BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision

Chenyu Yang[1*]   Yuntao Chen[2*]   Hao Tian[3*]   Chenxin Tao[1]   Xizhou Zhu[3]   Zhaoxiang Zhang[2,4]

Gao Huang[1]   Hongyang Li[5]   Yu Qiao[5]   Lewei Lu[3]   Jie Zhou[1]   Jifeng Dai[1,5✉]

[1]Tsinghua University   [2]Centre for Artificial Intelligence and Robotics, HKISI_CAS

[3]SenseTime Research   [4]Institute of Automation, Chinese Academy of Science (CASIA)

[5]Shanghai Artificial Intelligence Laboratory

Abstract

We present a novel bird's-eye-view (BEV) detector with perspective supervision, which converges faster and better suits modern image backbones. Existing state-of-the-art BEV detectors are often tied to certain depth pretrained backbones like VoVNet, hindering the synergy between booming image backbones and BEV detectors. To address this limitation, we prioritize easing the optimization of BEV detectors by introducing perspective view supervision. To this end, we propose a two-stage BEV detector, where proposals from the perspective head are fed into the bird's-eye-view head for final predictions. To evaluate the effectiveness of our model, we conduct extensive ablation studies focusing on the form of supervision and the generality of the proposed detector. The proposed method is verified with a wide spectrum of traditional and modern image backbones and achieves new SoTA results on the large-scale nuScenes dataset. The code shall be released soon.

1. Introduction

Bird's-eye-view(BEV) recognition models [17,21,25,27, 29,35,42] are a class of camera-based models for 3D object detection. They have attracted interest in autonomous driving as they can naturally integrate partial raw observations from multiple sensors into a unified holistic 3D output space. A typical BEV model is built upon an image backbone, followed by a view transformation module that lifts perspective image features into BEV features, which are further processed by a BEV feature encoder and some task-specific heads. Although much effort is put into designing the view transformation module [17,27,42] and incorporating an ever-growing list of downstream tasks [9,27] into the new recognition framework, the study of image

*: Equal contribution.

✉: Corresponding author, email: daijifeng@tsinghua.edu.cn.

backbones in BEV models receives far less attention. As a cutting-edge and highly demanding field, it is natural to introduce modern image backbones into autonomous driving. Surprisingly, the research community chooses to stick with VoVNet [13] to enjoy its large-scale depth pre-training [26]. In this work, we focus on unleashing the full power of modern image feature extractors for BEV recognition to unlock the door for future researchers to explore better image backbone design in this field.

However, simply employing modern image backbones without proper pre-training fails to yield satisfactory results. For instance, an ImageNet [6] pre-trained ConvNeXt-XL [23] backbone performs just on par with a DDAD-15M pretrained VoVNet-99 [26] for 3D object detection, albeit the latter has $3.5\times$ parameters of the former. We owe the struggle of adapting modern image backbones to the following issues: 1) The domain gap between natural images and autonomous driving scenes. Backbones pre-trained on general 2D recognition tasks fall short of perceiving 3D scenes, especially estimating depth. 2) The complex structure of current BEV detectors. Take BEVFormer [17] as an example. The supervision signals of 3D bounding boxes and object class labels are separated from the image backbone by the view encoder and the object decoder, each of which is comprised of transformers of multiple layers. The gradient flow for adapting general 2D image backbones for autonomous driving tasks is distorted by the stacked transformer layers.

In order to combat the difficulties mentioned above in adapting modern image backbones for BEV recognition, we introduce perspective supervision into BEVFormer, *i.e.* extra supervision signals from perspective-view tasks and directly applied to the backbone. It guides the backbone to learn 3D knowledge missing in 2D recognition tasks and overcomes the complexity of BEV detectors, greatly facilitating the optimization of the model. Specifically, we build a perspective 3D detection head [26] upon the backbone, which takes image features as input and directly predicts

the 3D bounding boxes and class labels of target objects. The loss of this perspective head, denoted as perspective loss, is added to the original loss (BEV loss) deriving from the BEV head as an auxiliary detection loss. The two detection heads are jointly trained with their corresponding loss terms. Furthermore, we find it natural to combine the two detection heads into a two-stage BEV detector, **BEV-Former v2**. Since the perspective head is full-fledged, it could generate high-quality object proposals in the perspective view, which we use as first-stage proposals. We encode them into object queries and gather them with the learnable ones in the original BEVFormer, forming hybrid object queries, which are then fed into the second-stage detection head to generate the final predictions.

We conduct extensive experiments to confirm the effectiveness and necessity of our proposed perspective supervision. The perspective loss facilitates the adaptation of the image backbone, resulting in improved detection performance and faster model convergence. While without this supervision, the model cannot achieve comparable results even if trained with a longer schedule. Consequently, we successfully adapt modern image backbones to the BEV model, achieving 63.4% NDS on nuScenes [2] test-set.

Our contributions can be summarized as follows:

- We point out that perspective supervision is key to adapting general 2D image backbones to the BEV model. We add this supervision explicitly by a detection loss in the perspective view.

- We present a novel two-stage BEV detector, BEV-Former v2. It consists of a perspective 3D and a BEV detection head, and the proposals of the former are combined with the object queries of the latter.

- We highlight the effectiveness of our approach by combining it with the latest developed image backbones and achieving significant improvements over previous state-of-the-art results on the nuScenes dataset.

2. Related Works

2.1. BEV 3D Object Detector

Bird's-eye-view (BEV) object detection has attracted more attention recently [17, 21, 25, 27, 29, 35, 42] due to its vast success in autonomous driving systems.

Early works including OFT [29], Pseduo LiDAR [35], and VPN [25] shed light on how to transform perspective features into BEV features but either for a single camera or on less well-known tasks. OFT [29] pioneered to adopt transformation from 2D image features to 3D BEV features for monocular 3D object detection. Pseudo LiDAR [35], as its name suggested, created pseudo point clouds through monocular depth estimation and camera intrinsics and processed them in the BEV space subsequently. VPN [25] was

the first to fuse multi-view camera inputs into a top-down view feature map for semantic segmentation.

Modern approaches enjoyed the convenience of integrating features from different perspective view sensors provided by 2D-3D view transformation. LSS [27] extended OFT by introducing a latent depth distribution during the pooling of BEV pillar features. Moreover, LSS pooled over six surrounding images compared with a single in OFT. Different from the 2D-to-3D lifting in LSS or the 3D-to-2D projection in OFT, CVT [42] utilized camera-aware positional encoding and dense cross attention to bridge perspective-view and BEV-view features. PETR [21] devised an approach without explicit BEV feature construction. Perspective feature maps are element-wisely fused with 3D positional embedding feature maps, and a subsequent DETR-style decoder is applied for object detection. BEVFormer [17] leveraged spatial cross-attention for view transformation and temporal self-attention for temporal feature fusion. The fully transformer-based structure of BEVFormer makes its BEV features more versatile than other methods, easily supporting non-uniform and non-regular sampling grids. Besides, as shown in Simple-BEV [7], multi-scale deformable attention [43] excels in all lifting strategies. So we choose to build our detector based on BEVFormer to exploit the strengths mentioned before.

Besides published works, there are many concurrent works due to the popularity of this field. BEVDet [10] introduced rich image-level and BEV-level augmentations for training. BEVStereo [14] and STS [37] both adopted a temporal stereo paradigm for better depth estimation. Polar-Former [11] came up with a non-cartesian 3D grid setting. SimpleBEV [7] compared different 2D-3D lifting methods.

Unlike existing works that mainly explore the designs for detectors, we focus on adapting modern image backbones into BEV recognition models.

2.2. Auxiliary Loss in Camera 3D Object Detection

Auxiliary losses are ubiquitous in monocular 3D object detection as most methods [15,20,26,30,32,33,40] are built upon 2D detectors like RetinaNet [18] and FCOS [31]. But those auxiliary losses seldom endowed any explicit meaning for 2D supervisions. MonoCon [20] made the most out of 2D auxiliary by utilizing up to 5 different 2D supervisions. As for BEV detectors, BEVDepth [15] utilized Li-DAR point clouds to supervise its intermediate depth network. MV-FCOS3D++ [32] introduced perspective supervision for training its image backbone, but the detector itself was supervised by BEV losses alone. SimMOD [41] used 2D auxiliary losses for its monocular proposal head.

Different from previous methods, our method adopted an end-to-end perspective supervision approach without using extra data such as LiDAR point clouds.
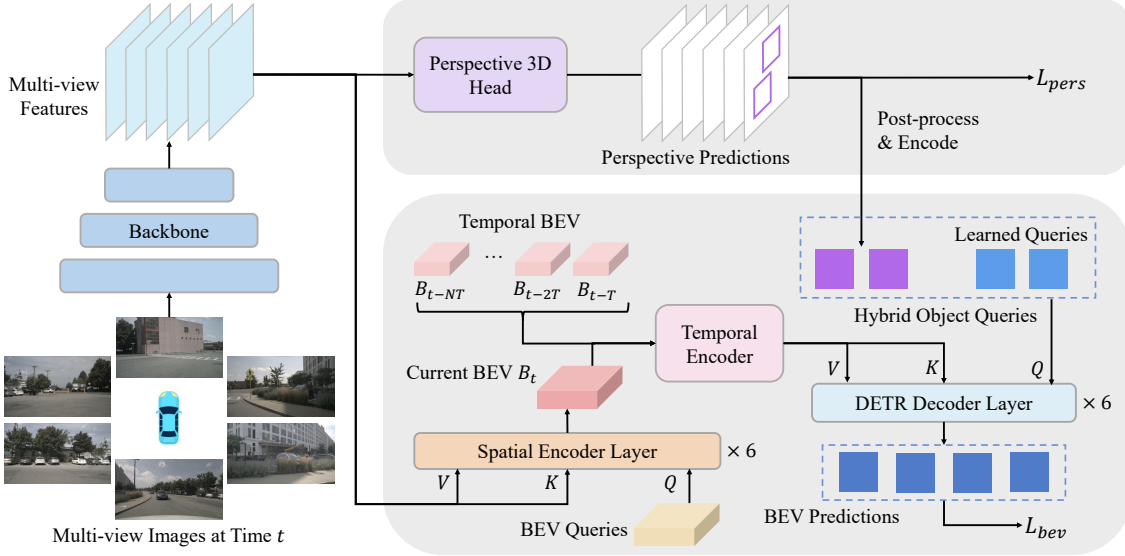
Figure 1. Overall architecture of BEVFormer v2. The image backbone generates features of multi-view images. The perspective 3D head makes perspective predictions which are then encoded as object queries. The BEV head is of encoder-decoder structure. The spatial encoder generates BEV features by aggregating multi-view image features, followed by the temporal encoder that collects history BEV features. The decoder takes hybrid object queries as input and makes the final BEV predictions based on the BEV features. The whole model is trained with the two loss terms of the two detection heads, L_{pers} and L_{bev}.

2.3. Two-stage 3D Object Detector

Although two-stage detectors are common in LiDAR-based 3D object detection [1, 5, 12, 16, 28, 38, 41], their application in camera-based 3D detection is far less well known. MonoDIS [30] used RoIAlign to extract image features from 2D boxes and to regress 3D boxes subsequently. SimMOD [41] employed a monocular 3D head for making proposals and a DETR3D [36] head for the final detection. However, using the same features from the perspective backbone in both stages provides no information gain for the second-stage head. We suppose that this is the main reason why two-stage detectors were far less popular in camera-based 3D detection. Instead, our two-stage detector utilizes features from both perspective and BEV view and thus enjoys information in both image and BEV space.

3. BEVFormer v2

Adapting modern 2D image backbones for BEV recognition without cumbersome depth pre-training could unlock many possibilities for downstream autonomous driving tasks. In this work, we propose BEVformer v2, a two-stage BEV detector that incorporates both BEV and perspective supervision for a hassle-free adoption of image backbones in BEV detection.

3.1. Overall Architecture

As illustrated in Fig. 1, BEVFormer v2 mainly consists of five components: an image backbone, a perspective 3D

detection head, a spatial encoder, a revamped temporal encoder, and a BEV detection head. Compared with the original BEVFormer [17], changes are made for all components except the spatial encoder. Specifically, all image backbones used in BEVFormer v2 are not pre-trained with any autonomous driving datasets or depth estimation datasets. A perspective 3D detection head is introduced to facilitate the adaptation of 2D image backbones and generate object proposals for the BEV detection head. A new temporal BEV encoder is adopted for better incorporating long-term temporal information. The BEV detection head now accepts a hybrid set of object queries as inputs. We combine the first-stage proposals and the learned object queries to form the new hybrid object queries for the second stage.

3.2. Perspective Supervision

We first analyze the problem of the bird's-eye-view models to explain why additional supervision is necessary. As illustrated in Fig. 2, a typical BEV model maintains grid-shaped BEV features, where each grid aggregates 3D information from multi-view image features at corresponding 2D pixels. After that, it predicts the 3D bounding boxes and class labels of the target objects based on the BEV features, and we name this supervision as BEV supervision. Take BEVformer [17] as an example, it consists of an encoder and a decoder, both of which are transformers of multiple layers. The encoder assigns each grid cell with a set of 3D reference points and projects them onto multi-view images as 2D reference points, around which it samples and ag-
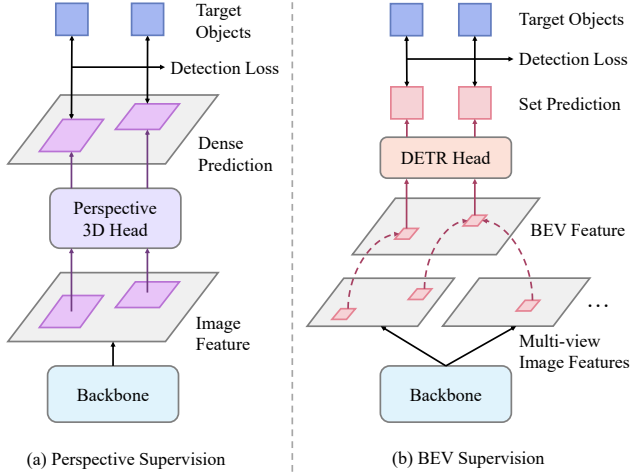
Figure 2. Comparison of perspective (a) and BEV (b) supervision. The supervision signals of the perspective detector are direct to the image feature, while those of the BEV detector are indirect.

gregates image features by spatial cross-attention. The decoder is a Deformable DETR [43] head. The underlying issue of BEV supervision is that it is indirect with respect to the image features. The gradient flow is distorted by the 3D-to-2D projection and the attentive sampling with multi-layer transformers, meaning that the supervision signals are largely separated from the image backbone. Therefore, inconsistency emerges during training that the BEV detection head relies on the 3D information contained in the image features, but it provides insufficient guidance for the backbone on how to encode such information.

Previous BEV methods do not severely suffer from this inconsistency, and they may not even realize this problem. This is because their backbones either have relatively small scales or have been pre-trained on 3D detection tasks with a monocular detection head. As shown in Fig. 2, in contrast to the BEV head, the perspective 3D head makes per-pixel predictions upon the image features, offering much richer supervision signals for adapting 2D image backbones. We define this supervision directly imposed on the image features as perspective supervision. We suppose that perspective supervision explicitly guides the backbone to perceive 3D scenes and extract useful information, e.g., the depths and orientations of the objects, overcoming the drawbacks of BEV supervision, thus is essential when training BEV models with modern image backbones.

3.3. Perspective Loss

As analyzed in the previous session, perspective supervision is the key to optimizing BEV models. In BEVformer v2, we introduce perspective supervision via an auxiliary perspective loss. Specifically, a perspective 3D detection head is built upon the backbone to detect target objects in the perspective view. We adopt an FCOS3D [33]-like detection head, which predicts the center location, size, orientation, and projected center-ness of the 3D bounding boxes. The detection loss of this head, denoted as perspective loss $\mathcal{L}_{pers}$, serves as the complement to the BEV loss $\mathcal{L}_{bev}$, facilitating the optimization of the backbone. The whole model is trained with a total objective

$$\mathcal{L}_{total} = \lambda_{bev}\mathcal{L}_{bev} + \lambda_{pers}\mathcal{L}_{pers}. \qquad (1)$$

3.4. Ravamped Temporal Encoder

BEVFormer uses recurrent temporal self-attention for incorporating historical BEV features. But the temporal encoder falls short of utilizing long-term temporal information, simply increasing the recurrent steps from 4 to 16 yields no extra performance gain.

We redesign the temporal encoder for BEVFormer v2 by a using simple warp and concatenate strategy. Given a BEV feature B_k at a different frame k, we first bi-linearly warp B_k into the current frame as B_k^t according to the reference frame transformation matrix $T_k^t = [\mathbf{R}|\mathbf{t}] \in \mathrm{SE3}$ between frame t and frame k. We then concatenate previous BEV features with the current BEV feature along the channel dimension and employ residual blocks for dimension reduction. To maintain a similar computation complexity as the original design, we use the same number of historical BEV features but increase the sampling interval. Besides benefiting from long-term temporal information, the new temporal encoder also unlocks the possibility of utilizing future BEV features in the offline 3D detection setting.

3.5. Two-stage BEV Detector

Though jointly training two detection heads has provided enough supervision, we obtain two sets of detection results separately from different views. Rather than take the predictions of the BEV head and discard those of the perspective head or heuristically combine two sets of predictions via NMS, we design a novel structure that integrates the two heads into a two-stage predicting pipeline, namely, a two-stage BEV detector. The object decoder in the BEV head, a DETR [3] decoder, uses a set of learned embeddings as object queries, which learns where the target objects possibly locate through training. However, randomly initialized embeddings take a long time to learn appropriate positions. Besides, learned object queries are fixed for all images during inference, which may not be accurate enough since the spatial distribution of objects may vary. To address these issues, the predictions of the perspective head are filtered by post-processing and then fused into the object queries of the decoder, forming a two-stage process. These hybrid object queries provide candidate positions with high scores (probability), making it easier for the BEV head to capture target objects in the second stage. The details of the decoder
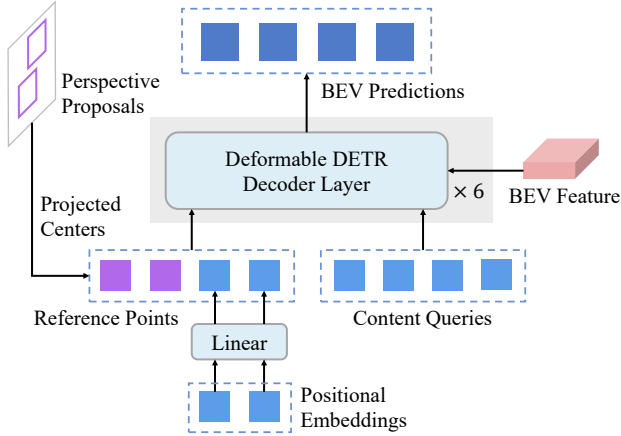
Figure 3. The decoder of the BEV head in BEVFromer v2. The projected centers of the first-stage proposals are used as per-image reference points (purple ones), and they are combined with per-dataset learnded content queries and positional embeddings (blue ones) as hybrid object queries.

with hybrid object queries will be described later. It should be noticed that the first-stage proposals are not necessarily from a perspective detector, e.g., from another BEV detector, but experiments show that only the predictions from the perspective view are helpful for the second-stage BEV head.

3.6. Decoder with Hybrid Object Queries

To fuse the first-stage proposals into the object queries of the second stage, the decoder of the BEV head in BEVformer v2 is modified based on the Deformable DETR [43] decoder used in BEVFormer [17]. The decoder consists of stacked alternated self-attention and cross-attention layers. The cross-attention layer is a deformable attention module [43] that takes the following three elements as input. (1) Content queries, the query features to produce sampling offsets and attention weights. (2) Reference points, the 2D points on the value feature as the sampling reference of each query. (3) Value features, the BEV feature to be attended. In the original BEVFormer [17], the content queries are a set of learned embeddings and reference points are predicted with a linear layer from a set of learned positional embeddings. In BEVformer v2, we obtain proposals from the perspective head and select a part of them via post-processing. As illustrated in Fig. 3, the projected box centers on the BEV plain of the selected proposals are used as per-image reference points and are combined with the per-dataset ones generated from positional embeddings. The per-image reference points directly indicate the possible positions of objects on the BEV plain, making it easier for the decoder to detect target objects. However, a small part of objects may not be detected by the perspective head due to occlusion or appearing at the boundary of two adjacent views. To avoid

missing these objects, we also keep the original per-dataset reference points to capture them by learning a spatial prior.

4. Experiments

4.1. Dataset and Metrics.

The nuScenes 3D detection benchmark [2] consists of 1000 multi-modal videos of roughly 20s duration each, and the key samples are annotated at 2Hz. Each sample consists of images from 6 cameras covering the full 360-degree field of view. The videos are split into 700 for training, 150 for validation, and 150 for testing. The detection task contains 1.4M annotated 3D bounding boxes of 10 object classes. The nuScenes computes the mean average precision (mAP) over four different thresholds using center distance on the ground plane, and it contains five true-positive metrics, namely, ATE, ASE, AOE, AVE, and AAE, for measuring translation, scale, orientation, velocity, and attribute errors, respectively. In addition, it also defines a nuScenes detection score (NDS) by combining the detection accuracy (mAP) with the five true-positive metrics.

4.2. Experimental Settings

We conduct experiments with multiple types of backbones: ResNet [8], DLA [39], VoVNet [13], and InternImage [34]. All the backbones are initialized with the checkpoints pre-trained on the 2D detection task of the COCO dataset [19]. Except for our modification, we follow the default settings of BEVFormer [17] to construct the BEV detection head. In Tab. 1 and Tab. 6, the BEV head utilizes temporal information with the new temporal encoder. For other experiments, we employ the single-frame version that only uses the current frame, like BEVFormer-S [17]. For the perspective 3D detection head, we adopt the implementation in DD3D [26] with camera-aware depth parameterization. The loss weight of perspective loss and BEV loss are set as $\lambda_{bev} = \lambda_{pers} = 1$. We use AdamW [24] optimizer and set the base learning rate as 2e-4 with layer-wise learning rate decay.

4.3. Benchmark Results

We compare our proposed BEVFormer v2 with existing state-of-the-art BEV detectors including BEVFormer [17], PolarFormer [11], PETRv2 [22], BEVDepth [15], and BEVStereo [14]. We report the 3D object detection results on the nuScenes test set in Tab. 1. The V2-99 [13] backbone used by BEVFormer, PolarFormer, BEVDepth, and BEVStereo have been pre-trained on the depth estimation task with extra data and then fine-tuned by DD3D [26] on the nuScenes dataset [2]. On the contrary, the InternImage [34] backbone we employ is initialized with the checkpoint from COCO [19] detection task without any 3D pre-training. InternImage-B has a similar number of parameters

Table 1. 3D detection results on the nuScenes *test* set of BEVFormer v2 and other SoTA methods.[†] indicates that V2-99 [13] was pre-trained on the depth estimation task with extra data [26]. [‡] indicates methods with CBGS which will elongate 1 epoch into 4.5 epochs. We choose to only train BEVFormer v2 for 24 epochs to compare fairly with previous methods.

Method	Backbone	Epoch	Image Size	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
BEVFormer [17]	V2-99[†]	24	900×1600	0.569	0.481	0.582	0.256	0.375	0.378	0.126
PolarFormer [11]	V2-99[†]	24	900×1600	0.572	0.493	0.556	0.256	0.364	0.440	0.127
PETRv2 [22]	GLOM	24	640×1600	0.582	0.490	0.561	0.243	0.361	0.343	0.120
BEVDepth [15]	V2-99[†]	90[‡]	640×1600	0.600	0.503	0.445	0.245	0.378	0.320	0.126
BEVStereo [14]	V2-99[†]	90[‡]	640×1600	0.610	0.525	0.431	0.246	0.358	0.357	0.138
BEVFormer v2	InternImage-B	24	640×1600	0.620	0.540	0.488	0.251	0.335	0.302	0.122
BEVFormer v2	InternImage-XL	24	640×1600	0.634	0.556	0.456	0.248	0.317	0.293	0.123

Table 2. The detection results of 3D detectors with different combinations of view supervision on the nuScenes *val* set. All models are trained without temporal information.

View Supervision	Backbone	Epoch	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
Perspective Only	ResNet-101	48	0.412	0.323	0.737	0.268	0.377	0.943	0.167
BEV Only	ResNet-101	48	0.426	0.355	0.751	0.275	0.429	0.847	0.215
Perspective & BEV	ResNet-101	48	0.451	0.374	0.730	0.270	0.379	0.773	0.205
BEV & BEV	ResNet-101	48	0.428	0.350	0.750	0.279	0.388	0.842	0.210

to V2-99, but better reflects the progress of modern image backbone design. We can observe that BEVFormer v2 with InternImage-B backbone outperforms all existing methods, showing that with the perspective supervision, backbones pre-trained on monocular 3D tasks are no longer necessary. BEVFormer v2 with InternImage-XL outperforms all entries on the nuScenes camera 3D objection leaderboard with 63.4% NDS and 55.6% mAP, surpassing the second-place method BEVStereo by 2.4% NDS and 3.1% mAP. This significant improvement reveals the huge benefit of unleashing the power of modern image backbone for BEV recognition.

4.4. Ablations and Analyses

4.4.1 Effectiveness of Perspective Supervision

To confirm the effectiveness of perspective supervision, we compare 3D detectors with different view supervision combinations in Tab. 2, including (1) Perspective & BEV, the proposed BEVFormer v2, a two-stage detector integrating a perspective head and a BEV head. (2) Perspective Only, the single-stage perspective detector in our model. (3) BEV Only, the single-stage BEV detector in our model without hybrid object queries. (4) BEV & BEV, a two-stage detector with two BEV heads, *i.e.*, replace the perspective head in our model with another BEV head that utilizes BEV features to make proposals for the hybrid object queries.

Compared with the Perspective Only detector, the BEV Only detector achieves better NDS and mAP by leveraging

multi-view images, but its mATE and mAOE are higher, indicating the underlying issues of BEV supervision. Our Perspective & BEV detector achieves the best performance and outperforms BEV Only detector with a margin of 2.5% NDS and 1.9% mAP. Specifically, the mATE, mAOE, and mAVE of Perspective & BEV detector are significantly lower than those of BEV Only detector. This remarkable improvement mainly derives from the following two aspects: (1) Backbones pre-trained on normal vision tasks cannot capture some properties of objects in 3D scenes, including depth, orientation, and velocity, while backbones guided by perspective supervision are capable of extracting information about such properties. (2) Compared to a fixed set of object queries, our hybrid object queries contain the first-stage predictions as reference points, helping the BEV head to locate target objects. To further ensure that the improvement is not brought by the two-stage pipeline, we introduce the BEV & BEV detector for comparison. It turns out that BEV & BEV is on par with BEV Only and is not comparable with Perspective & BEV. Therefore, only constructing the first-stage head and applying auxiliary supervision in the perspective view is helpful for BEV models.

4.4.2 Generalization of Perspective Supervision

The proposed perspective supervision is expected to benefit backbones of different architectures and sizes. We construct BEVFormer v2 on a series of backbones commonly used for 3D detection tasks: ResNet [8], DLA [39], VoVNet [13],

Table 3. The results of perspective supervision with different 2D image backbones on the nuScenes *val* set. 'BEV Only' and 'Perspective & BEV' are the same as Tab. 2. All the backbones are initialized with COCO [19] pretrained weights and all models are trained without temporal information.

Backbone	Epoch	View Supervision	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
ResNet-50	48	BEV Only	0.400	0.327	0.795	0.277	0.479	0.871	0.210
ResNet-50	48	Perspective & BEV	0.428	0.349	0.750	0.276	0.424	0.817	0.193
DLA-34	48	BEV Only	0.403	0.338	0.772	0.279	0.483	0.919	0.206
DLA-34	48	Perspective & BEV	0.435	0.358	0.742	0.274	0.431	0.801	0.186
ResNet-101	48	BEV Only	0.426	0.355	0.751	0.275	0.429	0.847	0.215
ResNet-101	48	Perspective & BEV	0.451	0.374	0.730	0.270	0.379	0.773	0.205
VoVNet-99	48	BEV Only	0.441	0.367	0.734	0.271	0.402	0.815	0.205
VoVNet-99	48	Perspective & BEV	0.467	0.396	0.709	0.274	0.368	0.768	0.196
InternImage-B	48	BEV Only	0.455	0.398	0.712	0.283	0.411	0.826	0.204
InternImage-B	48	Perspective & BEV	0.485	0.417	0.696	0.275	0.354	0.734	0.182

Table 4. Comparing models with BEV supervision only and with both Perspective & BEV supervision under different training epochs. The models are evaluated on the nuScenes *val* set. All models are trained without temporal information.

View Supervision	Backbone	Epoch	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
BEV Only	ResNet-50	24	0.379	0.322	0.803	0.280	0.549	0.954	0.240
		48	0.400	0.327	0.795	0.277	0.479	0.871	0.210
		72	0.410	0.335	0.771	0.280	0.458	0.848	0.216
Perspective & BEV	ResNet-50	24	0.414	0.351	0.732	0.271	0.505	0.899	0.204
		48	0.428	0.349	0.750	0.276	0.424	0.817	0.193
		72	0.428	0.351	0.741	0.279	0.419	0.835	0.196

and InternImage [34]. The results are reported in Tab. 3. Compared to pure BEV detector, BEVForemr v2 (BEV & perspective) boosts NDS by around 3% and mAP by around 2% for all the backbones, manifesting that it generalizes to different architectures and model sizes. We suppose that adding perspective supervision can be a general scheme for training BEV models, especially when adapting large-scale image backbones without any 3D pre-training.

4.4.3 Choice of Training Epochs

We train the BEV Only model and our BEVFormer v2 (BEV & Perspective) for different epochs to see how many the two models take to achieve convergence. Tab. 4 shows that our BEV & Perspective model converges faster than the BEV Only one, confirming that auxiliary perspective loss facilitates the optimization. The BEV Only model obtains marginal improvement if it is trained for more time. But the gap between the two models remains at 72 epochs and may not be eliminated even for longer training, which indicates that the image backbones cannot be well adapted by

BEV supervision alone. According to Tab. 4, training for 48 epochs is enough for our model, and we keep this fixed for other experiments unless otherwise specified.

4.4.4 Choice of Detection Heads

Various types of perspective and BEV detection heads can be used in our BEVFormer v2. We explore several representative methods to choose the best for our model: for the perspective head, the candidates are DD3D [26] and DETR3D [36]; for the BEV head, the candidates are Deformable DETR [43] and Group DETR [4]. DD3D is a single-stage anchor-free perspective head that makes dense per-pixel predictions upon the image feature. DETR3D, on the contrary, uses 3D-to-2D queries to sample image features and to propose sparse set predictions. However, according to our definition, it belongs to perspective supervision since it utilizes image features for the final prediction without generating BEV features, *i.e.*, the loss is directly imposed on the image features. As shown in Tab. 5, DD3D is better than DETR3D for the perspective head, which sup-

Table 5. Comparison of different choices for the perspective head and the BEV head in BEVFormer v2. The models are evaluated on the nuScenes *val* set. All models are trained without temporal information.

Perspective View	BEV View	Backbone	Epoch	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
DD3D	Deformable DETR	ResNet-50	48	0.428	0.349	0.750	0.276	0.424	0.817	0.193
DD3D	Group DETR	ResNet-50	48	0.445	0.353	0.725	0.276	0.366	0.767	0.180
DETR3D	Deformable DETR	ResNet-50	48	0.409	0.335	0.765	0.276	0.469	0.877	0.198
DETR3D	Group DETR	ResNet-50	48	0.423	0.351	0.743	0.279	0.466	0.844	0.201

Table 6. Ablation study of bells and whistles of BEVFormer v2 on the nuScenes *val* set. All models are trained with a ResNet-50 backbone and temporal information. 'Pers', 'IDA', 'Long', and 'Bi' denotes perspective supervision, image-level data augmentation, long temporal interval, and bi-directional temporal encoder, respectively.

Method	Epoch	Pers	IDA	Long	Bi	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
Baseline	24	✓				0.478	0.368	0.709	0.282	0.452	0.427	0.191
Image-level Data Agumentation	24	✓	✓			0.489	0.386	0.690	0.273	0.482	0.395	0.199
Longer Temporal Interval	24	✓	✓	✓		0.498	0.388	0.679	0.276	0.417	0.403	0.189
Bi-directional Temporal Encoder	24	✓	✓	✓	✓	0.529	0.423	0.618	0.273	0.413	0.333	0.181
All but Perspective	24		✓	✓	✓	0.507	0.397	0.636	0.281	0.455	0.356	0.190

ports our analysis in Sec. 3.2. Dense and direct supervision offered by DD3D is helpful for BEV models, while sparse supervision of DETR3D does not overcome the drawbacks of BEV heads. Group DETR head is an extension of Deformable DETR head that utilizes grouped object queries and self-attention within each group. Group DETR achieves better performance for the BEV head, but it costs more computation. Therefore, we employ DD3D head and Group DETR head in Tab. 1 and keep the same Deformable DETR head as BEVformer [17] in other ablations.

4.4.5 Ablations of Bells and Whistles

In Tab. 6, we ablate the bells and whistles employed in our BEVFormer v2 to confirm their contributions to the final result, including (1) Image-level data augmentation (IDA). The images are randomly flipped horizontally. (2) Longer temporal interval. Rather than use continuous frames with an interval of 0.5 seconds in BEVFormer [17], our BEV-Former v2 samples history BEV features with an interval of 2 seconds. (3) Bi-directional Temporal Encoder. For offline 3D detection, the temporal encoder in our BEVFormer v2 can utilize future BEV features. With longer temporal intervals, our model can gather information from more ego positions at different time stamps, which helps estimate the orientation of the objects and results in a much lower mAOE. In the offline 3D detection setting, the bi-directional temporal encoder could provide additional information from future frames and improves the performance of the model by a large margin. We also ablate the perspective supervision in case of applying all bells and whistles. As shown in Tab. 6, perspective supervision boosts NDS by 2.2 % and mAP by 2.6%, which contributes to the major improvement.

5. Conclusion

Existing works have paid much effort into designing and improving the detectors for bird's-eye-view (BEV) recognition models, but they usually get stuck to specific pretrained backbones without further exploration. In this paper, we aim to unleash the full power of modern image backbones on BEV models. We owe the struggle of adapting general 2D image backbones to the optimization problem of the BEV detector. To address this issue, we introduce perspective supervision into the BEV model by adding auxiliary loss from an extra perspective 3D detection head. In addition, we integrate the two detection heads into a two-stage detector, namely, BEVFormer v2. The full-fledged perspective head provides first-stage object proposals, which are encoded into object queries of the BEV head for the second-stage prediction. Extensive experiments verify the effectiveness and generality of our proposed method. The perspective supervision guides 2D image backbones to perceive 3D scenes of autonomous driving and helps the BEV model achieve faster convergence and better performance, and it is suitable for a wide range of backbones. Moreover, we successfully adapt large-scale backbones to BEV-Former v2, achieving new SoTA results on the nuScenes dataset. We suppose that our work paves the way for future researchers to explore better image backbone designs for BEV models.

Limitations. Due to computation and time limitations, we currently do not test our method on more large-scale image backbones. We have finished a preliminary verification of our method on a spectrum of backbones, and we will extend the model sizes in the future.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1090–1099, 2022. 3

[2] 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 multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 5

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision (ECCV)*, pages 213–229. Springer, 2020. 4

[4] Qiang Chen, Xiaokang Chen, Gang Zeng, and Jingdong Wang. Group detr: Fast training convergence with decoupled one-to-many label assignment. *arXiv preprint arXiv:2207.13085*, 2022. 7

[5] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1907–1915, 2017. 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 248–255, 2009. 1

[7] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? *arXiv preprint arXiv:2206.07959*, 2022. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016. 5, 6, 11

[9] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15273–15282, 2021. 1

[10] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2

[11] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformers. *arXiv preprint arXiv:2206.15398*, 2022. 2, 5, 6

[12] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018. 3

[13] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 13906–13915, 2020. 1, 5, 6, 11, 12

[14] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. *arXiv preprint arXiv:2209.10248*, 2022. 2, 5, 6

[15] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022. 2, 5, 6

[16] Zhichao Li, Feng Wang, and Naiyan Wang. Lidar r-cnn: An efficient and universal 3d object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7546–7555, 2021. 3

[17] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision (ECCV)*, pages 1–18. Springer, 2022. 1, 2, 3, 5, 6, 8, 12

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 2980–2988, 2017. 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision (ECCV)*, pages 740–755. Springer, 2014. 5, 7

[20] Xianpeng Liu, Nan Xue, and Tianfu Wu. Learning auxiliary monocular contexts helps monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1810–1818, 2022. 2

[21] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision (ECCV)*, pages 531–548. Springer, 2022. 1, 2

[22] 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. 5, 6

[23] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. 1

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 5

[25] 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*, 5(3):4867–4873, 2020. 1, 2

[26] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3142–3152, 2021. 1, 2, 5, 6, 7, 11, 12

[27] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision (ECCV)*, pages 194–210. Springer, 2020. 1, 2

[28] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 918–927, 2018. 3

[29] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. In *British Machine Vision Conference (BMVC)*, page 285, 2019. 1, 2

[30] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1991–1999, 2019. 2, 3, 11

[31] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 9627–9636, 2019. 2, 11

[32] Tai Wang, Qing Lian, Chenming Zhu, Xinge Zhu, and Wenwei Zhang. Mv-fcos3d++: Multi-view camera-only 4d object detection with pretrained monocular backbones. *arXiv preprint arXiv:2207.12716*, 2022. 2

[33] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 913–922, 2021. 2, 4

[34] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, and Yu Qiao. Internimage: Exploring large-scale vision foundation models with deformable convolutions. *arXiv preprint arXiv:2211.05778*, 2022. 5, 7, 11

[35] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8445–8453, 2019. 1, 2

[36] 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. 3, 7

[37] Zengran Wang, Chen Min, Zheng Ge, Yinhao Li, Zeming Li, Hongyu Yang, and Di Huang. Sts: Surround-view temporal stereo for multi-view 3d detection. *arXiv preprint arXiv:2208.10145*, 2022. 2

[38] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 11784–11793, 2021. 3

[39] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2403–2412, 2018. 5, 6, 11

[40] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3289–3298, 2021. 2

[41] Yunpeng Zhang, Wenzhao Zheng, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. A simple baseline for multi-camera 3d object detection. *arXiv preprint arXiv:2208.10035*, 2022. 2, 3

[42] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13760–13769, 2022. 1, 2

[43] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 4, 5, 7