JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

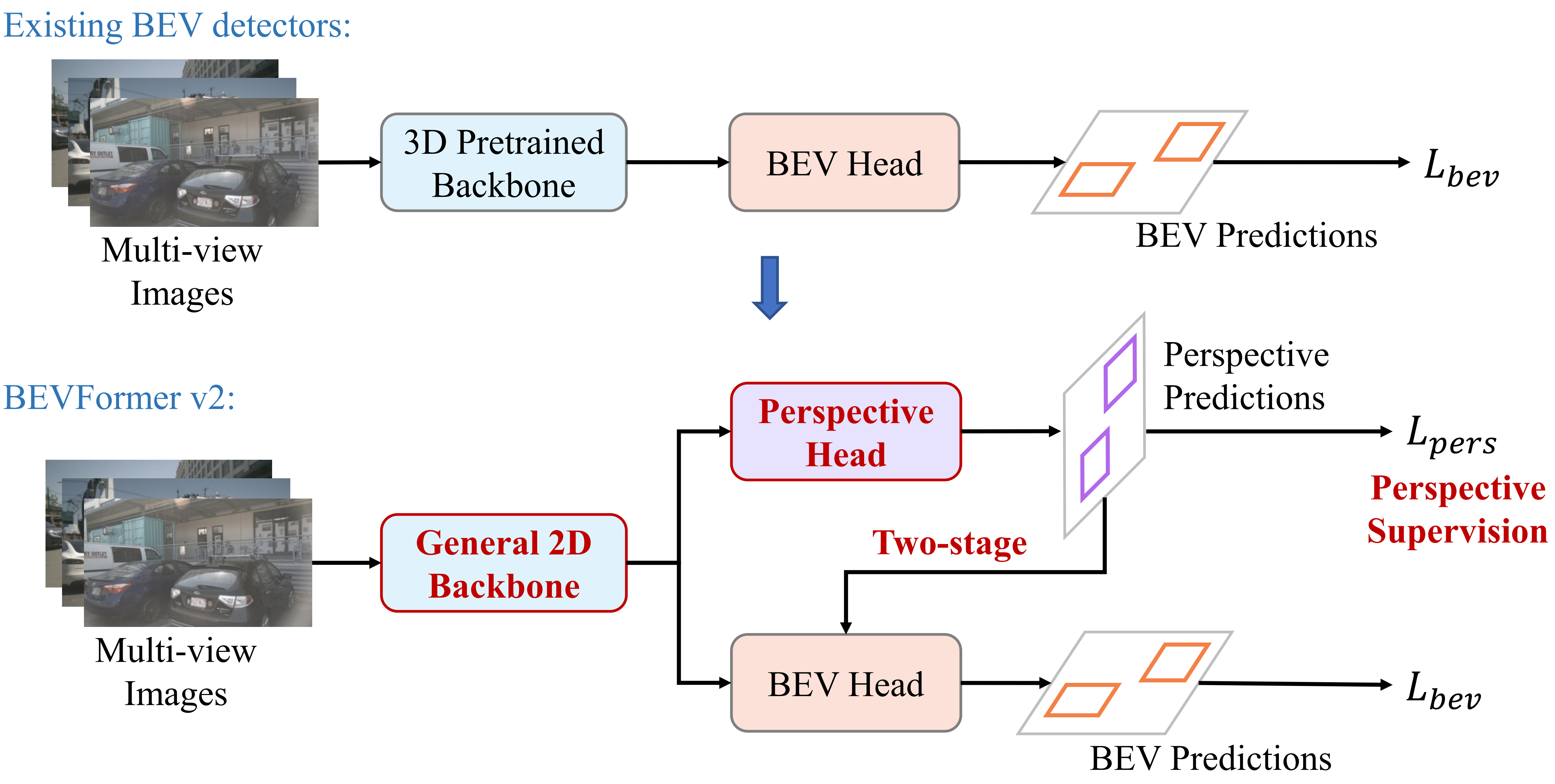
BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision

Chenyu Yang*, Yuntao Chen*, Hao Tian*, Chenxin Tao, Xizhou Zhu,
Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao,
Lewei Lu, Jie Zhou, Jifeng Dai✉

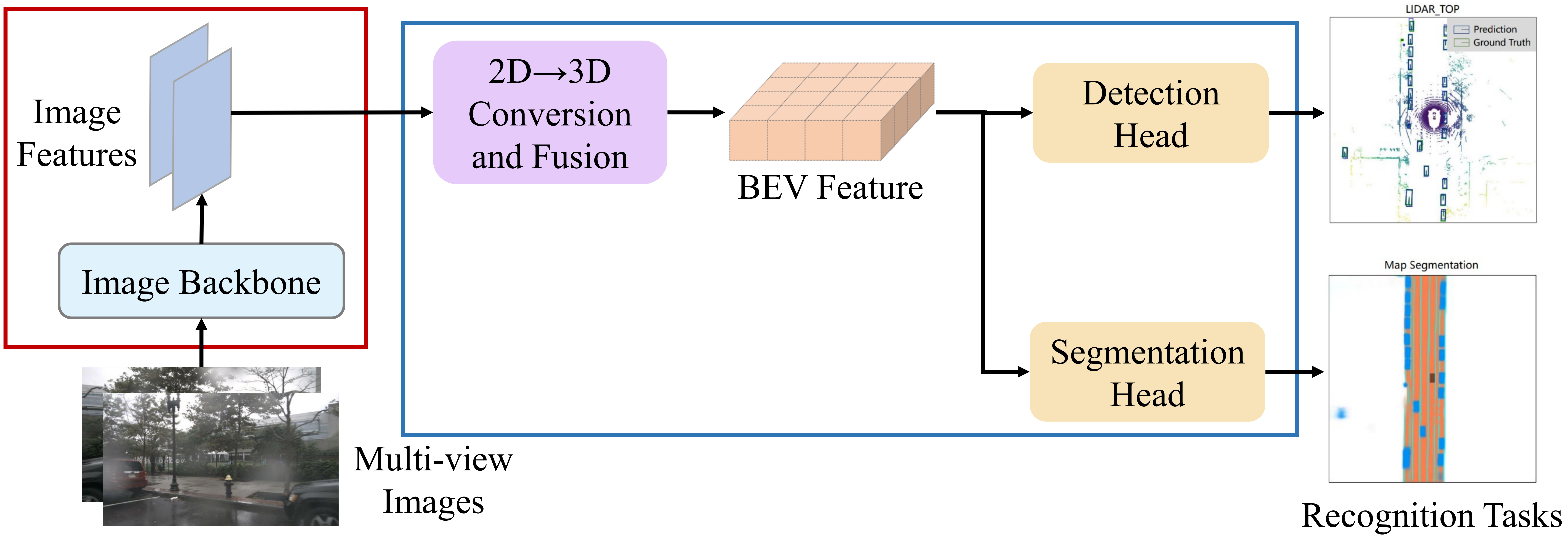
Poster ID: THU-AM-129

清華大學 Tsinghua University
中国科学院香港创新研究院 人工智能与机器人创新中心
商汤 sensetime
中国科学院自动化研究所 INSTITUTE OF AUTOMATION CHINESE ACADEMY OF SCIENCES
上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

Overview

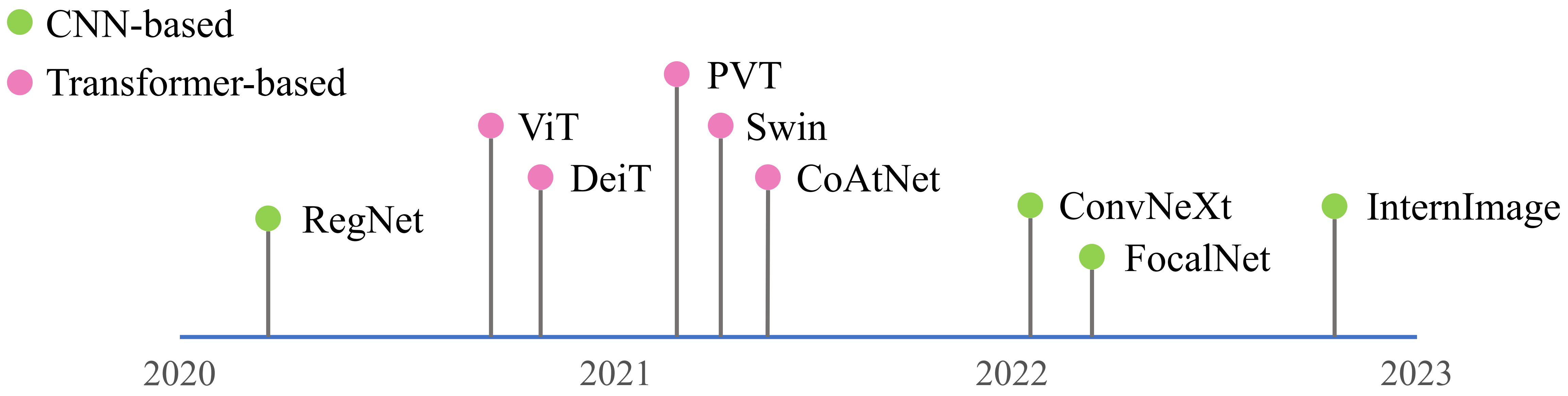

Existing BEV detectors:
Multi-view Images
3D Pretrained Backbone
BEV Head
BEV Predictions
L_{bev}
BEVFormer v2:
Multi-view Images
General 2D Backbone
Perspective Head
Perspective Predictions
L_{pers}
Perspective Supervision
Two-stage
BEV Head
BEV Predictions
L_{bev}

Background: Bird's-eye-view Recognition



- BEV recognition**:** construct **unified BEV feature** for recognition tasks
- Most existing works focus on how to **construct** and **utilize BEV feature**
- Our work focuses on providing better **image features from the backbone**

Background: Modern Image Backbones



Novel architectures and large scales

High performance on 2D recognition tasks

⬇

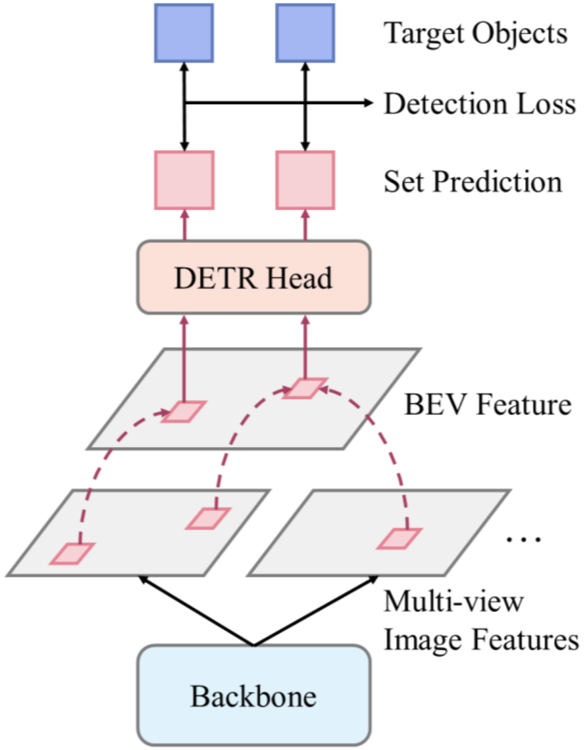
Unleash their power for BEV recognition

Problem:

Trained on general 2D vision tasks

Lack of specific 3D knowledge

Adapt 2D Backbone to BEV Recognition



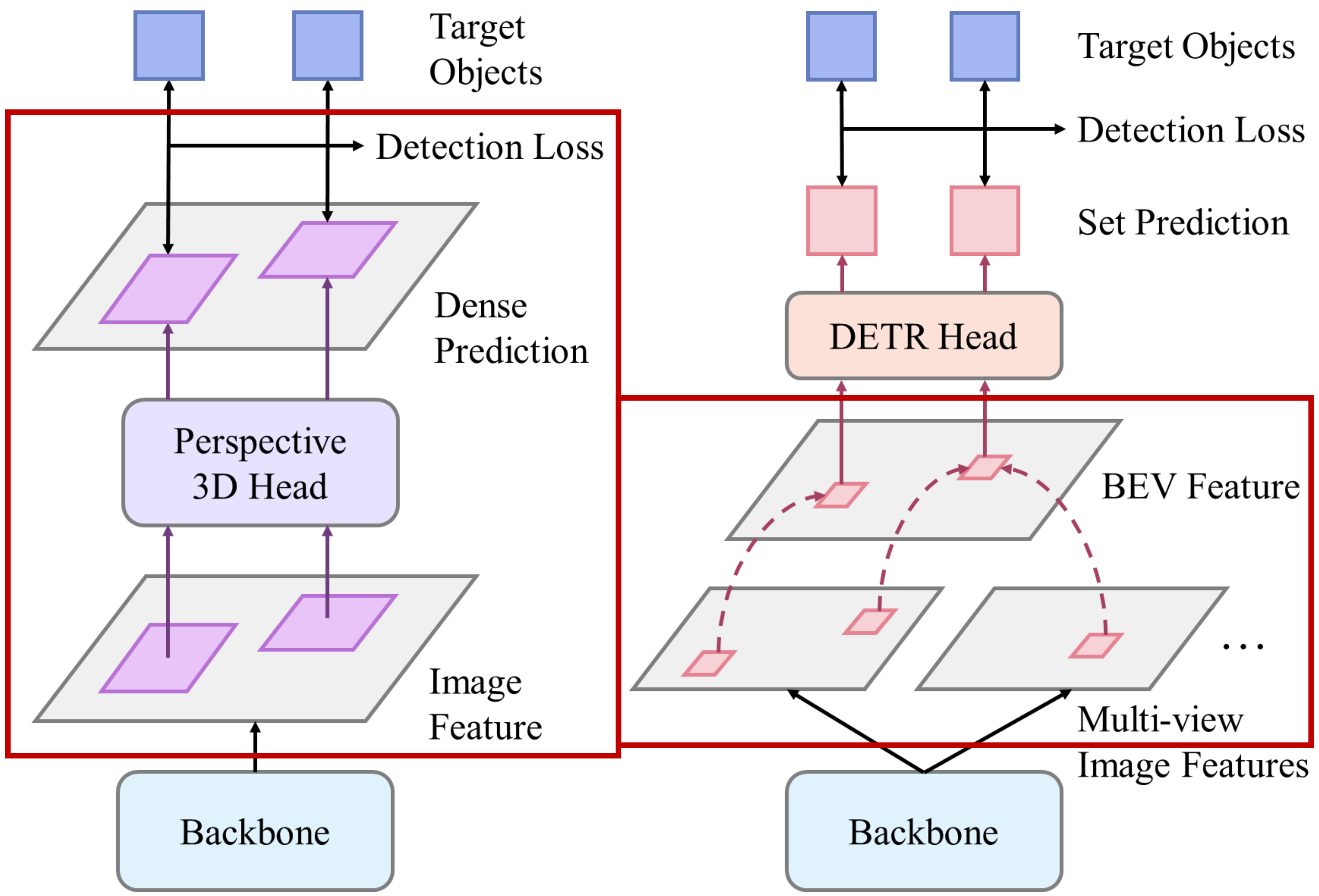
Challenge 1:

Domain gap between natural images and autonomous driving scenes

Challenge 2:

Optimization problem of BEV detectors: complex structure, indirect supervision

BEV Supervision vs. Perspective Supervision



(a) Perspective Supervision (b) BEV Supervision

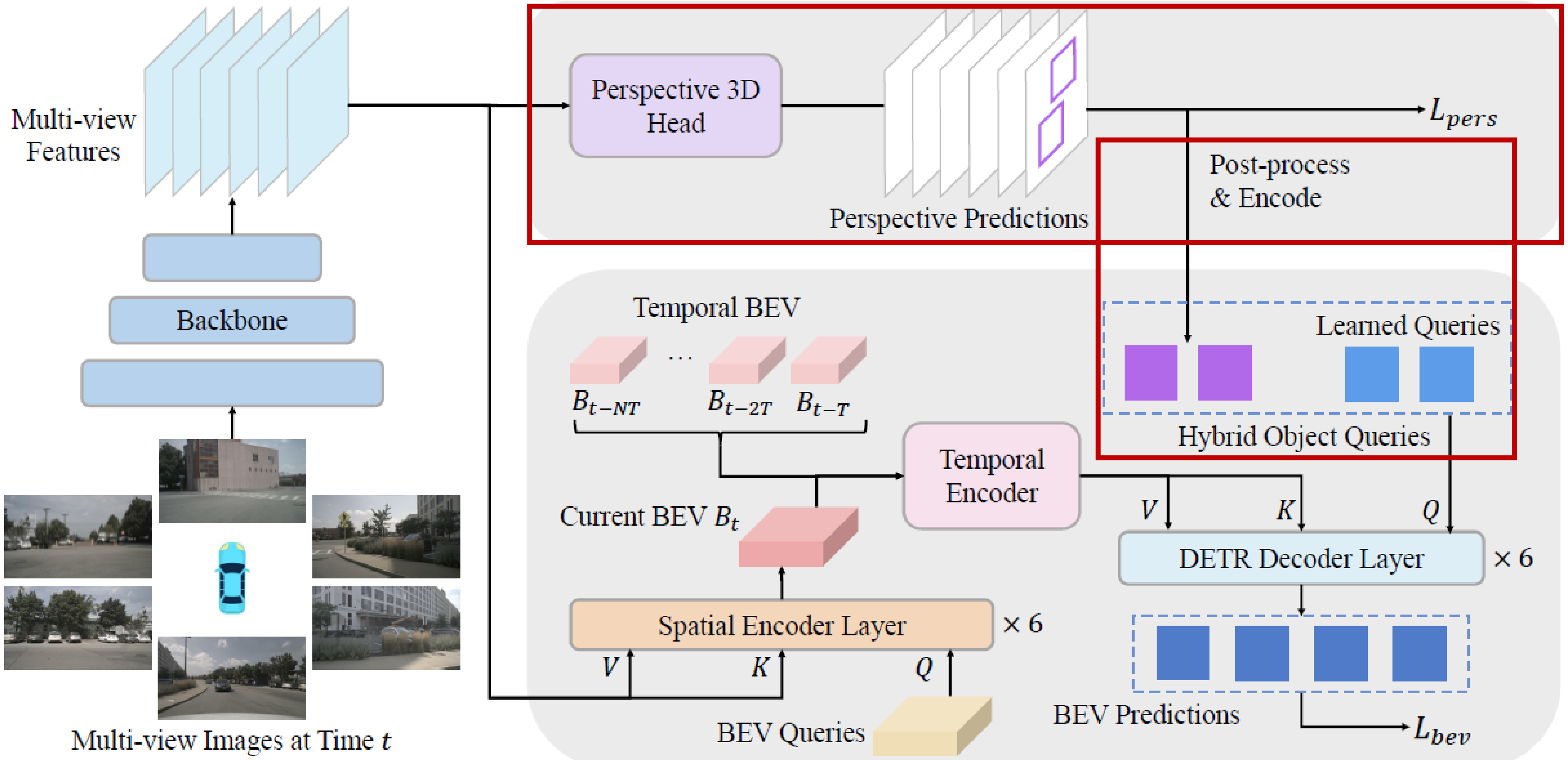
BEV supervision:

- **View transformation** and **attentive sampling** from image view to BEV
- Supervision signal **indirect** to the backbone

Perspective supervision:

- Per-pixel prediction upon the image feature
- **Direct** and **explicit** supervision to the backbone

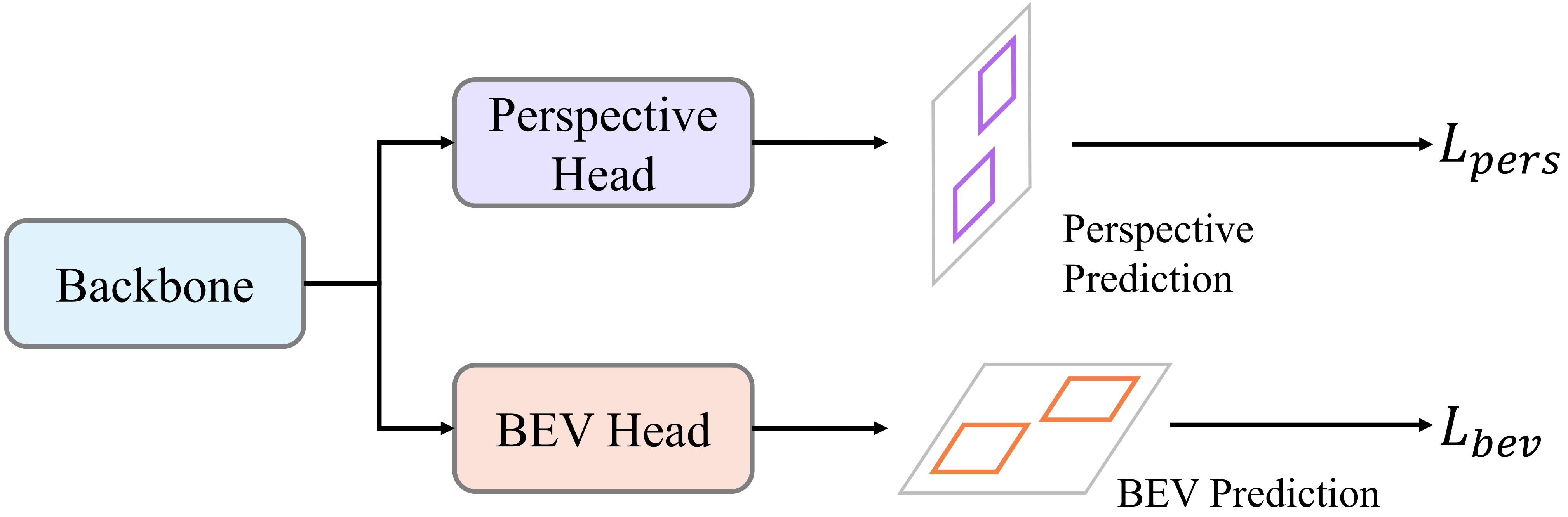
BEVFormer v2: Overall Archtecture



Key designs:

- Auxiliary loss for perspective supervision
- Two-stage pipeline with hybrid object query

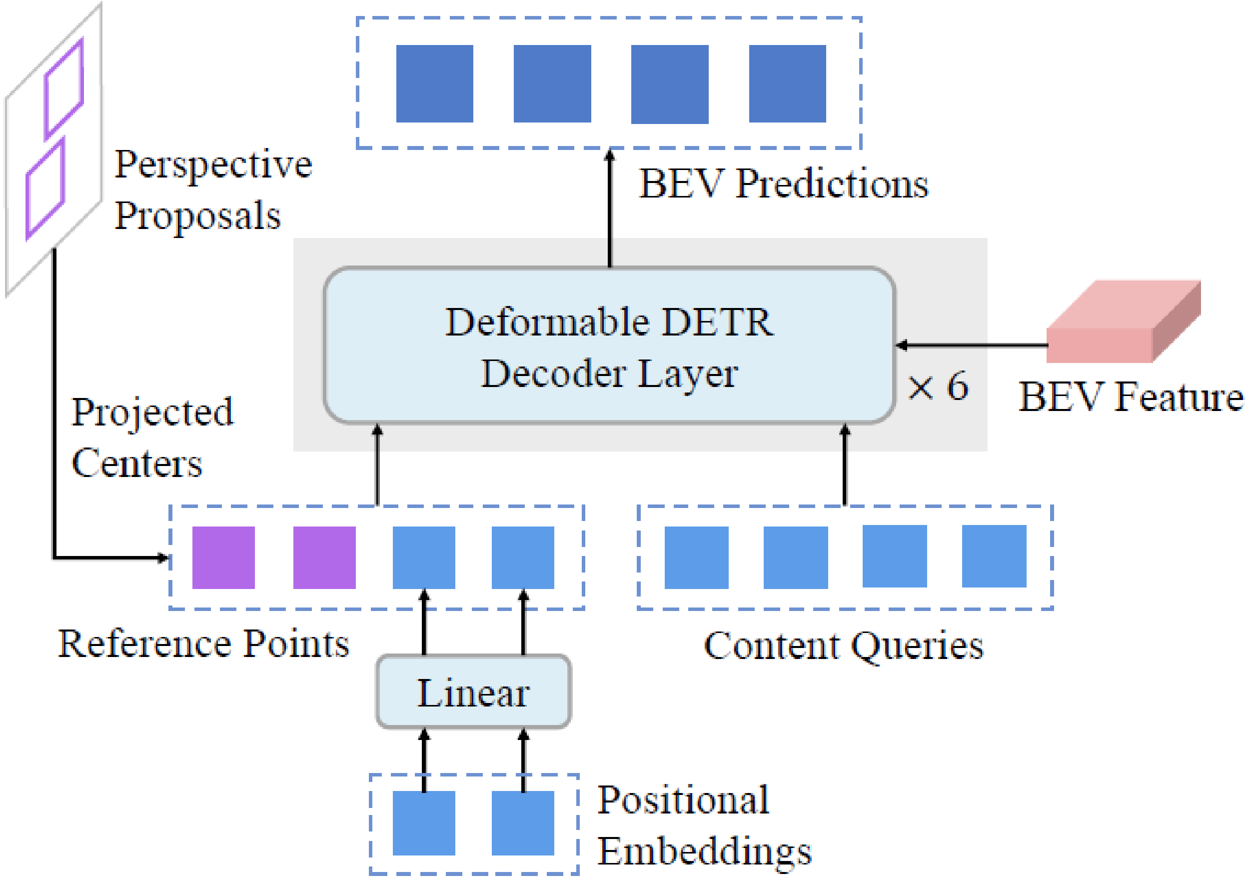
Auxiliary Loss for Perspective Supervision



- A **perspective head** and a **BEV head** on the backbone
- Joint training: predict the objects in two different views and compute losses
- A **auxiliary perspective loss** term for BEV model:

$$\mathcal{L}_{total} = \lambda_{bev}\mathcal{L}_{bev} + \lambda_{pers}\mathcal{L}_{pers}$$

Two-Stage Pipeline with Hybrid Object Query



Two Stage Pipeline:

1. First-stage perspective head makes predictions
2. Gather and post-process multi-view proposals
3. Take **projected centers** as **reference point**
4. Combine with learnable queries as **hybrid object queries**
5. Second-stage BEV head makes predictions

Exp: Performance on nuScenes

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 [27]. ‡ 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 [23]	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

- **nuScenes *test* set :** NDS **63.4** vs. **61.0**, mAP **55.6** vs. **52.5**
- **3D pretraining is not necessary**: outperform existing works with InternImage-B (size similar to V2-99, no 3D pretraining)

Exp: Ablation of Perspective Supervision

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 [20] 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

Generalization: **~3.0** NDS and **~2.0** mAP improvement for all backbones

Exp: Ablation of Perspective Supervision

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 [20] 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

Better perceive 3D scenes: errors of translation (**mATE**), orientation (**mAOE**), and velocity (**mAVE**) are significantly lower