiao, Zheng Zhu, Junjie Huang, Guan Huang, Dalong Du, and Xingang Wang. Hft: Lifting perspective representations via hybrid feature transformation. *arXiv preprint arXiv:2204.05068*, 2022. 3