

# Unleashing HyDRa: Hybrid Fusion, Depth Consistency and Radar for Unified 3D Perception

Philipp Wolters<sup>1,2</sup>    Johannes Gilg<sup>1</sup>    Torben Teepe<sup>1</sup>    Fabian Herzog<sup>1</sup>  
 Anouar Laouiichi<sup>1</sup>    Martin Hofmann<sup>2</sup>    Gerhard Rigoll<sup>1</sup>

<sup>1</sup>Technical University of Munich    <sup>2</sup>FusionRide

philipp.wolters@tum.de

## Abstract

*Low-cost, vision-centric 3D perception systems for autonomous driving have made significant progress in recent years, narrowing the gap to expensive LiDAR-based methods. The primary challenge in becoming a fully reliable alternative lies in robust depth prediction capabilities, as camera-based systems struggle with long detection ranges and adverse lighting and weather conditions. In this work, we introduce HyDRa, a novel camera-radar fusion architecture for diverse 3D perception tasks. Building upon the principles of dense BEV (Bird’s Eye View)-based architectures, HyDRa introduces a hybrid fusion approach to combine the strengths of complementary camera and radar features in two distinct representation spaces. Our Height Association Transformer module leverages radar features already in the perspective view to produce more robust and accurate depth predictions. In the BEV, we refine the initial sparse representation by a Radar-weighted Depth Consistency. HyDRa achieves a new state-of-the-art for camera-radar fusion of 64.2 NDS (+1.8) and 58.4 AMOTA (+1.5) on the public nuScenes dataset. Moreover, our new semantically rich and spatially accurate BEV features can be directly converted into a powerful occupancy representation, beating all previous camera-based methods on the Occ3D benchmark by an impressive 3.7 mIoU. Code and models are available at <https://github.com/phi-wol/hydra>.*

## 1. Introduction

Reliable and affordable 3D perception is an important cornerstone for safe and efficient operation of autonomous vehicles, enabling agents to navigate dynamic environments and complex scenarios. In this rapidly evolving field, deep learning-based Camera-LiDAR fusion has emerged as the de-facto standard for 3D reconstruction from multiple

sensors, with a proven track record in 3D object detection [37, 45, 77], semantic segmentation [12, 45, 49], and object tracking [8, 37].

While delivering highly accurate geometry sensing, LiDAR sensors still pose a high-cost barrier for large-scale deployment. In order to democratize access to autonomous driving systems, LiDAR-free, and vision-centric systems have gained significant traction in recent years and months. Especially with the advent of Bird’s Eye View (BEV) based architectures [56], camera-only methods have achieved remarkable progress [33, 62, 67]. Cameras provide dense and high-resolution semantic information, theoretically enabling the detection of objects at far distances. However, the main practical challenge lies in preserving the spatial information after 2D projection, leading to unstable depth estimation and localization errors [33, 55]. Another difficulty of passive sensors is posed by dynamic objects that are occluded, yet remain relevant for safe decision making.

Still underrepresented in this research domain is the incorporation of radar sensors. Especially camera-radar fusion has been neglected by the academic community [12]. As low-frequency active sensors based on time-of-flight principles, they exhibit a remarkable resilience to adverse weather and lighting conditions, offering metric measurements for perception ranges up to 300 meters [78]. Furthermore, by leveraging the Doppler effect, radar reflections can yield valuable velocity information about moving and occluded objects, underscoring the potential benefits as a camera-complementary sensor. Combining the two sensor types is key to unlocking the full potential of accessible 3D perception systems [60].

Nevertheless, the task of integrating both sensor types into a coherent and robust perception system is a challenging, non-trivial task. Effectively associating radar detections with camera features is complicated by the unique characteristics of the radar data. The absence of elevation information, the sparse nature of the radar returns, the presence of noise from multi-path reflections, the occurrence

of false positives, and the overall high measurement uncertainty make radar data difficult to process and associate with camera features [23, 60]. Consequently, the direct application of state-of-the-art LiDAR-like architectures on inherently sparse radar data has proven to be ineffective [51]. Despite being in point-cloud form, Yan et al. highlight that CMT [74], among the current leading methods on the nuScenes 3D object detection task, degrades in performance when incorporating the radar data into their pipeline. For this reason, we see the need for a significant shift in fusion paradigms when adopting radar. Traditionally, fusion approaches have been heavily biased towards LiDAR, with the bulk of the network’s capacity devoted to processing LiDAR point clouds, and camera data merely serving as a supplementary source of semantic information [19, 31, 65]. Leaving the LiDAR sensor out of the equation, new approaches significantly prioritize extracting dense semantic information from camera feeds and its projection into the BEV grid. The methodology relies heavily on advanced 2D feature extraction techniques [26, 43, 44] along with efficient view transformation and lifting modules [17, 34, 36]. In this new configuration, radar data assumes a role akin to the previous camera-based methods, enhancing the primary modality with sparse metric information and becoming a guidance module [12, 23].

Modern practices have focused on adaptive fusion in the BEV, reducing spatial misalignment during feature aggregation [23, 45]. Despite these improvements, they still struggle with robust depth estimation and rely partly on monocular depth cues for feature transformation [33]. Existing techniques, while innovative, do not fully capitalize on radar’s potential for unified robust depth sensing, often resulting in a dual-stage projection process where each modality stage relies on its own, potentially inconsistent, depth estimation. We argue that to fully unlock the potential, we must move the fusion stage even earlier. In this work, we propose **HyDRa**, a state-of-the-art camera-radar fusion architecture well equipped to tackle a variety of 3D perception tasks. Our contributions and important stepstones to enable LiDAR-free autonomous vehicles and advanced driver assistance systems are the following:

1. We introduce our **Height Association Transformer (HAT)** module and address the limitations of previous BEV-generating depth networks. By creating unified geometry-aware features in the perspective view, we significantly reduce the translation error of existing methods.
2. We upgrade the concept backprojection-as-refinement and establish a **Radar-weighted Depth Consistency (RDC)** for enhancing sparse fusion features in the BEV, tackling the issue of misaligned or unassociated features and occluded objects.
3. We extensively ablate our design choices and show that

each of our contributions significantly improves the performance of our model on the **nuScenes detection and tracking** challenges [2], but the synergy of all components is crucial for achieving best-in-class results.

4. **HyDRa** establishes a pioneering model for camera-radar based 3D semantic occupancy prediction on the Occ3D benchmark [61], pushing the upper bound of state-of-the-art low-cost vision-centric methods.

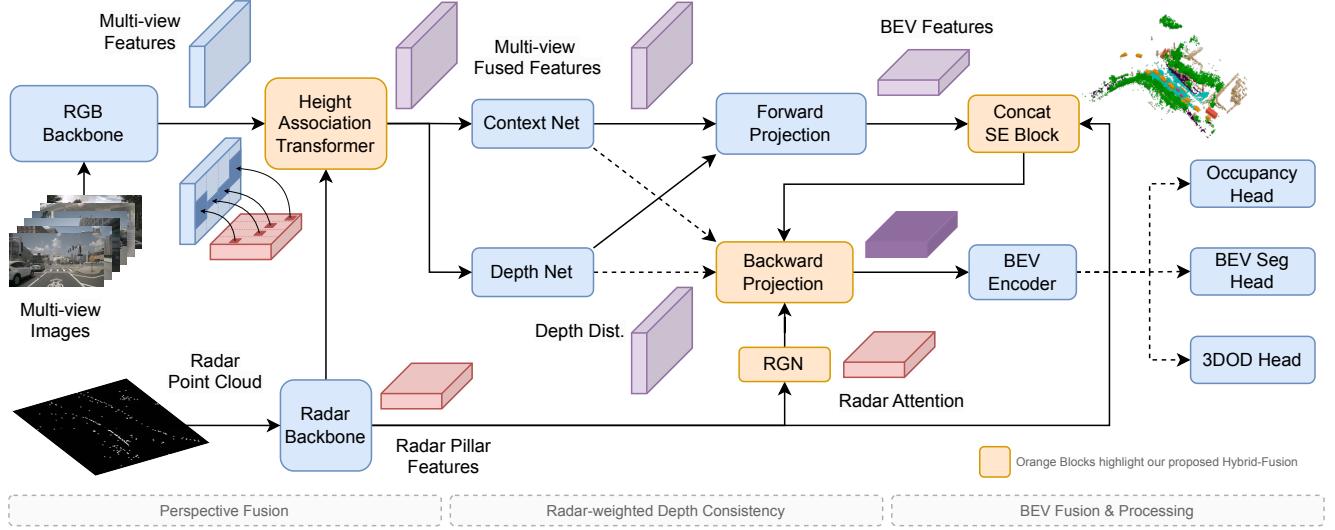
## 2. Related Work

### 2.1. Camera-based Architectures

**Dense BEV-based Architectures.** Shifting from perspective-view 3D object detection [50, 59, 68, 69, 86] to Birds-Eye-View (BEV) based perception [56, 57], the BEV-based architectures have become the new standard for 3D object detection and tracking [28]. With the introduction of differentiable lifting methods, the image-based output of a standard 2D convolutional neural networks [13, 44] or a transformer-based feature extractor [9, 43] is transformed into the more downstream task-friendly BEV ground plane representation [15, 86]. The BEVDet series [16–18] and following work [33, 45] have raised the bar for efficient and powerful forward projection pipeline, predicting an explicit, dense, and pixel-wise depth distribution and pooling the camera-frustum features in the BEV space. BEVDepth [33], SoloFusion [55], and BEVStereo [32] focused on localization by leveraging LiDAR supervision during training, investigating long-term 4D fusion and temporal stereo matching for better depth estimation. BEVFormer [34, 75] has shown that inverse modeling of the view transformation[58] is a valid alternative to the forward projection pipeline; by back-projecting a set of pre-defined BEV grid queries. Spatial deformable cross attention [64, 88] implicitly learns to model the 3D relations by interacting with regions of interest across multiple image feature maps.

FB-BEV [36] combines both approaches into the forward-backward projection paradigm, modeling a more precise projection relationship. It addresses the sparsity and inaccuracy inherent in both projection types by ensuring a Depth Consistency between the two modules. We will build upon this promising concept and extend it to the multi-modal domain by introducing a novel Radar-weighted Depth Consistency and incorporating radar in complementary representation spaces to enforce a more accurate depth estimation.

**Sparse BEV-based Architectures.** Instead of modeling the complete surrounding, a concurrent line of research has established the modeling of the scene as a sparse set of object queries [4, 27, 82], decoding 3D proposals via implicit attention [41, 42, 66, 70]. In addition to long-term



**Figure 1. Overall Architecture of HyDRa:** Modality-specific backbones extract multi-view image and radar pillar features. The features are fused in two representation spaces: Perspective View and BEV-Space. 1. The radar features are associated with the image features by the Height Association Transformer. With the resulting radar-informed dense depth distribution, the forward projection module generates a sparse BEV representation. 2: The splatted semantic BEV features and radar-BEV features are concatenated and fused by a SE block. 3: A depth-aware backward projection refines this representation, guided by radar attention weights before being distributed to task-specific heads.

temporal fusion [55], StreamPETR introduces an object-level temporal mechanism and query denoising [27, 82] to establish an efficient and powerful detection and tracking paradigm [67]. SparseBEV reformulates the design by scale-adaptive self-attention and spatio-temporal sampling inspired by AdaMixer[11]. Without explicit dense BEV feature construction, the model excels in detecting foreground objects, with comparable higher translation errors [55]. These sparse architectures cannot be easily applied for full environmental understanding, such as the drivable area and road boundaries [61, 62]. Hence, we focus on dense BEV-based architectures, as they are best suited for the full stack of planning-oriented end-to-end autonomous driving.

## 2.2. Multi-modal Architectures

Initial works on camera-radar fusion have been restricted by the view disparity between the sensor inputs. GrifNet [21], CRAFT [22] and Centerfusion [51] typically follow Region-of-Interest-pooled late-fusion approaches in the perspective or BEV-view. They focus on producing camera-based detections, refined by the offsets to associated radar point detections. Others directly concatenate view-aligned radar feature maps to high-level image features and feed the fused features into a standard 3D detector head [46, 72]. They rely on simple heuristics-based radar point duplication due to the missing height information. We address these shortcomings by designing an

end-to-end trainable association to image feature height dimension. X3KD [24] investigates a knowledge distillation framework distilling across modalities from a LiDAR-based teacher, whereas BEVGuide [49] and FUTR3D [5] employ BEV query-based feature sampler with a transformer decoder, but none of them leverages the radar at all for view transformation.

In general, fusing the radar in its natural 3D representation form of the BEV grid has proven to be most effective sofar [12, 23, 78]. RCBEV [85] empirically showed that using specialized or heavy point-cloud-processing feature backbones does not bring performance gains - hence why our state-of-the-art architecture leverages efficient pillar-encoding voxelization [25, 29].

The former state-of-the-art in radar-camera fusion, CRN [23] employs the same underlying architecture as BEVDepth [33]. Kim et al. introduce a multi-modal deformable attention mechanism, a BEV-fusion feature operator with a bigger receptive field than simple concatenation [45]. Additionally, they propose a radar-assisted view transformation (RVT) to complement a monocular depth estimation network with a radar-occupancy-based view projection, working in two separate streams. They were the first on nuScenes [2] to show that with an increasing perception range of up to 100m, radar-fusion-based methods outperform LiDAR at far distances [80]. While CRN's RVT is a step in the right direction, it still relies on two parallel but separate view transformations. We propose a more uni-

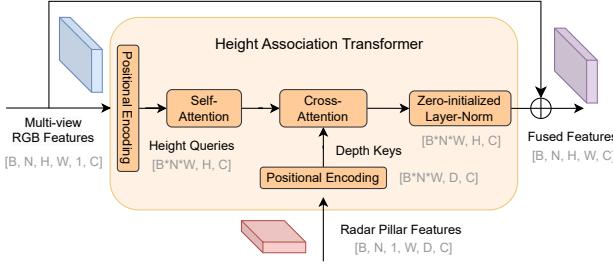


Figure 2. Overview of the Height Association Transformer. The radar fusion module exerts a pushing effect into the BEV.

fied approach to solve this spatial misalignment introduced by monocular depth cues by moving the fusion to a lower level and designing a hybrid fusion.

### 2.3. Occupancy Prediction

Occupancy Prediction is a new task emerging in the context of 3D perception with the goal of predicting the occupancy and semantic class of every voxel in the surrounding 3D environment [61, 62]. It is an important intermediate task in the full stack of planning-oriented end-to-end autonomous driving, positively affecting planning L2 error and collision rate. Semantic occupancy prediction even helps to outperform LiDAR-based planners [15]. Dense architectures (2.1) are well equipped to predict the 3D status of dynamic and static foreground as well as background voxels, as this task is heavily dependent on the quality of depth estimation [35, 48, 53, 81]. Instead of working on the 2D BEV space, the ground plane grid extends to 3D voxels, which are then predicted by 3D convolutional encoders. We show, that a camera-radar fused BEV representation is the best suited for this task, excelling over all previous camera-only methods.

## 3. HyDRa Architecture

We introduce **HyDRa**, a novel camera-radar fusion architecture aiming to reduce the depth prediction error and leading to state-of-the-art performance in 3D object detection and semantic occupancy prediction. The overall architecture is visualized in Figure 1 and consists of the following key components:

- **Modality-specific feature encoder:** Multi-view images are fed into a 2D encoder and converted to high-level feature maps. The radar point cloud is voxelized and encoded to one height level by point pillars [25].
- **Unified Depth Prediction:** Our newly proposed Height Association Transformer leverages cross-attention to associate pillar features (missing height, sparse depth) in the respective camera frustum with image columns (missing

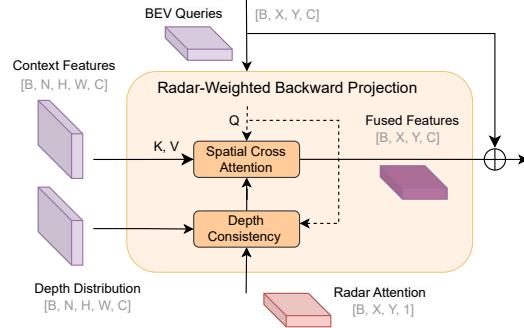


Figure 3. Details of the Radar-weighted Backward Projection. The radar features pull RGB information for refinement.

depth, dense height). The geometry-aware features are added residually and converted into a dense depth distribution.

- **BEV Fusion:** The forward projection module [18] generates the initial BEV representation and is concatenated with the radar pillars channels and fused by a simple Squeeze-and-Excitation layer [14, 37]. Cached history features [55] can be optionally fused.
- **Radar-guided Backward Projection:** The backward projection module [34] refines the initial sparse features, guided by the Radar-weighted Depth Consistency calculated from radar BEV features, its implicitly encoded location and radar-aware depth distribution.
- **Downstream Task Heads:** The fused BEV representation is encoded by additional residual blocks and fed into the respective task head.

### 3.1. Vision-Centric Foundation

In our method we propose an extension to existing vision-centric 3D perception systems. Thus we build upon the recent advancements in camera-based high-performance object detection (BEVDet Series [18, 33]) and the simple but effective concept of fusing radar and image features in the BEV space [12]. The setup follows our baseline method CRN [23]. The image encoder, multi-resolution feature pyramid neck [38], the Context-Net, Depth-Net, and BEV-Encoder [13] follow the same architecture choices as previous work [18, 33].

### 3.2. Height Association Transformer

Our method utilizes the strength of each modality to overcome the main challenge in vision-centric depth prediction. We propose a new transformer-based plug-in module to leverage the complementary radar pillar features (*cf.* Fig. 2). The primary challenge is associating the radar features  $R$  with the respective image features  $F$  in the perspective view. For comparison, CRN’s RVT [23] collapses the image features due to missing information in the height di-

Methods	Input	Backbone	Image Size	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
CenterPoint-P [80]	L	Pillars	-	59.8	49.4	0.320	0.262	0.377	0.334	0.198
CenterPoint-V [80]	L	Voxel	-	65.3	56.9	0.285	0.253	0.323	0.272	0.186
BEVDepth [33]	C	R50	256 × 704	47.5	35.1	0.639	0.267	0.479	0.428	0.198
FB-BEV [36]	C	R50	256 × 704	49.8	37.8	0.620	0.273	0.444	0.374	0.200
SOLOFusion [55]	C	R50	256 × 704	53.4	42.7	0.567	0.274	0.411	0.252	0.188
X3KD [24]	C+R	R50	256 × 704	53.8	42.3	-	-	-	-	-
StreamPETR [67]	C	R50	256 × 704	54.0	43.2	0.581	0.272	0.413	0.295	0.195
SparseBEV [39]	C	R50	256 × 704	54.5	43.2	0.606	0.274	<b>0.387</b>	0.251	0.186
CRN [23]	C+R	R50	256 × 704	56.0	49.0	0.487	0.277	0.542	0.344	0.197
<b>HyDRa</b>	C+R	R50	256 × 704	<b>58.5</b>	<b>49.4</b>	<b>0.463</b>	<b>0.268</b>	0.478	<b>0.227</b>	<b>0.182</b>
MVFusion [72]	C+R	R101	900 × 1600	45.5	38.0	0.675	0.258	0.372	0.833	0.196
FUTR3D [5]	C+R	R101	900 × 1600	50.8	39.9	-	-	-	0.561	-
BEVDepth [33]	C	R101	512 × 1408	53.5	41.2	0.565	0.266	0.358	0.331	0.190
BEVGuide [49]	C+R	R101	224 × 480	53.7	42.1	-	-	-	0.390	-
SparseBEV <sup>†</sup> [39]	C	R101	512 × 1408	59.2	50.1	0.562	0.265	0.321	0.243	0.195
StreamPETR <sup>†</sup> [67]	C	R101	512 × 1408	59.2	50.4	0.569	0.262	<b>0.315</b>	0.257	0.199
CRN [23]	C+R	R101	512 × 1408	59.2	52.5	0.460	0.273	0.443	0.352	<b>0.180</b>
<b>HyDRa</b>	C+R	R101	512 × 1408	<b>61.7</b>	<b>53.6</b>	<b>0.416</b>	<b>0.264</b>	0.407	<b>0.231</b>	0.186

Table 1. **3D Object Detection** on nuScenes val set. ‘L’, ‘C’, and ‘R’ represent LiDAR, Camera, and Radar, respectively. (<sup>†</sup> Backbones benefitted from perspective depth pre-training.)

dimension and uses the sparse encoded radar-only occupancy in the frustum to splat the features in the BEV space. Instead of using a separate view transformation or the noisy depth map as input [46], we aim to learn which part of the downsampled feature map should benefit from the noisy and ambiguous radar encoding with an extended receptive field along the complete image height.

Let  $F \in \mathbb{R}^{B \times N \times H \times W \times C}$  denote 2D image backbone output, while  $R \in \mathbb{R}^{B \times N \times 1 \times W \times C}$  describes the radar pillar features in the respective camera frustum.  $B$  is the batch size,  $N$  the number of camera views,  $H$  the downsampled feature height dimension,  $W$  the downsampled feature width,  $D$  the number of depth bins in the camera frustum and  $C$  the same-size embedding dimension. For efficiency, we query a single column  $w$  of discrete monocular height features  $F_{bnw} \in \mathbb{R}^{H \times 1 \times C}$  for the  $w$ -th respective radar feature keys and values  $R_{bnw} \in \mathbb{R}^{1 \times D \times C}$  on the ground plane. Therefore, we reshape the feature tensors to a new batch size  $B' = B \times N \times W$ . For each ‘sequence’ we apply a learnable positional embedding. Next, we encode the height bins with self-attention and decode and fuse the radar features via cross-attention [64]. A peak in activations on the depth bin plane is compared with promising height proposals, encoding the geometric relationship in the attention, inducing sparse but strong metric cues into the dense feature space. Either modality is missing crucial information that the other sensor can provide. To stabilize the learning, we apply a zero-initialized layer norm [1] to the fusion features, and add them residually to the original image fea-

tures. The resulting feature map  $F'$  is the input for the depth and context network, producing the final more robust depth distribution.

### 3.3. Radar-Weighted Depth Consistency

In our method we tackle two challenges of spatial misalignments and projection inconsistencies when integrating the depth information from multiple sensors. CRN [23] tries to compensate for spatial misalignments with a global receptive field with their Multi-modal Deformable Cross Attention (MDCA) module, which solely operates within the BEV. Instead, we propose concatenating the initial set of lifted and splatted BEV features with the radar pillar channels and refining them with our radar-weighted back-projection pipeline. Extending the back-projection-as-refinement [36] to the multi-modal domain, we enforce the consistency between the two projection spaces on both sides of the view disparity [85]. We upgrade the full cycle of Depth Consistency three-fold to enhance the projection quality, leveraging the synergy of our HAT module and a lightweight radar guidance network (RGN). With a small  $3 \times 3$  convolution followed by a sigmoid, the RGN encodes the radar-only BEV features to additional attention weights  $r$  (cf. Fig. 3).

The overall concept relies on the classic projection of 3D BEV points  $(x, y, z)$  onto 2D image coordinates  $(u, v)$  with projection matrix  $P$ . Every BEV proposal location  $Q_{x,y}$  implicitly encodes a depth value  $d_Q$  for the corresponding

Methods	Input	Backbone	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
PointPillars [25]	L	Pillars	55.0	40.1	0.392	0.269	0.476	0.270	0.102
CenterPoint [80]	L	Voxel	67.3	60.3	0.262	0.239	0.361	0.288	0.136
KPConvPillars [63]	R	Pillars	13.9	4.9	0.823	0.428	0.607	2.081	1.000
CenterFusion [51]	C+R	DLA34	44.9	32.6	0.631	0.261	0.516	0.614	0.115
MVFusion [72]	C+R	V2-99	51.7	45.3	0.569	0.269	0.410	0.781	0.203
CRAFT [22]	C+R	DLA34	52.3	41.1	0.467	0.268	0.456	0.519	<b>0.114</b>
BEVDepth [33]	C	ConvNeXt-B	60.9	52.0	0.445	0.243	0.352	0.347	0.127
SOLOFusion [55]	C	ConvNeXt-B	61.9	54.0	0.453	0.257	0.376	0.276	0.148
BEVFormerV2 [75]	C	InternImage-B	62.0	54.0	0.479	0.248	0.320	0.295	0.123
CRN [23]	C+R	ConvNeXt-B	62.4	<b>57.5</b>	0.416	0.264	0.456	0.365	0.130
StreamPETR [67]	C	V2-99	63.6	55.0	0.479	<b>0.239</b>	<b>0.317</b>	<b>0.241</b>	0.119
SparseBEV [39]	C	V2-99	63.6	55.6	0.485	0.244	0.332	0.246	0.117
<b>HyDRa</b>	C+R	V2-99	<b>64.2</b>	57.4	<b>0.398</b>	0.251	0.423	0.249	0.122

Table 2. **3D Object Detection** on the nuScenes [2] test set. By default, V2-99 [26, 54] is pre-trained for depth estimation, whereas ConvNext-B [44] is initialized from ImageNet-22K [6]. The larger ConvNext Base backbone typically leads to better performance [33].

camera image projection.

$$d \cdot [u \ v \ 1]^T = P \cdot [x \ y \ z \ 1]^T \quad (1)$$

By weighting the cross-attention [88] with the consistency  $w_C$  (scalar product) of the predicted depth distribution  $d_P$  and implicit query depth value  $d_Q$  (converted to a distribution), we enrich the otherwise sparse, misaligned, and especially unmatched single-modal features.

$$w_c = d_P \cdot d_Q \quad (2)$$

Our HAT module tackles the left side of Equation 1 by strengthening the explicit depth (pushing effect [12] of radar). BEV-fusion activates better proposal locations  $Q_{x,y}$  for the right side of Equation 1 (pulling effect [12] of radar). The deformable cross-attention brings them closer together and enforces depth consistency on both sides of the equation. We follow the notation of [36] and adapt to the radar-weighted Spatial Cross-Attention (SCA):

$$\text{SCA}(Q_{x,y}, F) = \sum_{i=1}^{N_c} \sum_{j=1}^{N_{\text{ref}}} \mathcal{F}_d(Q_{x,y}, \mathcal{P}_i(x, y, z_j), F_i) \cdot w_c^{ij} \cdot r_{x,y}, \quad (3)$$

with  $\mathcal{F}_d$  denoting the deformable attention, dynamically sampling features around the projection point  $\mathcal{P}_i(x, y, z_j)$  on the image feature map  $F_i$ , weighted by depth consistency  $w_c$  and radar attention  $r$ . Equation 3 highlights where the complementary hybrid fusion approach is beneficial for a more robust depth prediction. Furthermore, we redesigned the ordering of the modules by moving the backprojection after the temporal fusion to capitalize on partially occluded objects indicated not only by radar reflections but also propagated history features.

### 3.4. Down-stream Tasks

Our final BEV representation can be leveraged for multiple 3D perception tasks, generalizing well for 3D object detection, 3D multi-object tracking (MOT), and 3D semantic occupancy prediction. To decode the features into bounding box characteristics (3D location, 3D size, yaw orientation, and 2D velocity vector), we follow standard practice for dense architectures and use a simple anchor-free center-based head [80]. Similar to [23], we utilize the same box-based tracking-by-detection approach of CenterPoint [80], which relies on a velocity-based greedy distance matching.

Our powerful fused BEV features can be directly converted into a rich semantic occupancy output without the need to forward projecting features into the full 3D voxel cube or a 3D-convolution-based BEV encoder. We use a one-layer  $1 \times 1$  Channel-To-Height convolution [81] to create the final representation for our occupancy head [35]. We simply enlarge the final BEV features  $F \in \mathbb{R}^{B \times X \times Y \times C}$  in the channel dimension  $C$  to  $C' = Z \times C$  and reshape the feature map to the new height dimension  $Z$ , resulting in  $F' \in \mathbb{R}^{B \times X \times Y \times Z \times C'}$ . This new unstacked representation is fed into the low-cost occupancy head [35].

## 4. Experiments

### 4.1. Dataset and Metrics

We perform extensive experiments and ablations on HyDRa using the nuScenes dataset [2] and the Occ3D benchmark [61] and challenge our model with other relevant camera- and radar-based methods on three different tasks.

**nuScenes** The nuScenes dataset [2] is the main large-scale

Methods	Input	Backbone	Image Size	mIoU↑	Bus	Car	Ped.	Truck	MC
MonoScene [3]	C	R101	900 × 1600	6.1	4.9	9.4	3.0	7.2	4.0
TPVFormer [20]	C	R101	900 × 1600	27.8	40.8	45.9	18.9	34.2	20.0
BEVFormerOcc [34]	C	R101	900 × 1600	26.9	40.4	42.4	21.8	30.7	23.9
CTF-Occ [61]	C	R101	900 × 1600	28.5	38.3	42.2	22.7	31.1	24.5
SparseOcc [40]	C	R50	256 × 704	30.9	32.9	43.3	23.4	29.3	23.8
BEVStereo-OCC [18]	C	R50	256 × 704	36.1	42.1	49.6	21.5	37.1	17.4
OctreeOcc [48]	C	R50	256 × 704	37.4	-	-	-	-	-
FlashOcc (M5) [81]	C	R50	256 × 704	39.0	41.6	50.5	23.8	37.4	23.0
RenderOcc [53]	C	R50	256 × 704	39.7	-	-	-	-	-
FB-OCC [35, 36]	C	R50	256 × 704	40.7	45.3	50.5	28.2	37.9	29.0
<b>HyDRa</b>	C+R	R50	256 × 704	<b>44.4</b>	<b>52.3</b>	<b>56.3</b>	<b>35.1</b>	<b>44.1</b>	<b>35.9</b>

Table 3. **3D Occupancy Prediction** on Occ3D based on nuScenes val set (CVPR2023 Occupancy Challenge [61] [62]). In addition to the mIoU based on 16 different classes, we also highlight the IoU of the five dynamic classes car, bus, pedestrian (ped.), truck and motorcycle (MC). An overview of all classes (including drivable area and vegetation etc.) can be found in the Appendix.

research benchmark for camera-radar based 3D perception. The dataset contains 1000 diverse and complex scenarios from the Boston and Singapore urban environments, containing left- and right-handed traffic and adverse conditions like night and rain. Annotated at 2Hz, the dataset provides 6 RGB cameras, 5 mmWave radar sensors, and 1 LiDAR sensor. nuScenes provides a rich set of metrics to assess the quality of 3D perception algorithms: mean Average Precision [10], mean Average Translation, Scale, Orientation, Velocity, and Attribute Error. The NDS is a weighted average of all other metrics. The AMOTA score [71] ranks trackers on the benchmark and evaluates the overall 3D multi object tracking performance.

**Occ3D** The nuScenes Occ3D benchmark [61, 62] evaluates the quality of arbitrarily shaped 3D foreground and background objects. The goal is to predict the complete 3D scene geometry around the ego-car, consisting of a voxelized representation for 18 classes. The benchmark reports the mean Intersection over Union (mIoU) over the number of true positive, false positive, and false negative voxel predictions.

## 4.2. Implementation Details

Referring to the literature, we train our model in three different scaling settings following the best practices of previous work. We set the default backbones to ResNet50 [13] with an input resolution of  $256 \times 704$ , 118 depth categories, and a pillar and BEV grid size of 0.8m, leading to  $128 \times 128$  BEV space. We leverage the AdamW [47] optimizer with a batch size of 64. Scaling up the model backbones and image resolutions to ResNet101 [13] ( $512 \times 1408$ ) and V2-99 [26] ( $640 \times 1600$ ), the BEV space is doubled to a standard of  $256 \times 256$ . For comparison to the latest camera-based methods on the test set, we opt for the smaller and more efficient

Methods	Input	Backbone	AMOTA↑	AMOTP↓	FP↓	FN↓	IDS↓
CenterPoint [80]	L	Voxel	63.8	0.555	18612	22928	760
UVTR [30]	C	V2-99	51.9	1.125	14994	39209	2204
ByteTrackV2 [83]	C	V2-99	56.4	1.005	18939	33531	<b>704</b>
StreamPETR [76]	C	ConvNeXt-B	56.6	0.975	21268	<b>31484</b>	784
CRN [23]	C+R	ConvNeXt-B	56.9	<b>0.809</b>	16822	41093	946
<b>HyDRa</b>	C+R	V2-99	<b>58.4</b>	0.950	<b>13996</b>	32950	791

Table 4. **3D Object Tracking** on nuScenes test set.

V2-99 backbone.

We build upon the code base of the BEVDet Series [18] and in particular employ the efficient implementation of BEVPoolV2 [17] as a forward projection view transformer. We train for 20 epochs following the memory-efficient sequential sampling and settings of [55] without CBGS [87]. This reduces the step-to-step sampling diversity but speeds up the training time significantly. Regarding the backward projection module, we follow the configuration of FB-BEV [36]. To trade off the number of parameters, we half the backbone output and depth net middle channels. Our HAT module is implemented by one layer of vanilla self- and cross-attention working on 16-times downsampled height queries. This leads to sequences of 16, 32, and 40 height bins for the three model sizes. For temporal fusion, we follow [36, 39, 67] and use the last eight history frames. Data augmentations are the same as BEVDet [18] and CRN [23], but no test-time augmentations are used. For a fair evaluation, each model is trained only on the respective task and leverages the same task head and supervision signals as the baseline methods.

## 4.3. Main Results

**3D Object Detection.** In Table 1 we report the results of our HyDRa model on the nuScenes validation set. With a ResNet50 backbone, we achieve a new state-of-the-art of



Figure 4. Qualitative comparison of the semantic occupancy prediction in a challenging night scenario. The top row shows the front-view input cameras. We compare FB-OCC [35] with our proposed HyDRA . While the baseline struggles to distinguish different objects at distance, HyDRA showcases spatial consistency and robustness of the detected cars (orange), truck (red) and pedestrian (blue).

58.5 NDS, surpassing the previous best score of 56.0 NDS by CRN [23] by 2.5 NDS. Scaling the model and image resolution to ResNet101 and  $512 \times 1408$ , we maintain the large margin and increase the NDS to 61.7, outperforming all other camera- and radar-based methods on the val split. We want to highlight the impressive improvement in the mean Average Translation Error (mATE) of 0.416, which can be attributed to our model’s robust depth estimation concept. Our proposed HyDRA surpasses CRN with the more powerful ConvNext-Base[33] backbone by 1.4 NDS due to strengthened translation and velocity estimation, leading to a new state-of-the-art of 64.2 NDS on the nuScenes test set (see Table 2). Utilizing a bigger backbone, e.g., Vit-L [9, 67], our method can surpass voxel-based CenterPoint [80]. Qualitative results are published in the Appendix.

Additionally, we show in Table 5 that HyDRA generalizes on the recent View-of-Delft validation set [52], utilizing a more advanced 4D millimeter-wave radar. HyDRA outperforms the best method by an increase of +4.57 mAP in the entire annotation area and +4.06 mAP in the region of interest (compared to the CRN-inspired LXL model [73]).

**3D Multi Object Tracking.** We summarize the tracking results in Table 4. Benefiting from accurate object localization and high-quality velocity estimation, the center-based tracking performance of HyDRA is competitive with the state-of-the-art methods, leading to the best trade-off between false positives (FP), false negatives (FN), and identity switches (IDS) on the nuScenes test set, once more

narrowing the gap to LiDAR-based CenterPoint.

### 3D Occupancy Prediction.

Dense architectures with strong depth prediction excel in occupancy prediction [61, 62]. Our BEV-based architecture showcases superior performance on the Occ3D benchmark, outperforming all other camera-based methods by a large margin of 3.7 mIoU (see Table 3). It works remarkably well on challenging dynamic objects as visualized in Figure Fig. 4. Complementary camera-radar fusion is the key to unlocking the potential of vision-centric 3D sensing. While sparse representations perform well in detection (*cf.* SparseBEV [39] in Table 2), they underperform in occupancy prediction (*cf.* Sparseocc [40] in Table 3). This emphasizes the importance of a dense representation for full 3D scene understanding.

### 4.4. Ablation Studies

To validate the individual contribution of each of our proposed designs and components, we conducted detailed ablation studies. The ablation studies and baseline training of Tab. 6, Tab. 7, and Tab. 8 are performed in the front-view setting. Input for the model are the three forward-facing cameras and the corresponding radar sensors, which reduces the data and computation size by a factor of two. The ablations are conducted on the nuScenes validation set.

**Effect of the Range View Fusion.** We develop a simple baseline similar to Radiant [46] to show that using radar in a unified depth net brings more performance than separate branches. We extend BEVDepth [33] with an additional

Methods	Input	AP in the Entire Annotated Area (%)				AP in the Region of Interest (%)			
		Car	Pedestrian	Cyclist	mAP	Car	Pedestrian	Cyclist	mAP
PointPillar [25]	R	37.06	35.04	63.44	45.18	70.15	47.22	85.07	67.48
RadarPillarNet [84]	R	39.30	35.10	63.63	46.01	71.65	42.80	83.14	65.86
RCFusion [84]	C+R	41.70	38.95	68.31	49.65	71.87	47.50	88.33	69.23
SMURF [79]	R	42.31	39.09	71.50	50.97	71.74	50.54	86.87	69.72
LXL [73]	C+R	42.33	49.48	77.12	56.31	72.18	58.30	88.31	72.93
<b>HyDRa</b>	C+R	52.83	56.57	73.25	<b>60.88</b>	80.65	62.90	87.43	<b>76.99</b>

Table 5. **3D object detection** results on View-of-Delft [52]. The ROI ( $(x, y, z) | -4m < x < 4m, z < 25m$ ) is the driving corridor close to the ego-vehicle. The IoU thresholds are set to 0.5 for cars and 0.25 for pedestrians and cyclists (official evaluation settings).

Methods	Input	BEVF	4D	Depth	NDS $\uparrow$	mAP $\uparrow$	mATE $\downarrow$	mAVE $\downarrow$
BEVDepth [33]	C	$\times$	$\checkmark$	$\checkmark$	44.4	32.1	0.737	0.344
CRN RVT [23]	C+R	$\times$	$\checkmark$	$\checkmark$	45.6	38.3	0.619	0.510
BEVDepth + RF	C+R	$\times$	$\checkmark$	$\checkmark$	48.1	37.2	0.614	<b>0.284</b>
BEVDepth + HAT	C+R	$\times$	$\checkmark$	$\checkmark$	<b>50.3</b>	<b>39.8</b>	<b>0.586</b>	0.287
CRN [23]	C+R	$\checkmark$	$\checkmark$	$\checkmark$	48.9	39.7	0.616	0.361
RDC	C+R	$\checkmark$	$\checkmark$	$\checkmark$	52.7	40.5	0.534	0.261
<b>HyDRa</b>	C+R	$\checkmark$	$\checkmark$	$\checkmark$	<b>53.6</b>	<b>41.2</b>	<b>0.493</b>	<b>0.257</b>

Table 6. **Ablation of fusion strategies:** RF - Baseline Range-Fusion of projected radar depth maps; CRN with occupancy-based Radar View Transformation (RVT); HyDRa with Radar-weighted Depth Consistency (RDC) and Height Association Transformer (HAT) is our full model. We train and evaluate in the front-view setting.

ResNet18 [13] encoding the radar in the same image view. By projecting radar points into the image plane, we create a range image of the radar point cloud, converting 3D points into a pixel location. Instead of RGB channels, radar channels are used, with every non-radar pixel set to zero. This pseudo-image is input to a smaller ResNet18 encoder, and each output stage is concatenated to the respective RGB encoder output.

We show in Table 6 that this unified approach already outperforms the RVT of CRN. We indicate, that the velocity estimation is also benefitting from a perspective fusion, compared to pure radar-occupancy-guided transformation of the semantic features. Processing the mostly sparse stripe of radar points with 2D convolutions is neither effective nor efficient. To overcome the limitations of the naive baseline, we propose a more sophisticated approach, the HAT module, which is a novel way of fusing radar and camera features in the perspective view and is capable of significantly improving the depth estimation of camera-based models. Compared to related work, we do not have to rely on heuristic height extensions [46, 51] or neglect it at all [12] by collapsing image features [23].

**Effect of Radar Weighted Depth Consistency.** Radar-informed depth consistency is a crucial concept of our architecture, showing significant and consistent improvements over the baseline architecture of CRN. Especially

Methods	Input	4D	Depth	NDS $\uparrow$	mAP $\uparrow$	mATE $\downarrow$	mAVE $\downarrow$
BEVDepth [33]	C	$\times$	$\times$	33.3	26.7	0.803	0.930
CRN [23]	C+R	$\times$	$\times$	42.5	<b>35.0</b>	0.607	0.584
<b>HyDRa</b>	C+R	$\times$	$\times$	<b>45.8</b>	34.7	<b>0.569</b>	<b>0.482</b>

Table 7. Comparison of BEVDepth, CRN and HyDRa without temporal fusion or depth supervision. We train and evaluate in the front-view setting.

Exp.	Setting	Input	Radar Attn.	NDS $\uparrow$	mAP $\uparrow$	mATE $\downarrow$	mAVE $\downarrow$
A	BP $\rightarrow$ 4D $\rightarrow$ RF	C+R	$\times$	52.0	41.1	0.551	0.261
B	RF $\rightarrow$ BP $\rightarrow$ 4D	C+R	$\checkmark$	52.7	41.2	0.527	0.281
C	RF $\rightarrow$ 4D $\rightarrow$ BP	C+R	$\checkmark$	<b>53.6</b>	<b>41.2</b>	<b>0.493</b>	<b>0.257</b>

Table 8. Ablation of component order (BP: Backprojection, 4D: Temporal Fusion, RF: Radar BEV-Fusion). Radar Attention for Backprojection is not available in setting A. This reports our full model, including HAT and RDC. We train and evaluate in the front-view setting.

adding the HAT module into this paradigm leverages synergistic effects, leading to strong metric depth understanding (*cf.* Tab. 6). Without temporal or depth supervision, as ablated in Table 7, HyDRa is the superior concept for increased depth and velocity sensing, showing a decent improvement over the camera-only baseline. This could be important for time-critical situations e.g. collision avoidance [78].

**Order Matters.** In Table 8, we show that the order of the fusion matters. The radar BEV features should be fused before any 4D aggregation and backward projection. Enriching the queries with radar and temporal information, poses more meaningful cues for refinement, without additional computational cost as the modules are working on the same BEV channel dimensions.

#### 4.5. Limitations

In future endeavors, we want to explore the potential of object-level motion modeling [67] within our dense paradigm, as the ego-motion compensation [16] does not account for highly dynamic objects. Moreover, the orientation poses a bottleneck as sparse radar points only give a

rough estimate of the object’s location. Therefore, 4D imaging radars pose an interesting research opportunity [7, 52].

## 5. Conclusion

We introduce **HyDRa**, the new state-of-the-art hybrid fusion paradigm, excelling in various 3D perception tasks and showing a promising path for future research in the field of radar-based detection. HyDRa beats the former state-of-the-art in camera-radar fusion by a clear margin of 1.8 NDS and naturally generalizes to the first radar-enhanced occupancy prediction model, improving upon the best camera baseline by 3.7 mIoU. With this line of work, we want to contribute to safer autonomous driving due to better handling of low-visible objects, more robust depth estimation, and accurate velocity estimation.

## References

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5
- [2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liang, Qiang Xu, Anush Krishnan, Yu Pan, Giacomo 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*, 2020. 2, 3, 6
- [3] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 7
- [4] 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*, pages 213–229. Springer, 2020. 2
- [5] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 5
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6
- [7] Fangqiang Ding, Andras Palffy, Dariu M. Gavrila, and Chris Xiaoxuan Lu. Hidden gems: 4d radar scene flow learning using cross-modal supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9340–9349, 2023. 10
- [8] Shuxiao Ding, Eike Rehder, Lukas Schneider, Marius Cordts, and Juergen Gall. 3dmotformer: Graph transformer for online 3d multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9784–9794, 2023. 1
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 8
- [10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 7
- [11] Ziteng Gao, Limin Wang, Bing Han, and Sheng Guo. Adamixer: A fast-converging query-based object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5364–5373, 2022. 3
- [12] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2759–2765, 2023. 1, 2, 3, 4, 6, 9
- [13] 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*, pages 770–778, 2016. 2, 4, 7, 9
- [14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 4
- [15] Yihan Hu, Jiazhai Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhui Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 2, 4
- [16] Junjie Huang and Guan Huang. BEVDet4D: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 2, 9
- [17] Junjie Huang and Guan Huang. Bevpoolv2: A cutting-edge implementation of bevdet toward deployment. *arXiv preprint arXiv:2211.17111*, 2022. 2, 7
- [18] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. In *arXiv preprint arXiv:2112.11790*, 2021. 2, 4, 7
- [19] Junjie Huang, Yun Ye, Zhujin Liang, Yi Shan, and Dalong Du. Detecting as labeling: Rethinking lidar-camera fusion in 3d object detection. *arXiv preprint arXiv:2311.07152*, 2023. 2
- [20] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 7
- [21] Youngseok Kim, Jun Won Choi, and Dongsuk Kum. Grif net: Gated region of interest fusion network for robust 3d object detection from radar point cloud and monocular image. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10857–10864. IEEE, 2020. 3
- [22] Youngseok Kim, Sanmin Kim, Jun Won Choi, and Dongsuk Kum. CRAFT: Camera-Radar 3D Object Detection with Spatio-Contextual Fusion Transformer. In *AAAI*, 2023. 3, 6

- [23] Youngseok Kim, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, and Dongsuk Kum. Crn: Camera radar net for accurate, robust, efficient 3d perception. In *ICCV*, 2023. [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#)
- [24] Marvin Klingner, Shubhankar Borse, Varun Ravi Kumar, Behnaz Rezaei, Venkatraman Narayanan, Senthil Yogamani, and Fatih Porikli. X3kd: Knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection. In *CVPR*, 2023. [3](#), [5](#)
- [25] 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. [3](#), [4](#), [6](#), [9](#)
- [26] Youngwan Lee, Joong-won Hwang, Sangrok Lee, Yuseok Bae, and Jongyoul Park. An energy and gpu-computation efficient backbone network for real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. [2](#), [6](#), [7](#)
- [27] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022. [2](#), [3](#)
- [28] Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhui Wang, Lewei Lu, Huijie Wang, Jia Zeng, Zhiqi Li, Jiazhi Yang, Hanming Deng, et al. Delving into the devils of bird’s-eye-view perception: A review, evaluation and recipe. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. [2](#)
- [29] Jinyu Li, Chenxu Luo, and Xiaodong Yang. Pillarnext: Rethinking network designs for 3d object detection in lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17567–17576, 2023. [3](#)
- [30] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *NeurIPS*, 35, 2022. [7](#)
- [31] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Ji-quan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17182–17191, 2022. [2](#)
- [32] Yinzhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1486–1494, 2023. [2](#)
- [33] Yinzhao 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. In *AAAI*, 2023. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [8](#), [9](#)
- [34] Zhiqi Li, Wenhui 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. In *ECCV*, 2022. [2](#), [4](#), [7](#)
- [35] Zhiqi Li, Zhidong Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. [4](#), [6](#), [7](#), [8](#)
- [36] Zhiqi Li, Zhidong Yu, Wenhui Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. In *ICCV*, 2023. [2](#), [5](#), [6](#), [7](#)
- [37] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35, 2022. [1](#), [4](#)
- [38] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. [4](#)
- [39] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *ICCV*, 2023. [5](#), [6](#), [7](#), [8](#)
- [40] Haisong Liu, Haiguang Wang, Yang Chen, Zetong Yang, Jia Zeng, Li Chen, and Limin Wang. Fully sparse 3d panoptic occupancy prediction. *arXiv preprint arXiv:2312.17118*, 2023. [7](#), [8](#)
- [41] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *ECCV*, pages 531–548, 2022. [2](#)
- [42] Yingfei Liu, Junjie Yan, Fan Jia, Shuaolin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023. [2](#)
- [43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. [2](#)
- [44] 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*, pages 11976–11986, 2022. [2](#), [6](#)
- [45] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023. [1](#), [2](#), [3](#)
- [46] Yunfei Long, Abhinav Kumar, Daniel Morris, Xiaoming Liu, Marcos Castro, and Punarjay Chakravarty. Radiant: Radar-image association network for 3d object detection. In *AAAI*, 2023. [3](#), [5](#), [8](#), [9](#)
- [47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. [7](#)
- [48] Yuhang Lu, Xinge Zhu, Tai Wang, and Yuexin Ma. Octreeocc: Efficient and multi-granularity occupancy predic-

- tion using octree queries. *arXiv preprint arXiv:2312.03774*, 2023. 4, 7
- [49] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Be-guided multi-modality fusion for driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 3, 5
- [50] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017. 2
- [51] Ramin Nabati and Hairong Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1527–1536, 2021. 2, 3, 6, 9
- [52] Andras Palfy, Ewoud Pool, Srimannarayana Baratam, Julian FP Kooij, and Dariu M Gavrila. Multi-class road user detection with 3+1d radar in the view-of-delft dataset. *IEEE Robotics and Automation Letters*, 7(2):4961–4968, 2022. 8, 9, 10
- [53] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 4, 7
- [54] 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*, pages 3142–3152, 2021. 6
- [55] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. In *ICLR*, 2023. 1, 2, 3, 4, 5, 6, 7
- [56] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pages 194–210, 2020. 1, 2
- [57] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021. 2
- [58] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. *arXiv preprint arXiv:1811.08188*, 2018. 2
- [59] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kuntschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1991–1999, 2019. 2
- [60] Arvind Srivastav and Soumyajit Mandal. Radars for autonomous driving: A review of deep learning methods and challenges. *arXiv preprint arXiv:2306.09304*, 2023. 1, 2
- [61] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 2024. 2, 3, 4, 6, 7, 8
- [62] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hamming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *ICCV*, 2023. 1, 3, 4, 7, 8
- [63] Michael Ulrich, Sascha Braun, Daniel Köhler, Daniel Niederlöhrner, Florian Faion, Claudius Gläser, and Holger Blume. Improved orientation estimation and detection with hybrid object detection networks for automotive radar. In *Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 111–117, 2022. 6
- [64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 5
- [65] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020. 2
- [66] Shihao Wang, Xiaohui Jiang, and Ying Li. Focal-petr: Embracing foreground for efficient multi-camera 3d object detection. *IEEE Transactions on Intelligent Vehicles*, 2023. 2
- [67] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 3, 5, 6, 7, 8, 9
- [68] 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*, pages 913–922, 2021. 2
- [69] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 2
- [70] 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. 2
- [71] Xinshuo Weng and Kris Kitani. A baseline for 3d multi-object tracking. *arXiv preprint arXiv:1907.03961*, 1(2):6, 2019. 7
- [72] Zizhang Wu, Guilian Chen, Yuanzhu Gan, Lei Wang, and Jian Pu. Mvfusion: Multi-view 3d object detection with semantic-aligned radar and camera fusion. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 3, 5, 6
- [73] Weiyi Xiong, Jianan Liu, Tao Huang, Qing-Long Han, Yuxuan Xia, and Bing Zhu. Lxl: Lidar excluded lean 3d object detection with 4d imaging radar and camera fusion. *IEEE Transactions on Intelligent Vehicles*, 2023. 8, 9
- [74] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In

- Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18268–18278, 2023. 2
- [75] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird’s-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 6
- [76] Jinrong Yang, En Yu, Zeming Li, Xiaoping Li, and Wenbing Tao. Quality matters: Embracing quality clues for robust 3d multi-object tracking. *arXiv preprint arXiv:2208.10976*, 2022. 7
- [77] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. *Advances in Neural Information Processing Systems*, 35:1992–2005, 2022. 1
- [78] Shanliang Yao, Runwei Guan, Xiaoyu Huang, Zhuoxiao Li, Xiangyu Sha, Yong Yue, Eng Gee Lim, Hyungjoon Seo, Ka Lok Man, Xiaohui Zhu, and Yutao Yue. Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review, 2023. 1, 3, 9
- [79] Shanliang Yao, Runwei Guan, Xiaoyu Huang, Zhuoxiao Li, Xiangyu Sha, Yong Yue, Eng Gee Lim, Hyungjoon Seo, Ka Lok Man, Xiaohui Zhu, et al. Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review. *IEEE Transactions on Intelligent Vehicles*, 2023. 9
- [80] 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*, pages 11784–11793, 2021. 3, 5, 6, 7, 8
- [81] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023. 4, 6, 7
- [82] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2023. 2, 3
- [83] Yifu Zhang, Xinggang Wang, Xiaoqing Ye, Wei Zhang, Jincheng Lu, Xiao Tan, Errui Ding, Peize Sun, and Jingdong Wang. Bytetrackv2: 2d and 3d multi-object tracking by associating every detection box. *arXiv preprint arXiv:2303.15334*, 2023. 7
- [84] Lianqing Zheng, Sen Li, Bin Tan, Long Yang, Sihan Chen, Libo Huang, Jie Bai, Xichan Zhu, and Zhixiong Ma. Rcfusion: Fusing 4d radar and camera with bird’s-eye view features for 3d object detection. *IEEE Transactions on Instrumentation and Measurement*, 2023. 9
- [85] Taohua Zhou, Junjie Chen, Yining Shi, Kun Jiang, Mengmeng Yang, and Diange Yang. Bridging the view disparity between radar and camera features for multi-modal fusion 3d object detection. *IEEE Transactions on Intelligent Vehicles (IEEE Trans. Intell. Veh.)*, 2023. 3, 5
- [86] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2
- [87] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. 7
- [88] 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*, 2021. 2, 6