# 2D and 3D Recognition with Transformers

Jingdong Wang
Chief Scientist for Computer Vision, Baidu
IEEE/IAPR Fellow, ACM Distinguished Member
https://jingdongwang2017.github.io/

2023.06.18

# ViT: Vision Transformer

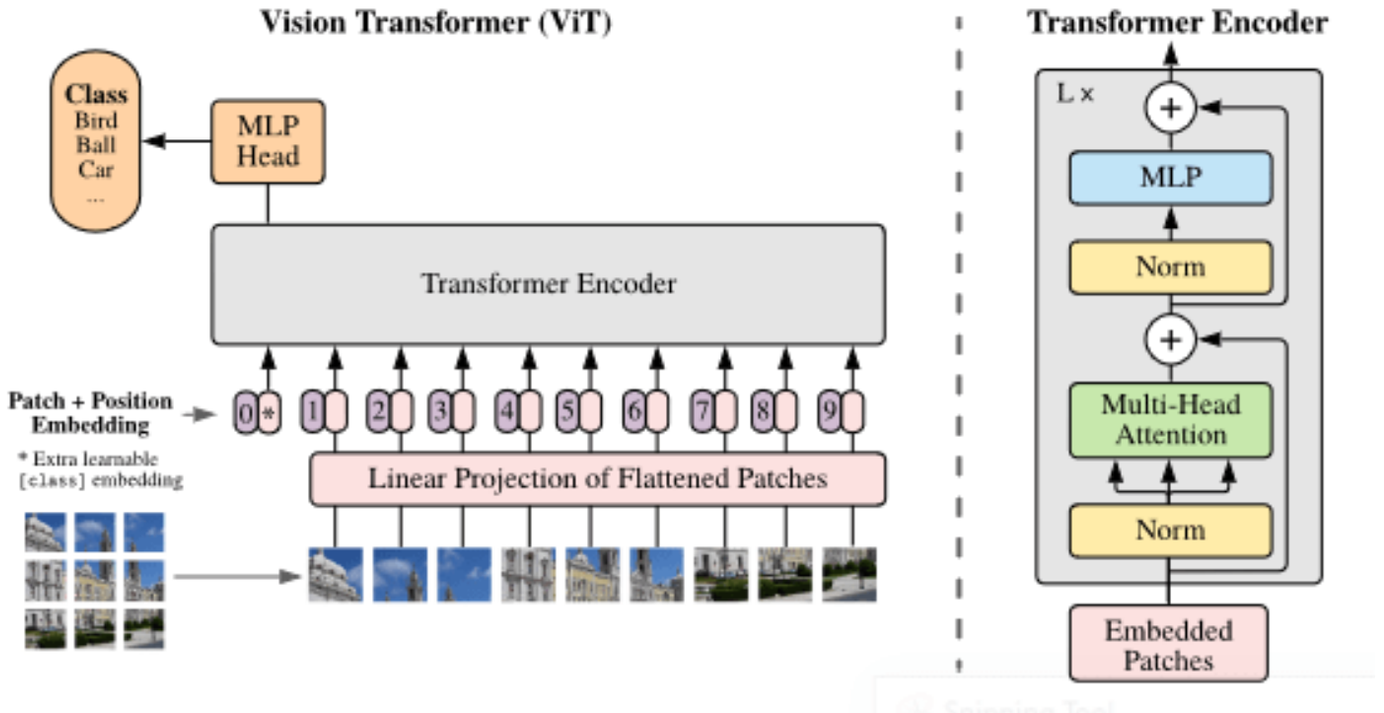



Figure from Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

# Transformer for Recognition

## Transformer encoder

- ViT
- DeiT
- Swin
- …

## Transformer decoder

- Semantic segmentation
- Object detection
- Multi-view 3D detection
- …

Encoder → Decoder → Result

# Transformer Encoder for Visual Recognition

## Vision Transformer: Attention

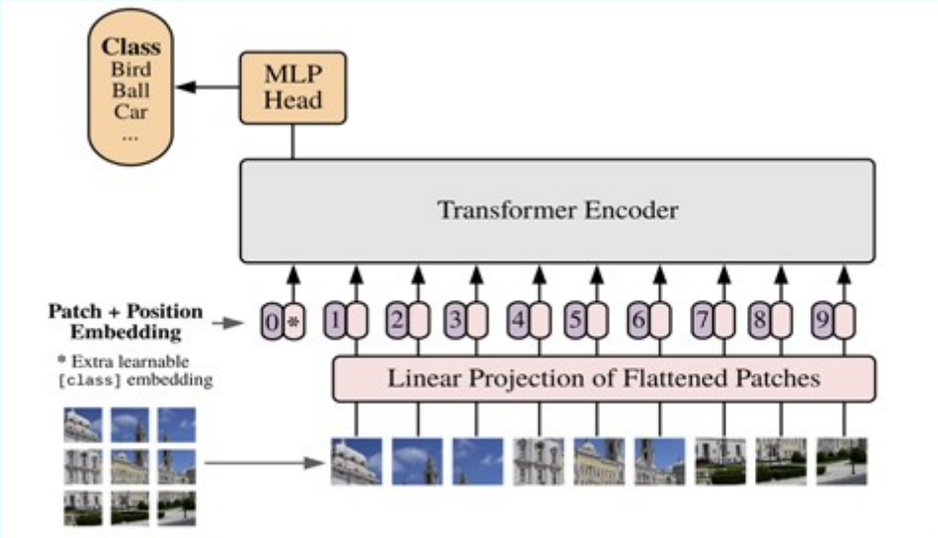
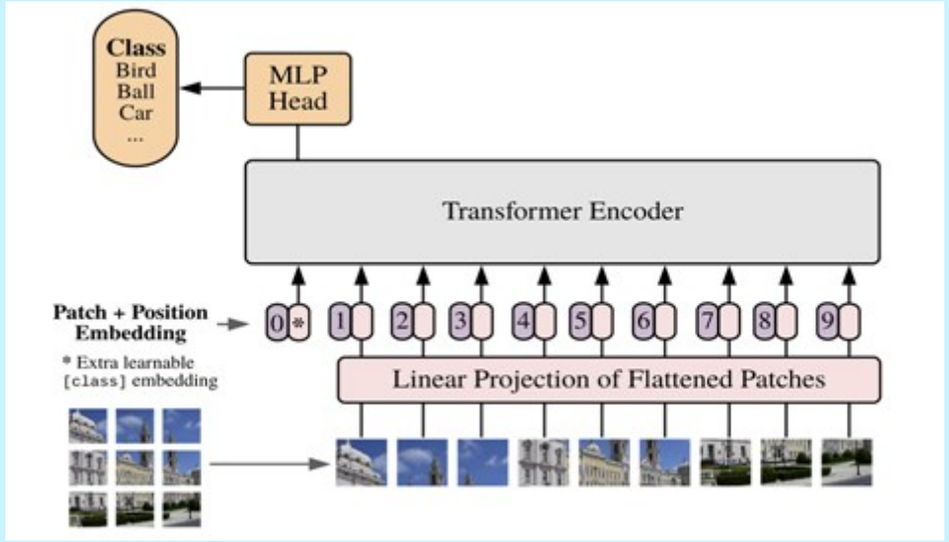
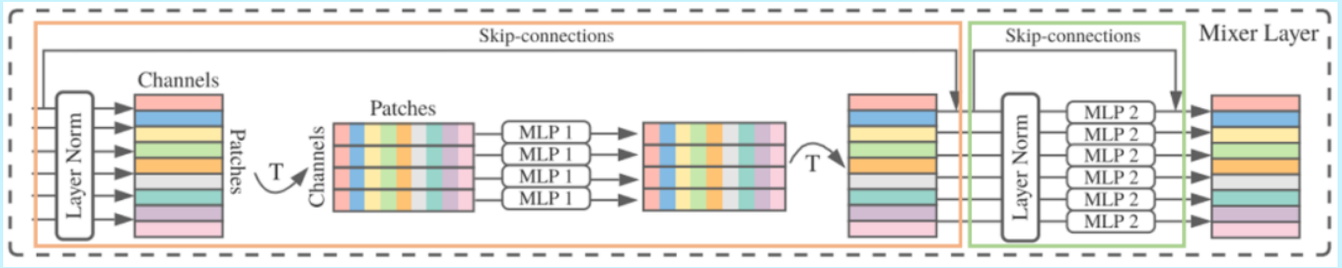



Figure from Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

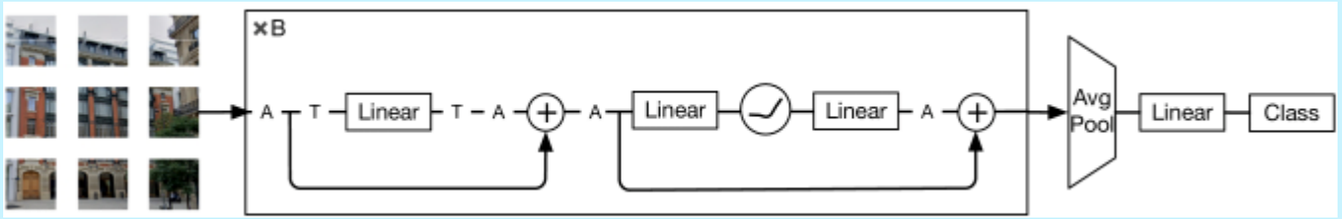
## Vision Transformer: Attention




## MLP-Mixer




## ResMLP




## Swin: local attention

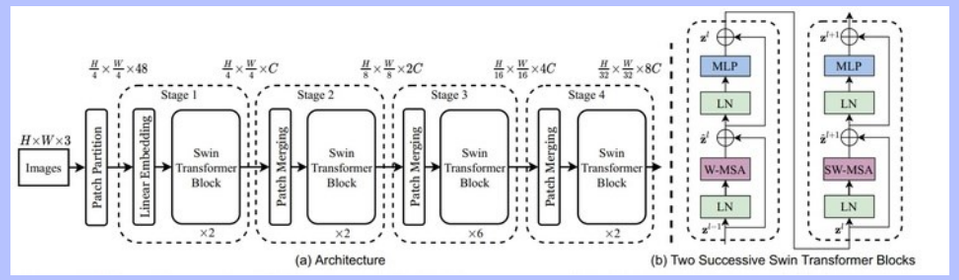
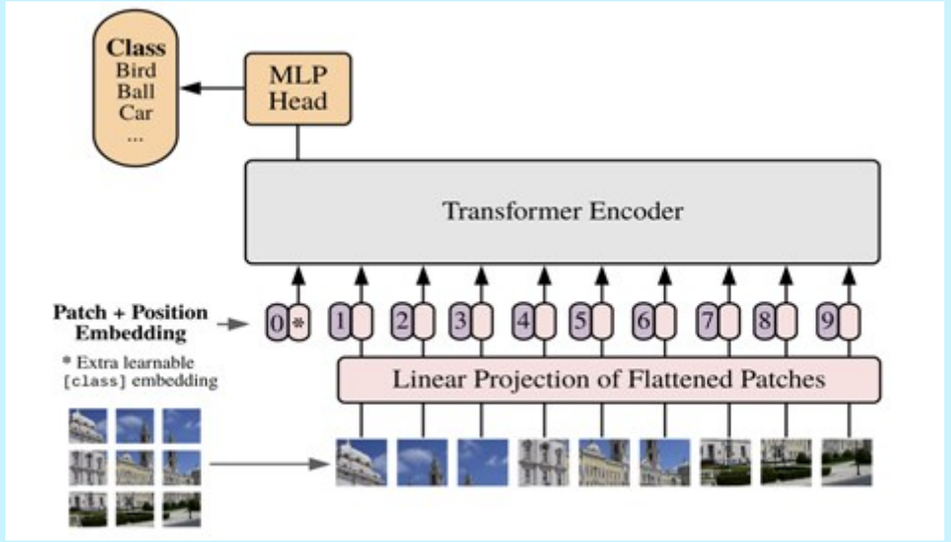
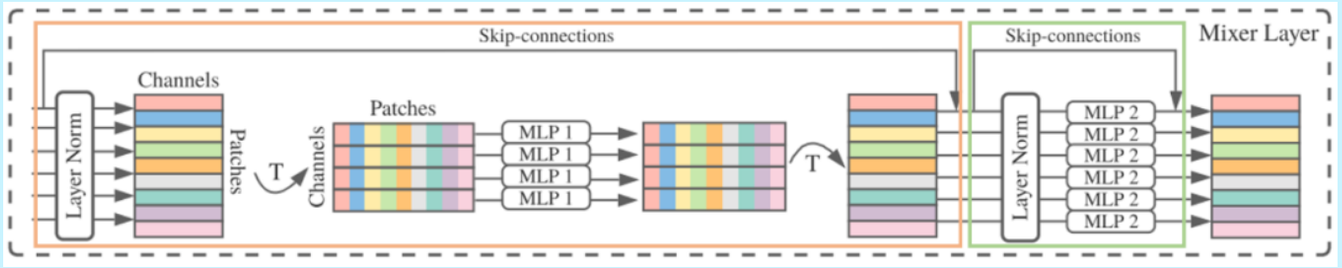



Figure from Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

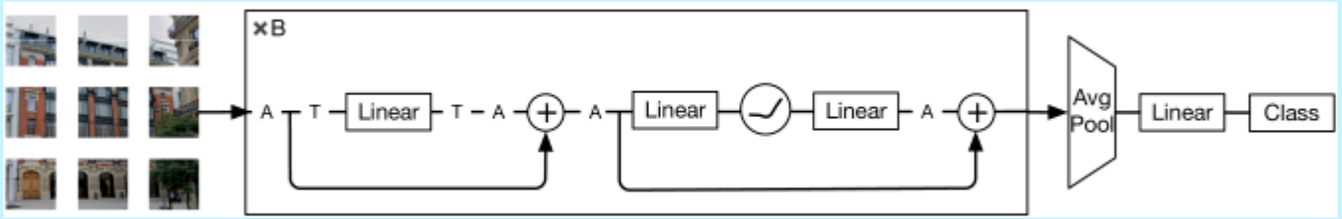
## Vision Transformer: Attention



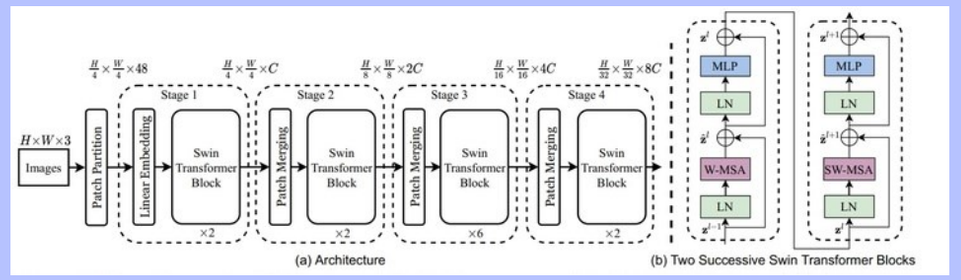

## MLP-Mixer



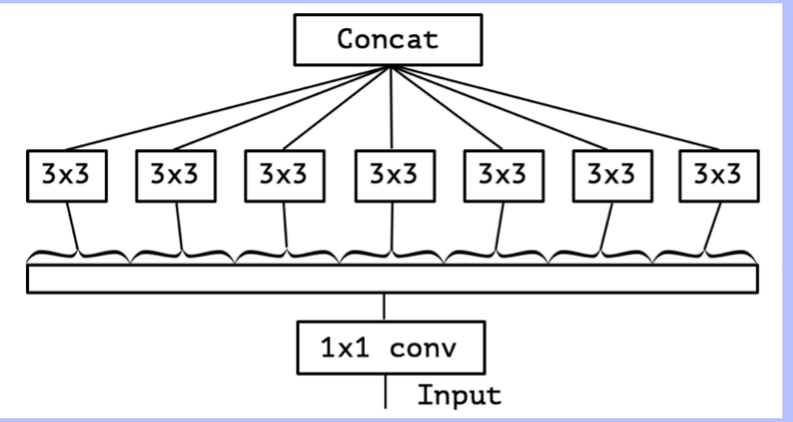

## ResMLP




## Swin: local attention



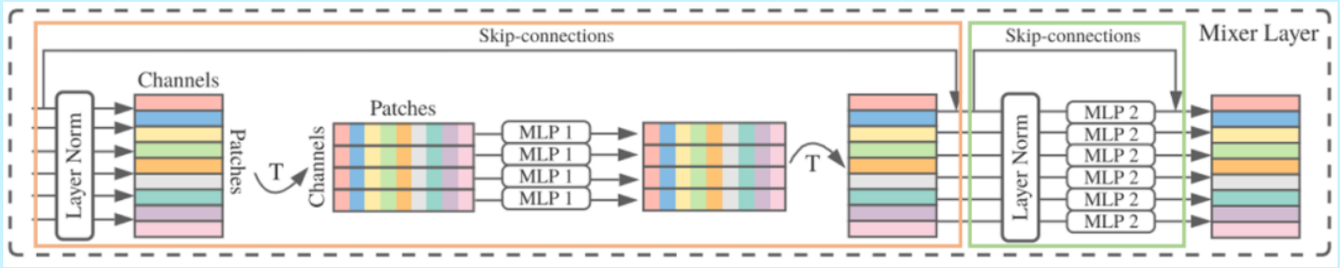

## Depth-wise separable conv.




Qi Han, Zejia Fan, Qi Dai, Lei Sun, Ming-Ming Cheng, Jiaying Liu, Jingdong Wang: Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight. ICLR 2022

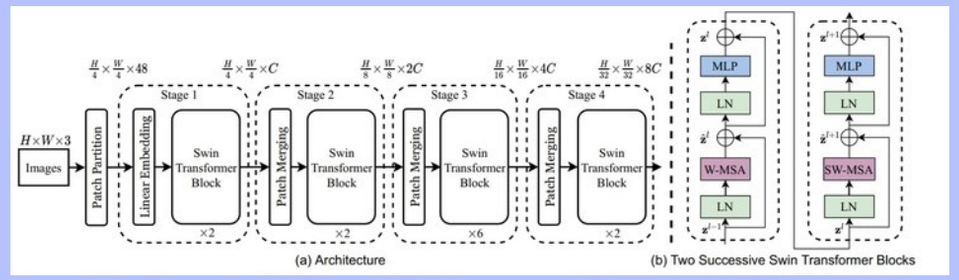
## Vision Transformer: Attention


## MLP-Mixer

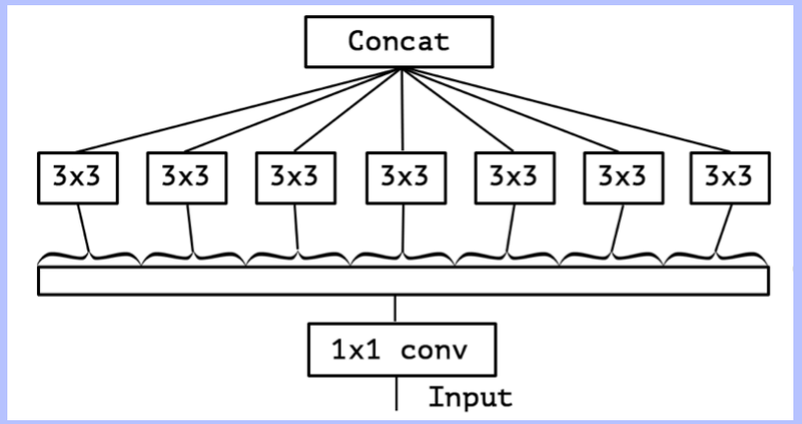



Local transformer attention is equivalent to depth-wise convolution

## Swin: local attention




## Depth-wise separable conv.




Qi Han, Zejia Fan, Qi Dai, Lei Sun, Ming-Ming Cheng, Jiaying Liu, Jingdong Wang: Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight. ICLR 2022

# DWNet (2021) and ConvNeXt (2022) studied the equivalence between local attention and depth-wise convolution

## DWNet in June 2021 (ours)

**ON THE CONNECTION BETWEEN LOCAL ATTENTION AND DYNAMIC DEPTH-WISE CONVOLUTION**

**Qi Han**[1*] **Zejia Fan**[2*] **Qi Dai**[3†] **Lei Sun**[3] **Ming-Ming Cheng**[1] **Jiaying Liu**[2]
**Jingdong Wang**[4†]
TKLNDST, CS, Nankai Univerisy[1], Peking University[2], Microsoft Research Asia[3], Baidu Inc.[4]

Same architecture and training setting with Swin Transformer

## ConvNeXt in January 2022 (Meta and UC Berkeley)

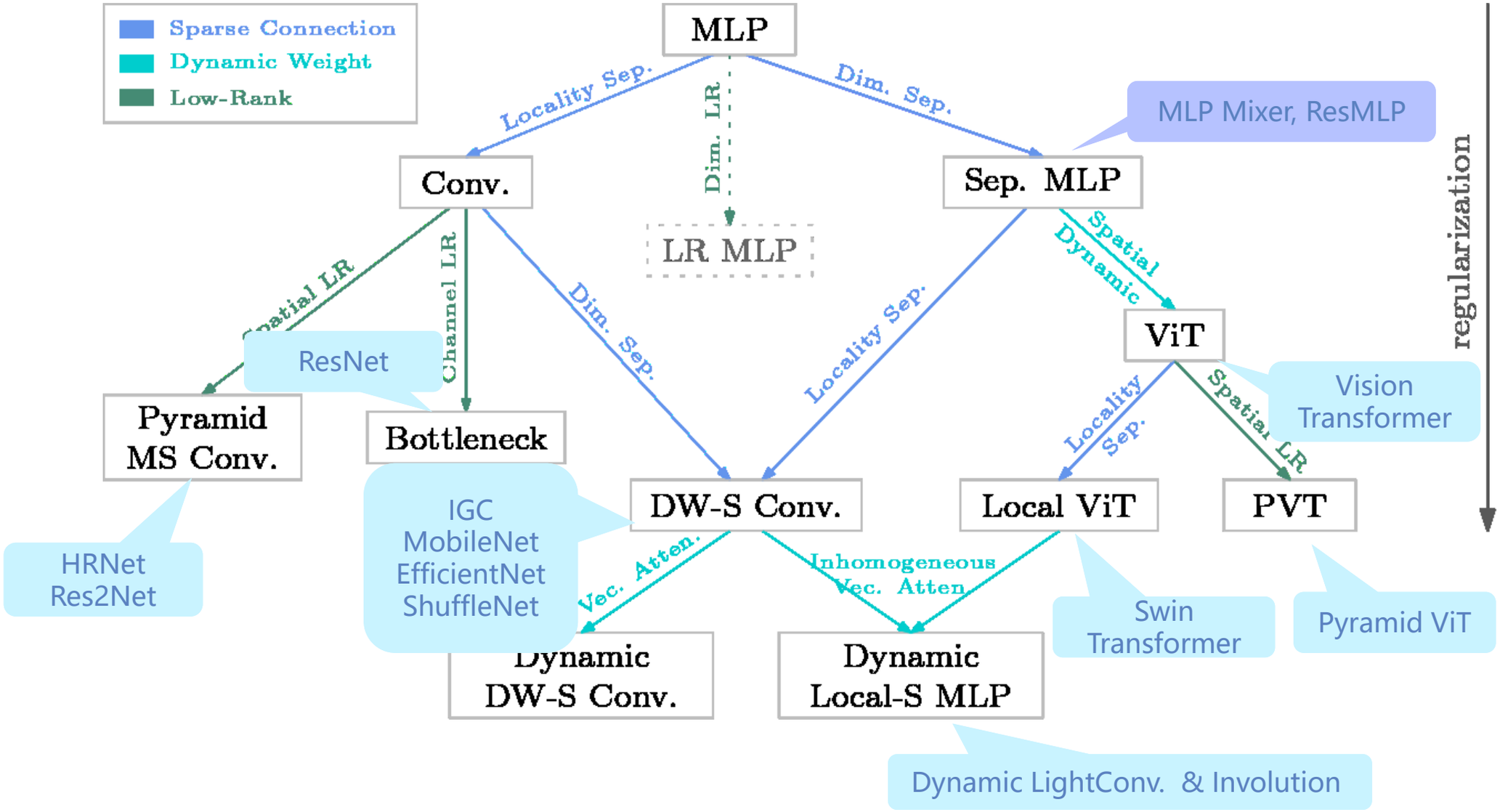
**A ConvNet for the 2020s**

Zhuang Liu[1,2*] Hanzi Mao[1] Chao-Yuan Wu[1] Christoph Feichtenhofer[1] Trevor Darrell[2] Saining Xie[1†]
[1]Facebook AI Research (FAIR) [2]UC Berkeley

Code: https://github.com/facebookresearch/ConvNeXt

Carefully tuned micro design and training setting

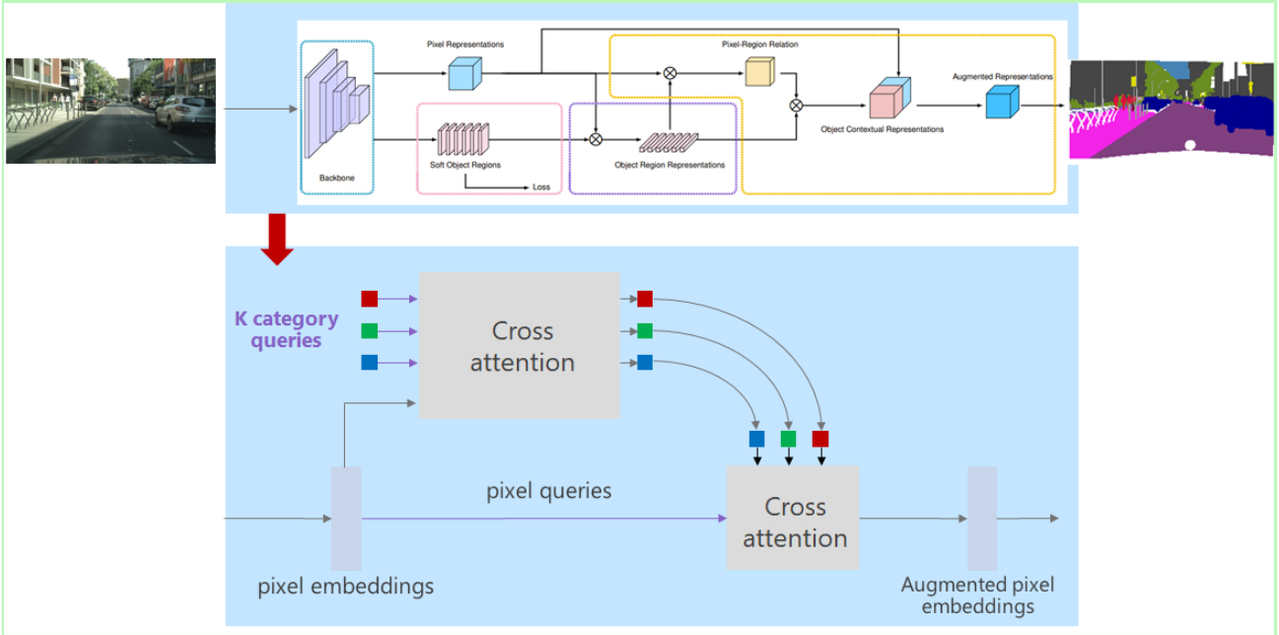
# Relation Graph for Typical Networks



# Query-based Dense Recognition with Transformers

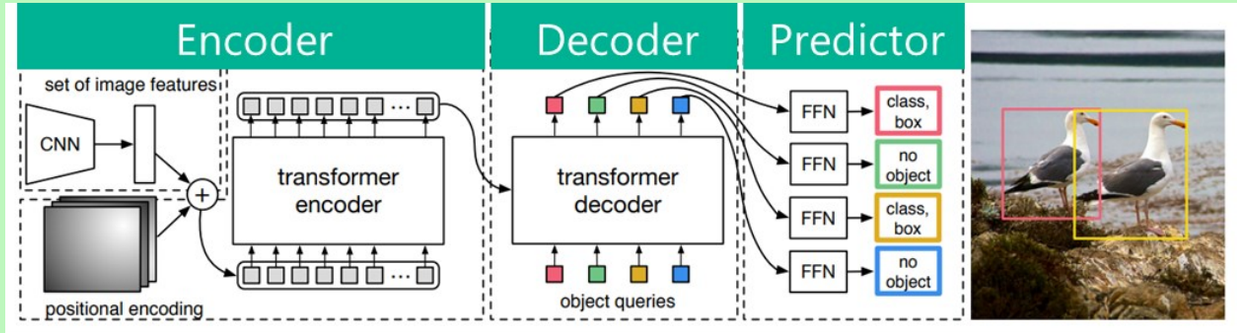
# Semantic Segmentation and Object Detection

## OCRNet for semantic segmentation




Yuhui Yuan, Xilin Chen, Jingdong Wang. Object-Contextual Representations for Semantic Segmentation.
ECCV 2020

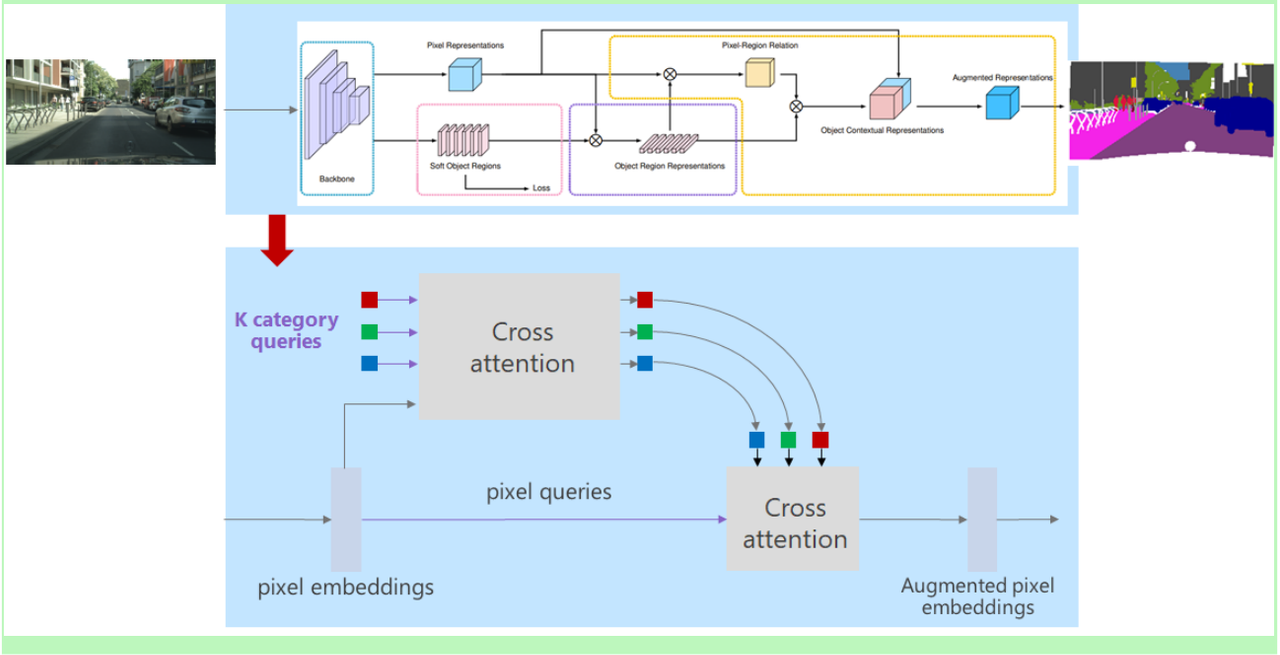
## DETR for object detection




Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko: End-to-End Object Detection with Transformers.
ECCV 2020

# OCRNet: Introduction of Category Queries for Semantic Segmentation

## OCRNet for semantic segmentation




Yuhui Yuan, Xilin Chen, Jingdong Wang. Object-Contextual Representations for Semantic Segmentation.
ECCV 2020

Motivation: the label of a pixel is the category of the object that the pixel belongs to

OCRNet: the first approach to introduce category queries for semantic segmentation

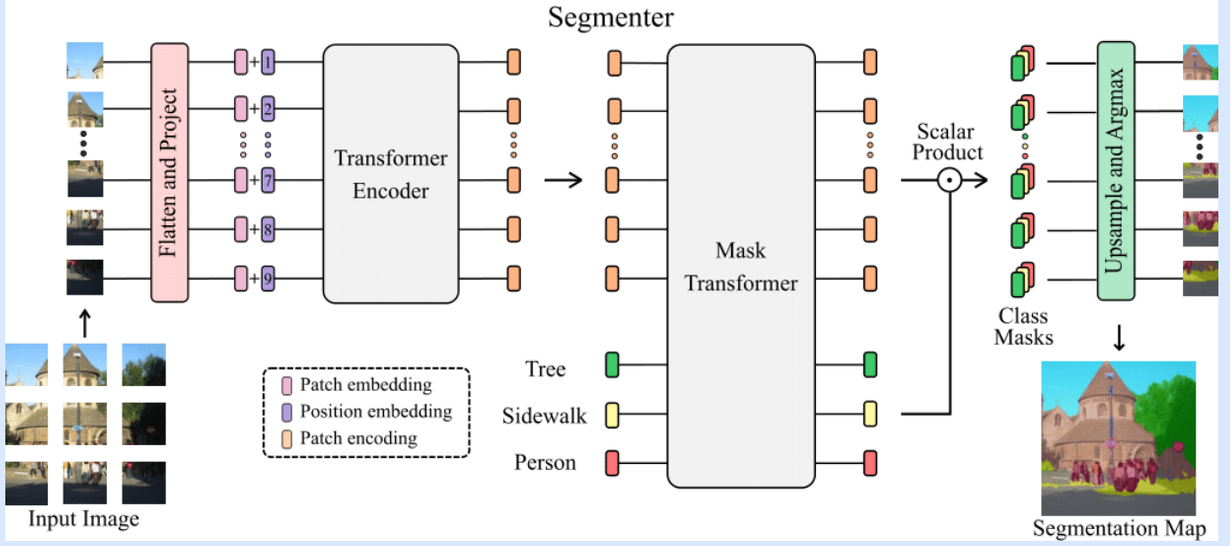
Cross-attention:

- ❑ Learn image-dependent category embeddings (OCR)
- ❑ Category embeddings as queries, and pixel embeddings as keys and values

Cross-attention:

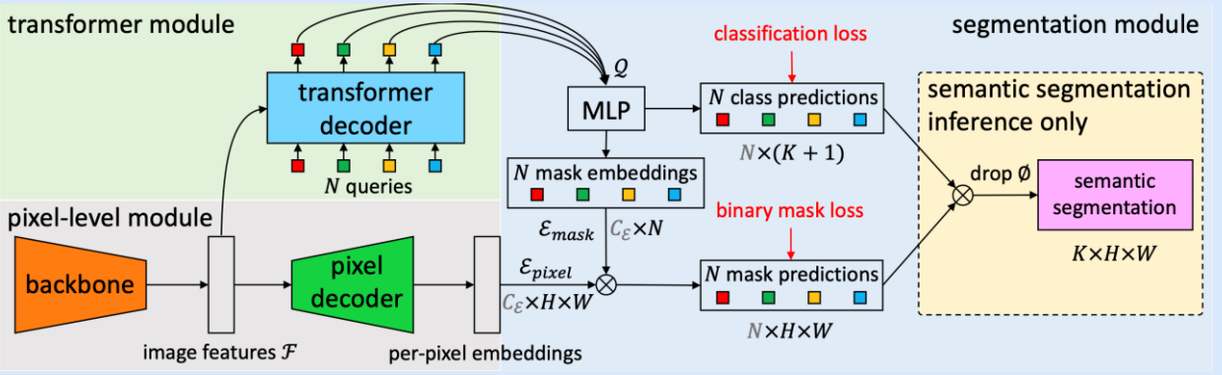
- ❑ Augment the pixel embeddings using OCRs
- ❑ Pixel embeddings as queries, and OCRs as keys and values

# More Transformer Methods for Semantic Segmentation in 2021




## Segmenter

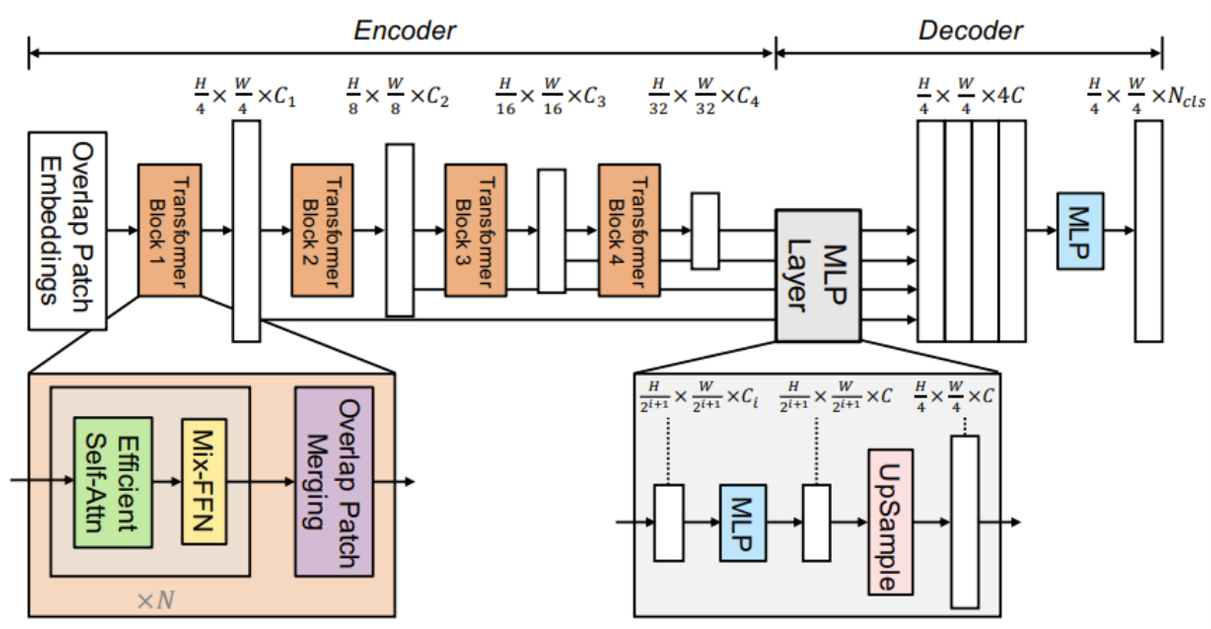
Robin Strudel, Ricardo Garcia Pinel, Ivan Laptev, Cordelia Schmid: Segmenter: Transformer for Semantic Segmentation. ICCV 2021: 7242-7252




## MaskFormer

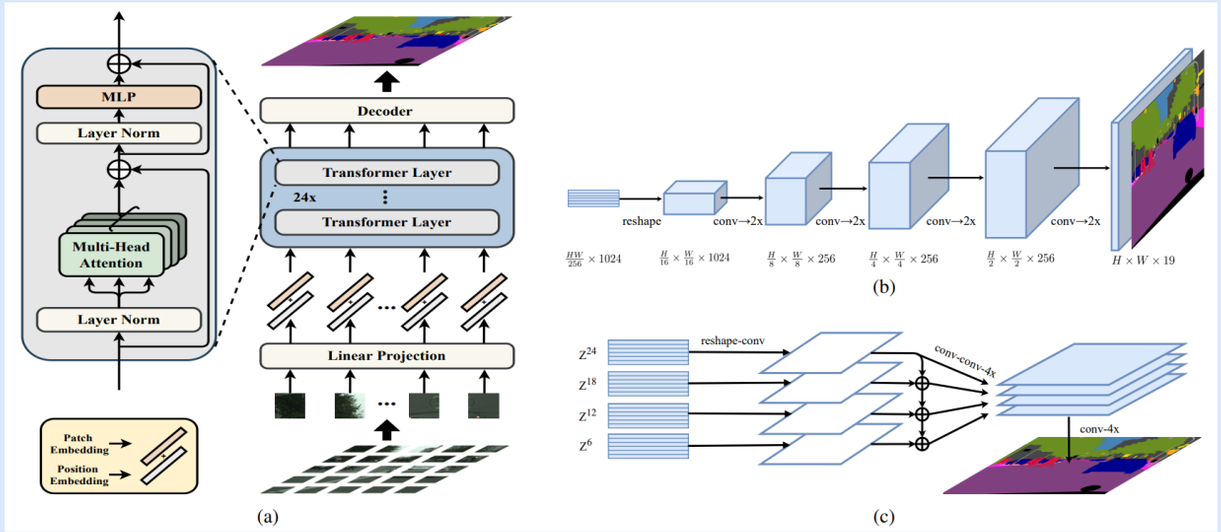
Bowen Cheng, Alexander G. Schwing, Alexander Kirillov: Per-Pixel Classification is Not All You Need for Semantic Segmentation. NeurIPS 2021: 17864-17875

# Transformer Encoder for Semantic Segmentation in 2021




## SegFormer

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo:
SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers.
NeurIPS 2021: 12077-12090

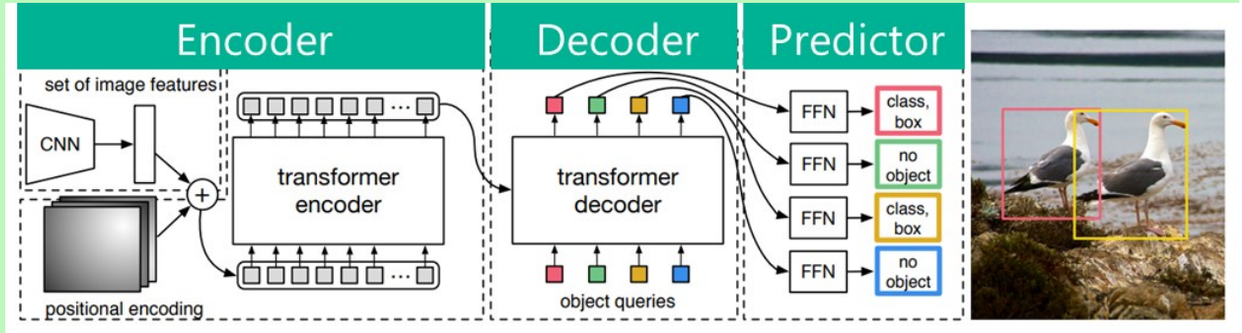



## SETR

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, Li Zhang:
Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Transformers. CVPR 2021: 6881-6890

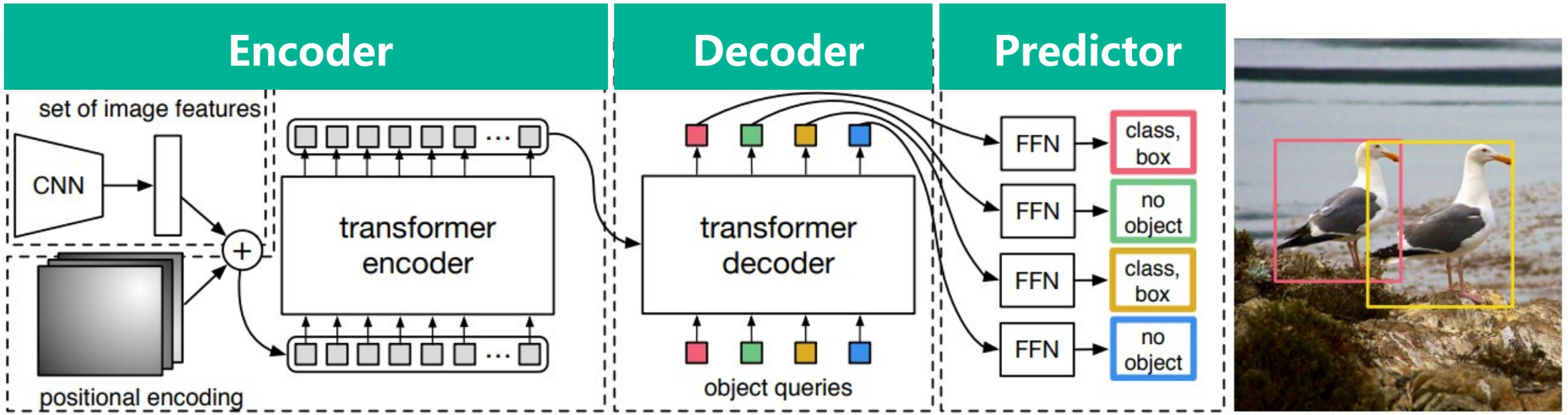
# Semantic Segmentation and Object Detection

## DETR for object detection



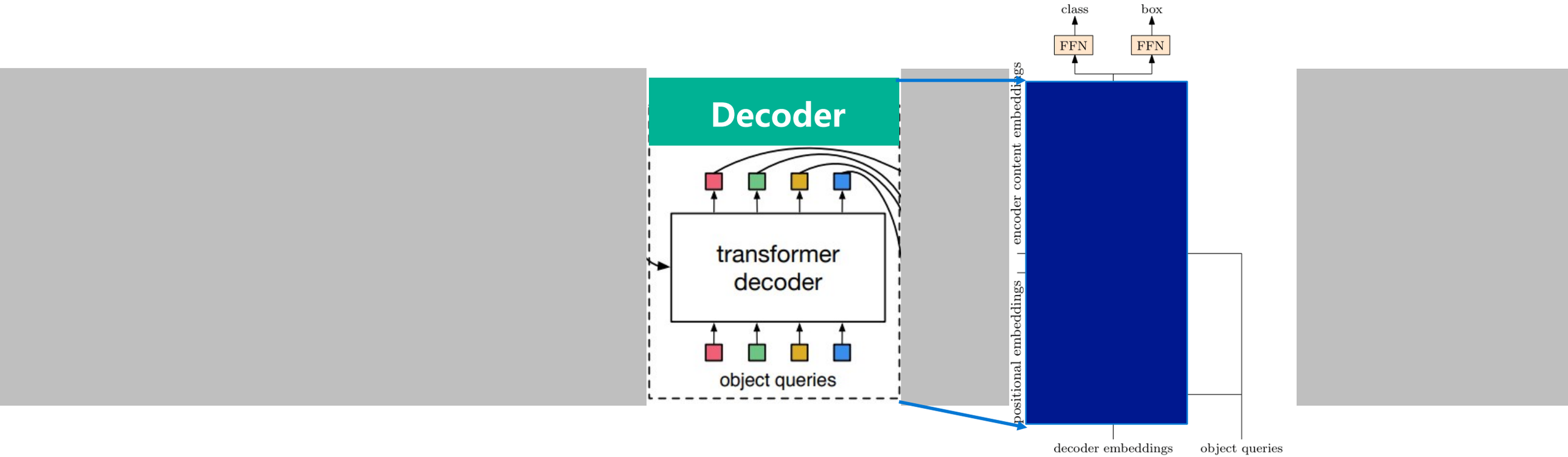

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko: End-to-End Object Detection with Transformers. ECCV 2020

# Detection Transformer (DETR)




Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko: End-to-End Object Detection with Transformers. ECCV (1) 2020: 213-229

# Detection Transformer (DETR)

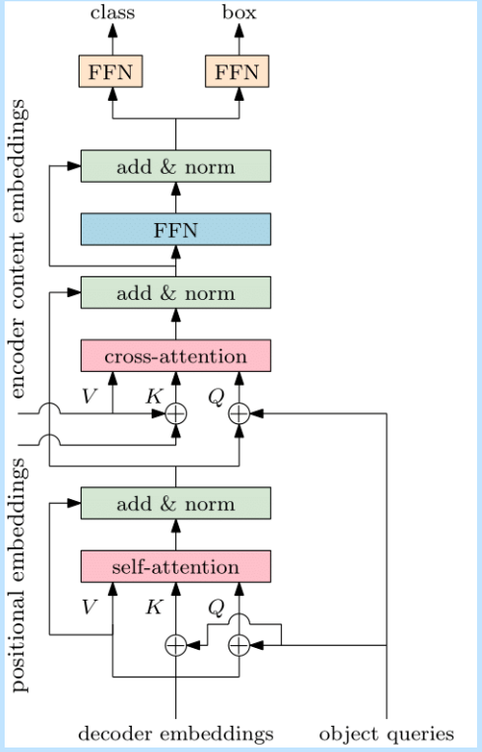



Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko: End-to-End Object Detection with Transformers. ECCV (1) 2020: 213-229

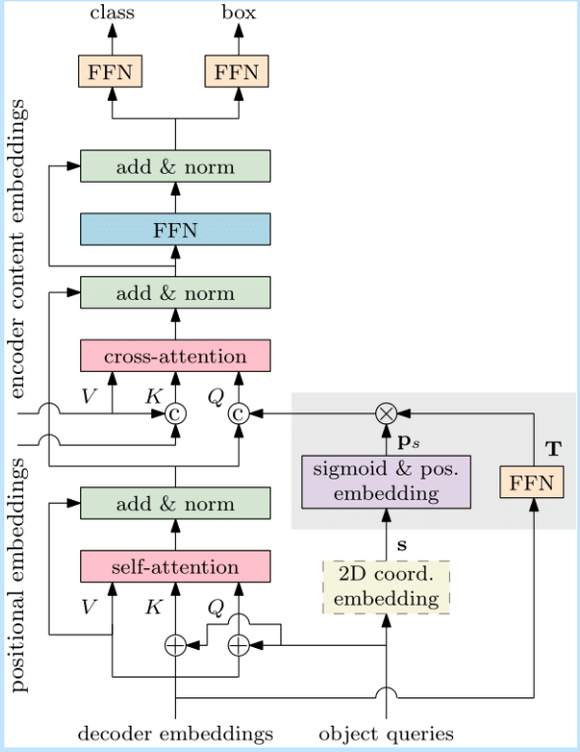
# DETR：Two Key Designs

| Object queries | One-to-one assignment |
| --- | --- |
| Detection as a search problem<br>- Localize the four extremities (box detection)<br>- Select a region inside the object (classification)<br><br>Object queries are learned as model parameters<br>- Same for all the images | One ground-truth object is assigned to one queries<br>- Necessary for removing NMS<br>- All the other queries are viewed as negative |
| Conditional DETR: Learning conditional spatial queries | Group DETR: Group-wise one-to-many assignment |

# Accelerating DETR: Conditional DETR




**DETR**



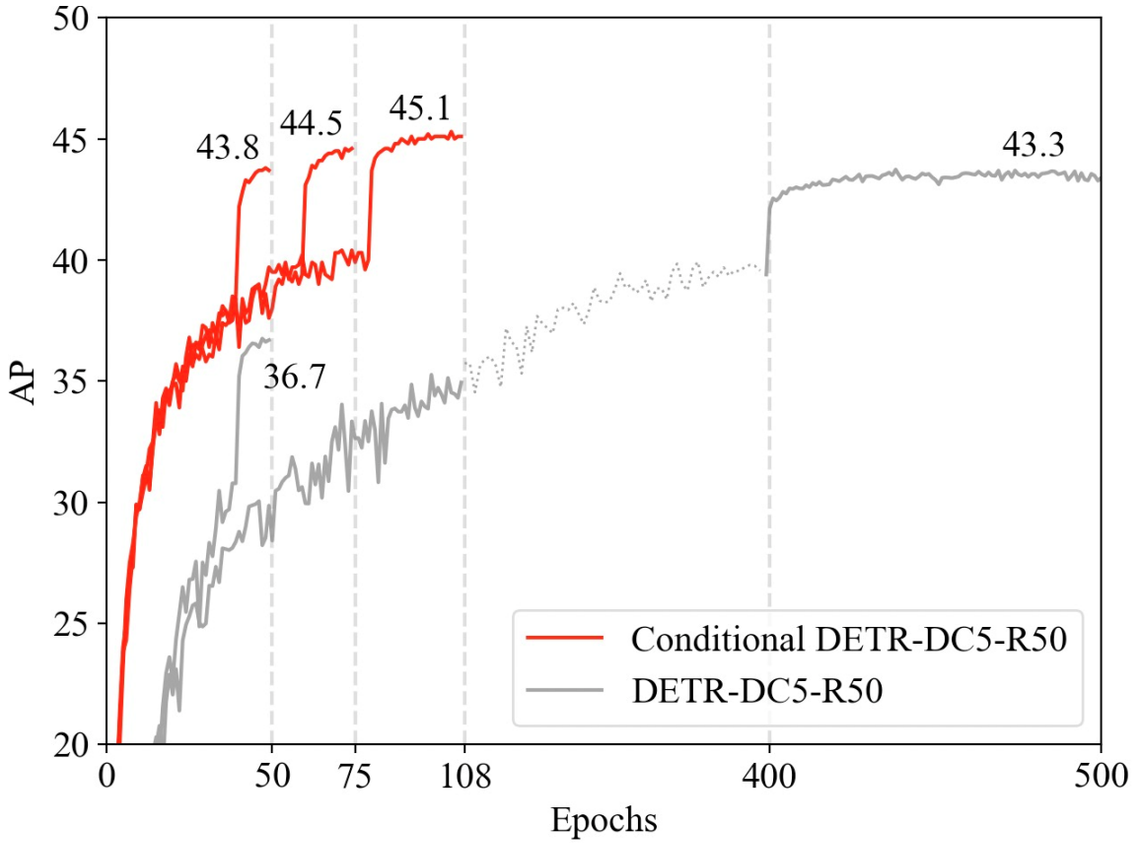

**Conditional DETR**: Decompose the queries into ***content and spatial queries***. **~10X faster**

Cross-attention maps: spatial, content, combined

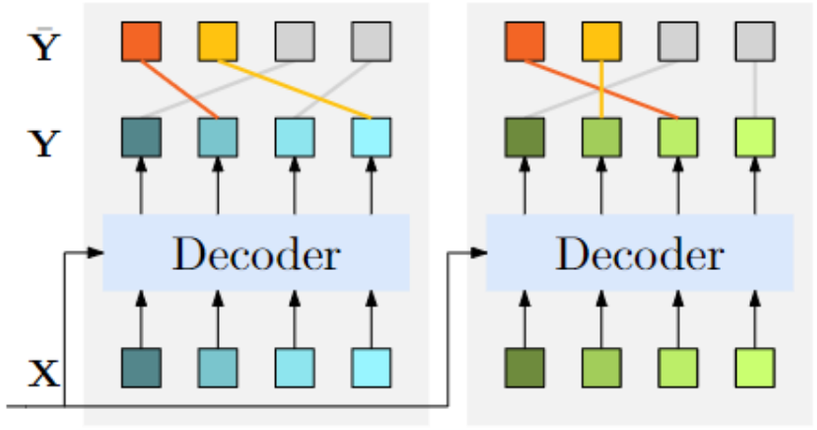
https://github.com/atten4vis

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, Jingdong Wang. Conditional DETR for Fast Training Convergence. ICCV 2021: 3631-3640

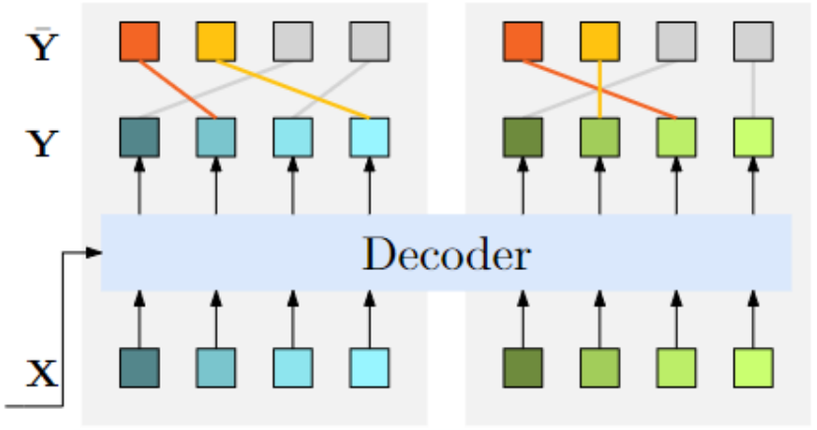
# Conditional DETR: Convergence is 10x Faster for DC5 – R50



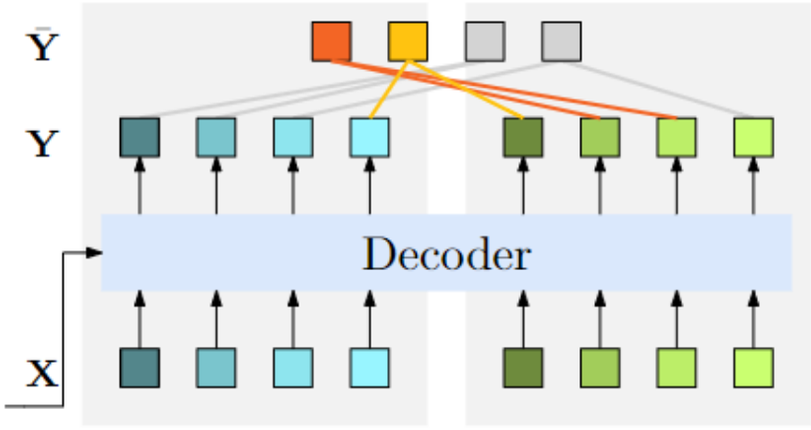
# Accelerating DETR: Group DETR




Group DETR



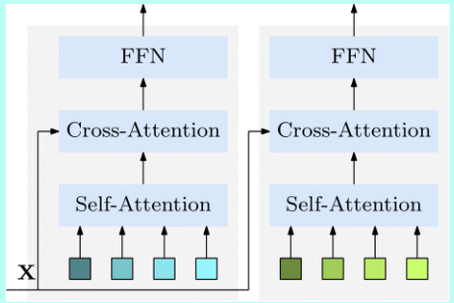

Group-wise assignment

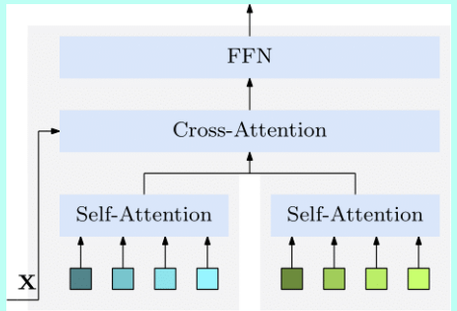



Naïve o2m assignment

Main design for Group DETR

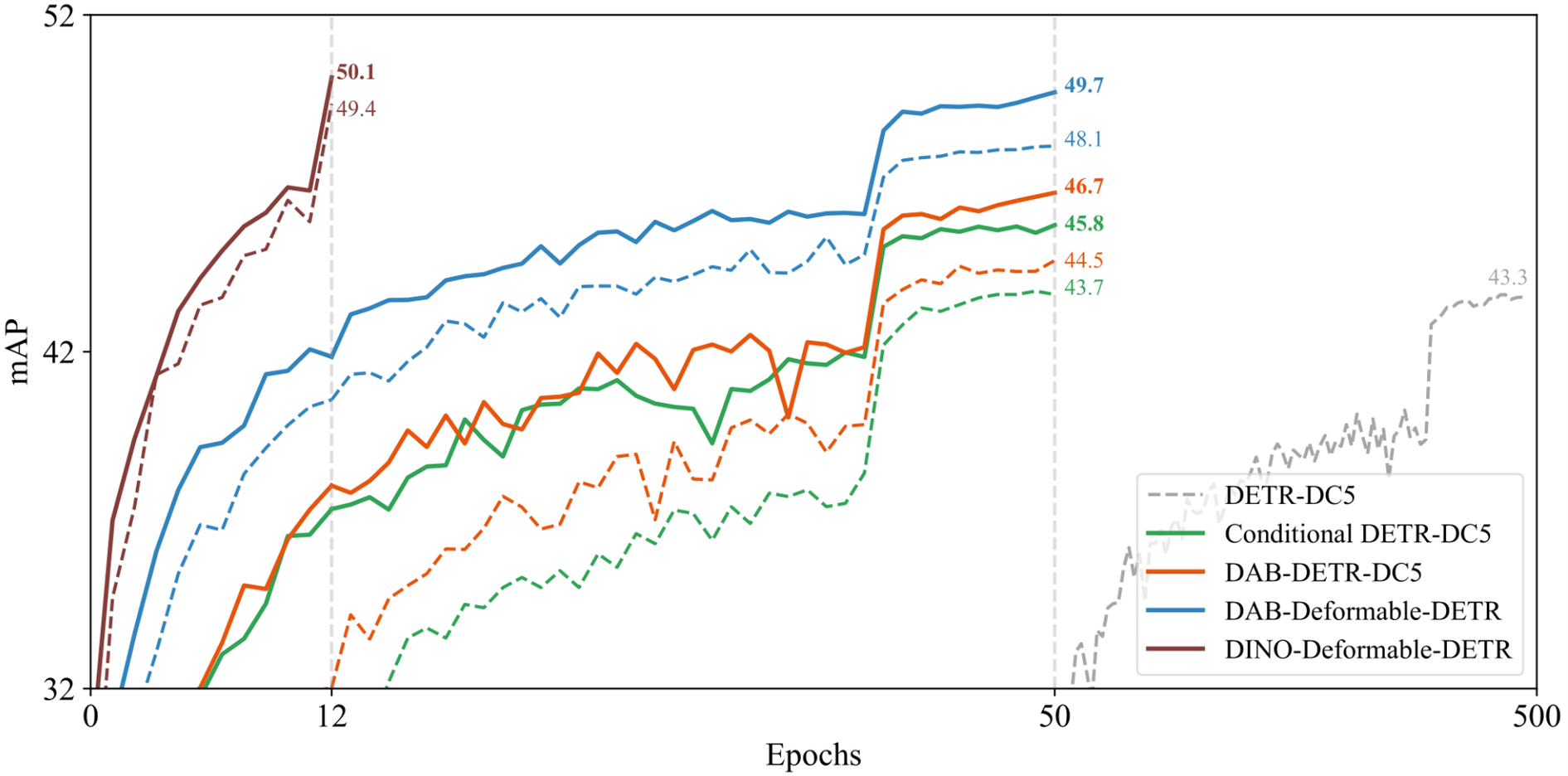
- ❑ More groups of queries
- ❑ Group-wise o2m assignment
- ❑ Separate self-attention




Qiang Chen, Xiaokang Chen, Gang Zeng, Jingdong Wang. Group DETR: Fast Training Convergence with Decoupled One-to-Many Label Assignment. CoRR abs/2207.13085 (2022)

# Group DETR: Faster Convergence



# Group DETR for Multi-view 3D Object Detection

| Method | w/ Group DETR | #Epochs | NDS | mAP |
|---|---|---|---|---|
| PETR | | 24 | 42.0 | 37.4 |
| PETR | √ | 24 | 45.0 **(+3.0)** | 38.8 **(+1.4)** |
| PETR v2 | | 24 | 50.3 | 40.7 |
| PETR v2 | √ | 24 | 51.3 **(+1.0)** | 41.9 **(+1.2)** |
| PETR v2 | | 36 | 50.8 | 41.3 |
| PETR v2 | √ | 36 | 52.3 **(+1.5)** | 42.7 **(+1.4)** |

# Group DETR v2: Encoder-Decoder Pretraining

Encoder

- ViT-Huge (628M)
- Multi-scale feature maps (4-scales)

Decoder

- DINO decoder
- Group DETR

Encoder pretraining and finetuning

- Pretraining with self-supervised method, e.g., CAE, on ImageNet-1K
- Finetuning on ImageNet-1K

Decoder pretraining and finetuning

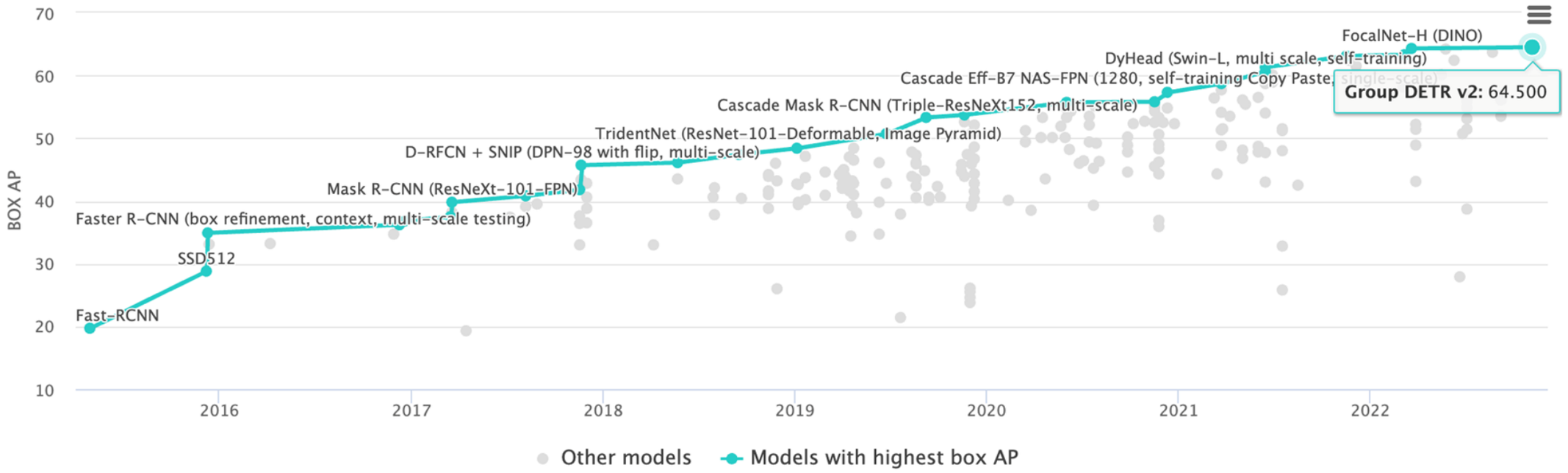
- Pretraining on Object365
- Finetuning on COCO

Qiang Chen, Jian Wang, Chuchu Han, Shan Zhang, Zexian Li, Xiaokang Chen, Jiahui Chen, Xiaodi Wang, Shuming Han, Gang Zhang, Haocheng Feng, Kun Yao, Junyu Han, Errui Ding, Jingdong Wang. Group DETR v2: Strong Object Detector with Encoder-Decoder Pretraining. CoRR abs/2211.03594 (2022)

# Group DETR v2: Encoder-Decoder Pretraining on Object 365

## 标准数据集 COCO 目标检测，首次达到64.5

11.8.2022



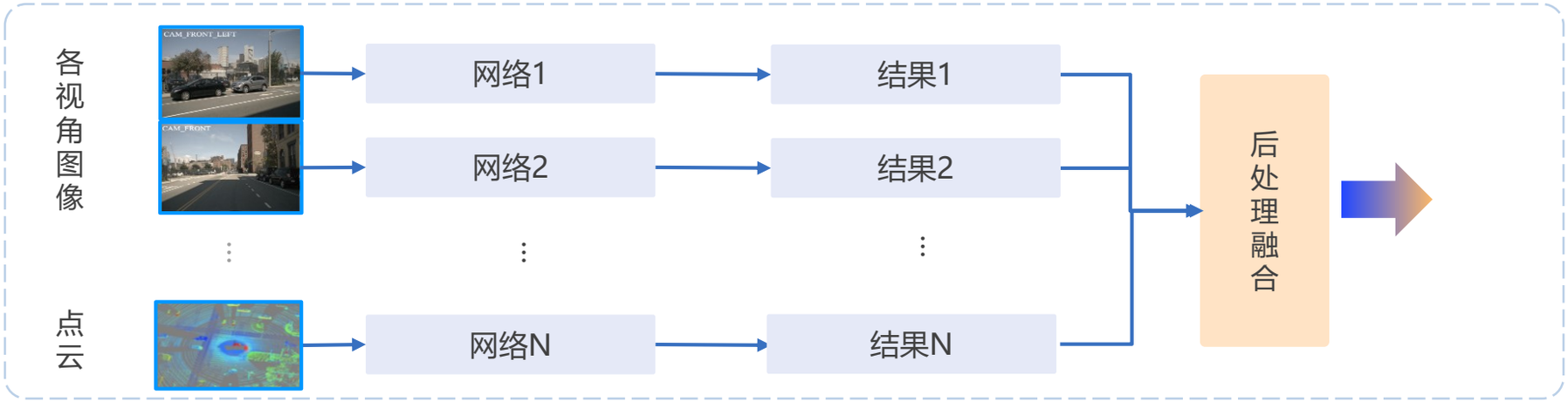

Qiang Chen, Jian Wang, Chuchu Han, Shan Zhang, Zexian Li, Xiaokang Chen, Jiahui Chen, Xiaodi Wang, Shuming Han, Gang Zhang, Haocheng Feng, Kun Yao, Junyu Han, Errui Ding, Jingdong Wang. Group DETR v2: Strong Object Detector with Encoder-Decoder Pretraining. CoRR abs/2211.03594 (2022)

# 3D Recognition with Transformers for Autonomous Driving

# 传统感知 VS UniBEV 感知

**传统感知**

**分而治之**

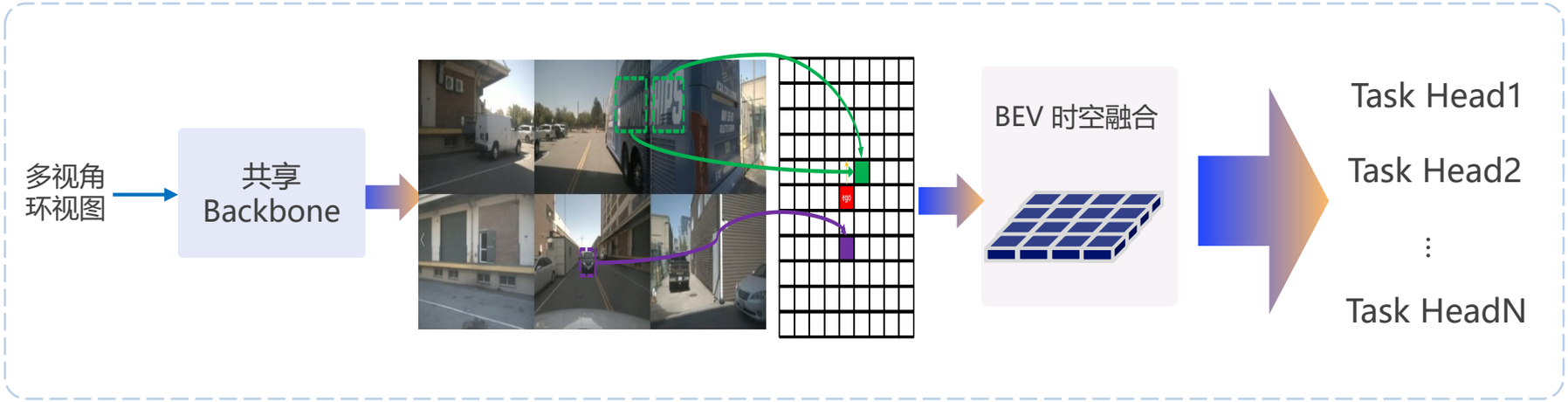



算力有限

后处理复杂

**基于 UniBEV 的感知**

✓

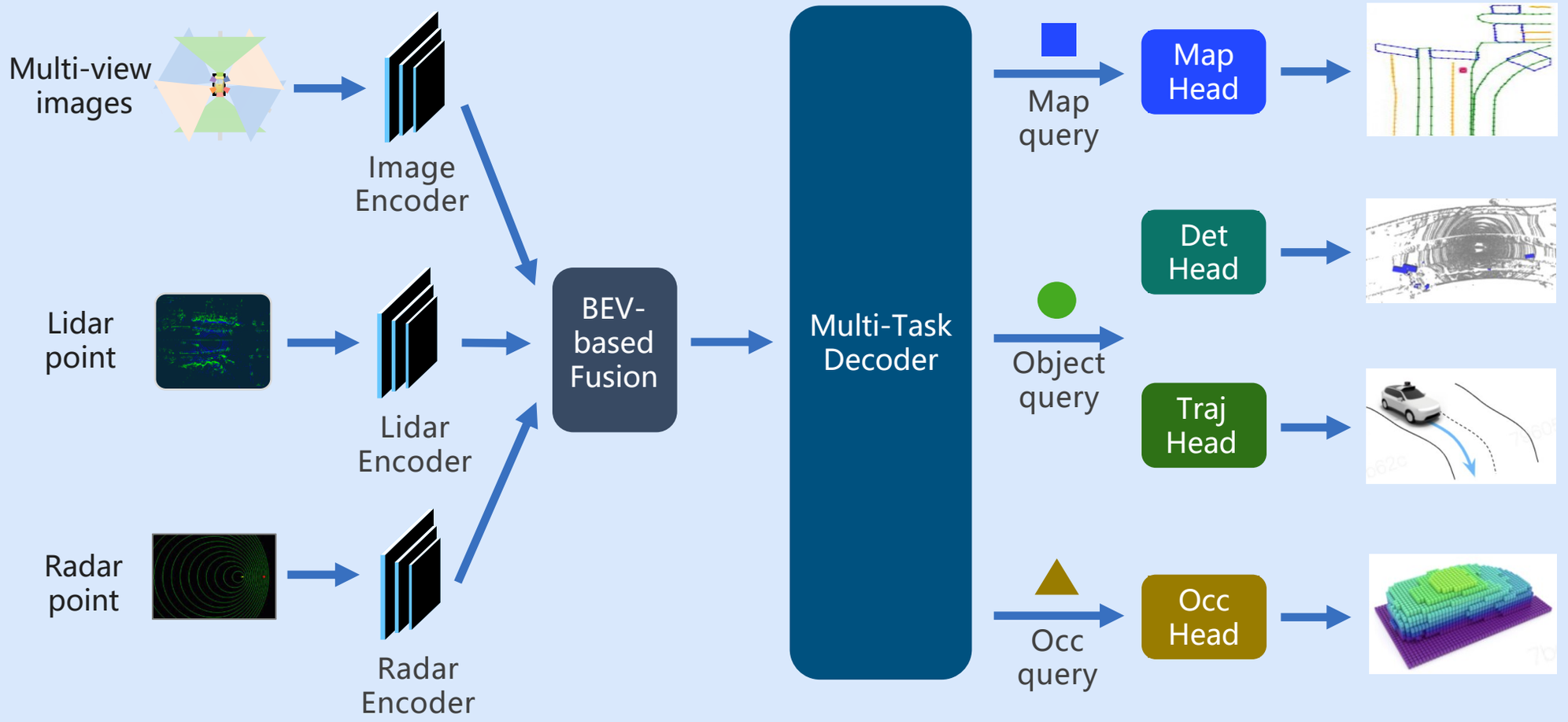



共享+交互

无复杂后处理

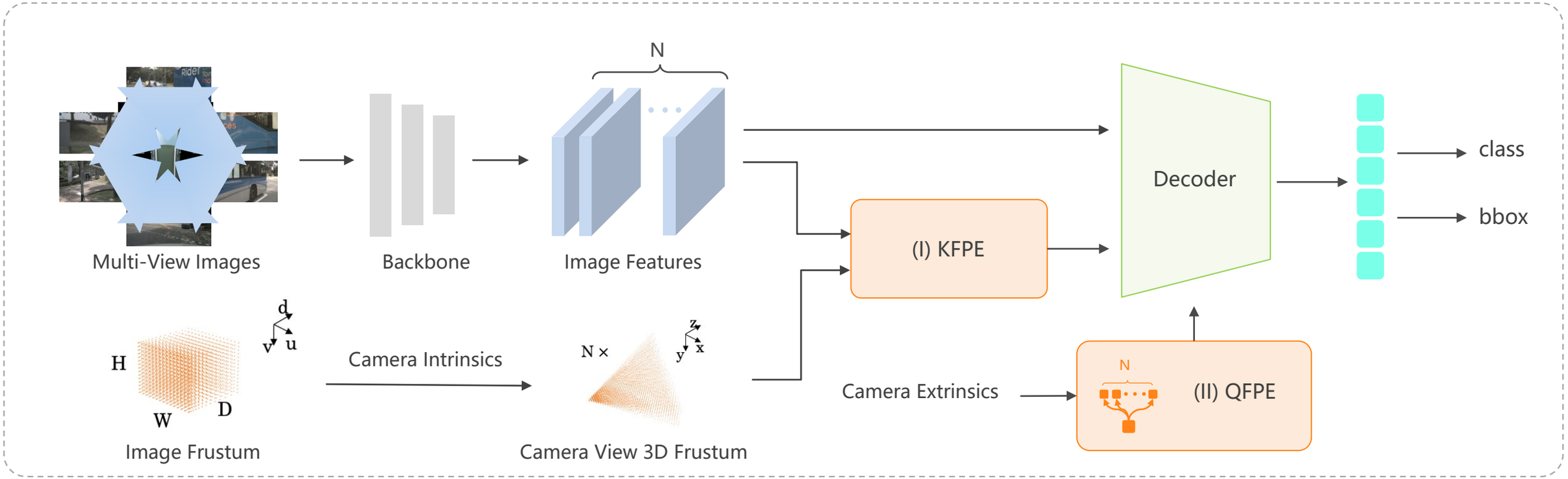
基于 BEV 鸟瞰图视角的统一特征空间表达

# 基于 UniBEV 的 **多任务+多模态** 统一融合方案



# UniBEV-CAPE:
# Camera View Position Embedding for Multi-View 3D Object Detection

- DETR-style paradigm: sparse object queries  in 3D space
- Main point: Camera View Position Embedding

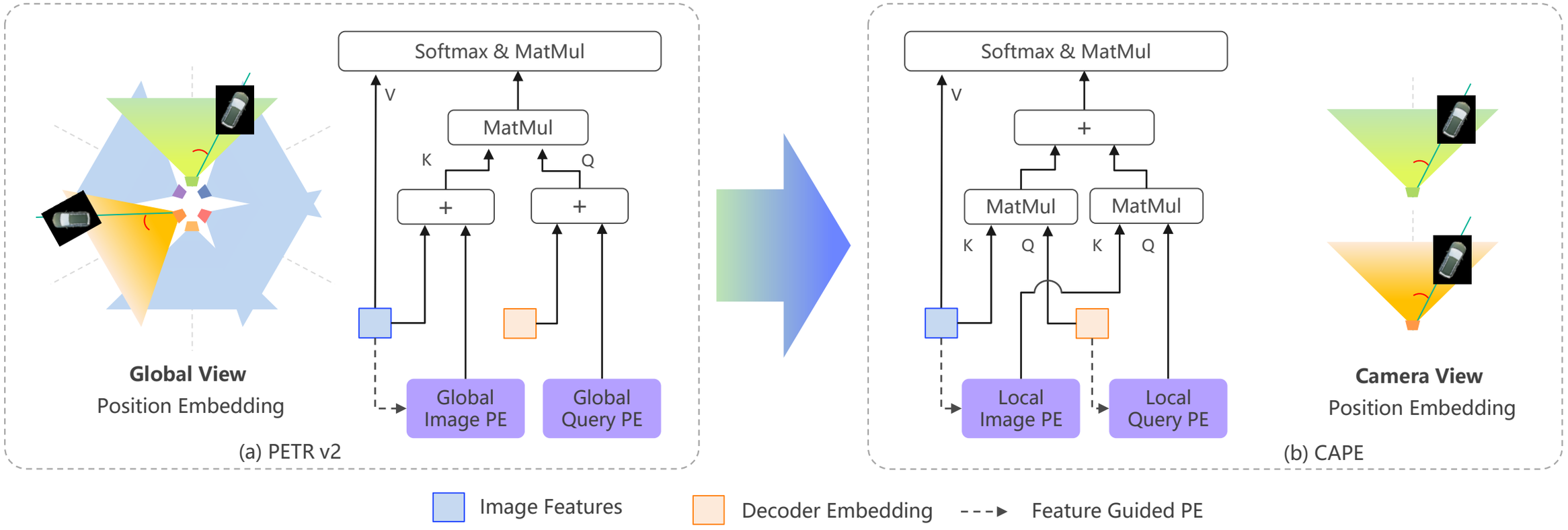



Kaixin Xiong, Shi Gong, Xiaoqing Ye, Xiao Tan, Ji Wan, Errui Ding, Jingdong Wang, Xiang Bai. CAPE: Camera View Position Embedding for Multi-View 3D Object Detection. CVPR 2023

CAPE code：https://github.com/PaddlePaddle/Paddle3D

# UniBEV-CAPE:
# Camera View Position Embedding for Multi-View 3D Object Detection

- Global view position embedding vs camera view position embedding
- Plain attention vs bilateral attention (borrowed from Conditional DETR)




Kaixin Xiong, Shi Gong, Xiaoqing Ye, Xiao Tan, Ji Wan, Errui Ding, Jingdong Wang, Xiang Bai. CAPE: Camera View Position Embedding for Multi-View 3D Object Detection. CVPR 2023

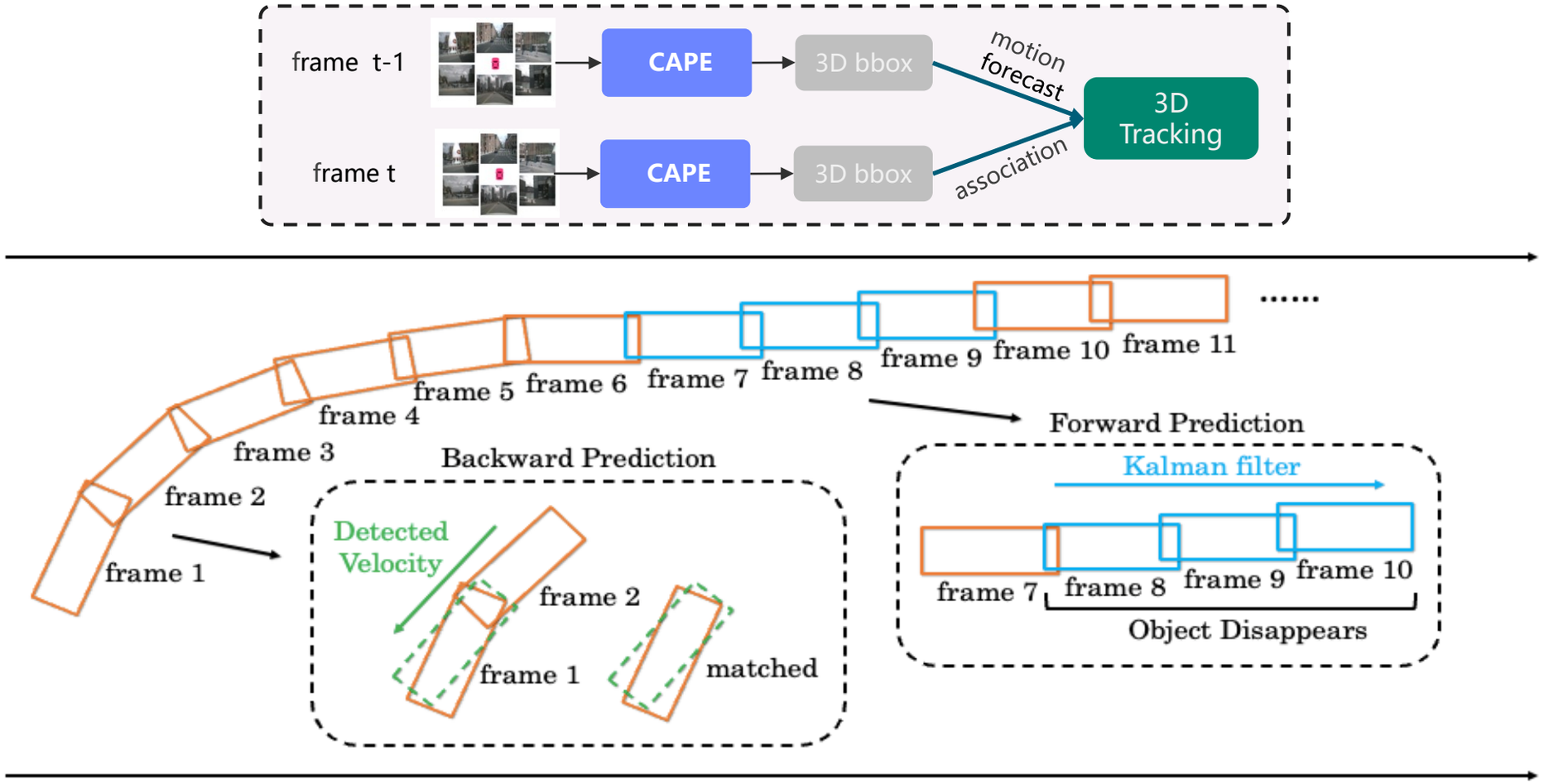
CAPE code：https://github.com/PaddlePaddle/Paddle3D

# UniBEV-CAPE: Results on nuScenes Test

| Method | Temporal | Year | Backbone | NDS | mAP |
|---|---|---|---|---|---|
| DETR3D | ✗ | CORL2022 | Res-101 | 47.9 | 41.2 |
| PETR | ✗ | ECCV2022 | V2-99 | 50.4 | 44.1 |
| **CAPE** | ✗ | CVPR2023 | V2-99 | 52.0 (+1.6) | 45.8 (+1.7) |
| BEVFormer | ✓ | ECCV2022 | V2-99 | 56.9 | 48.1 |
| BEVDet4D | ✓ | arXiv2022 | Swin-B | 56.9 | 45.1 |
| PETR v2 | ✓ | arXiv2022 | V2-99 | 58.2 | 49.0 |
| **CAPE-T** | ✓ | CVPR2023 | V2-99 | 61.0 (+2.8) | 52.5 (+3.5) |
| | | | ViT-Huge | 62.8 (+4.6) | 55.3 (+6.3) |

CAPE code：https://github.com/PaddlePaddle/Paddle3D

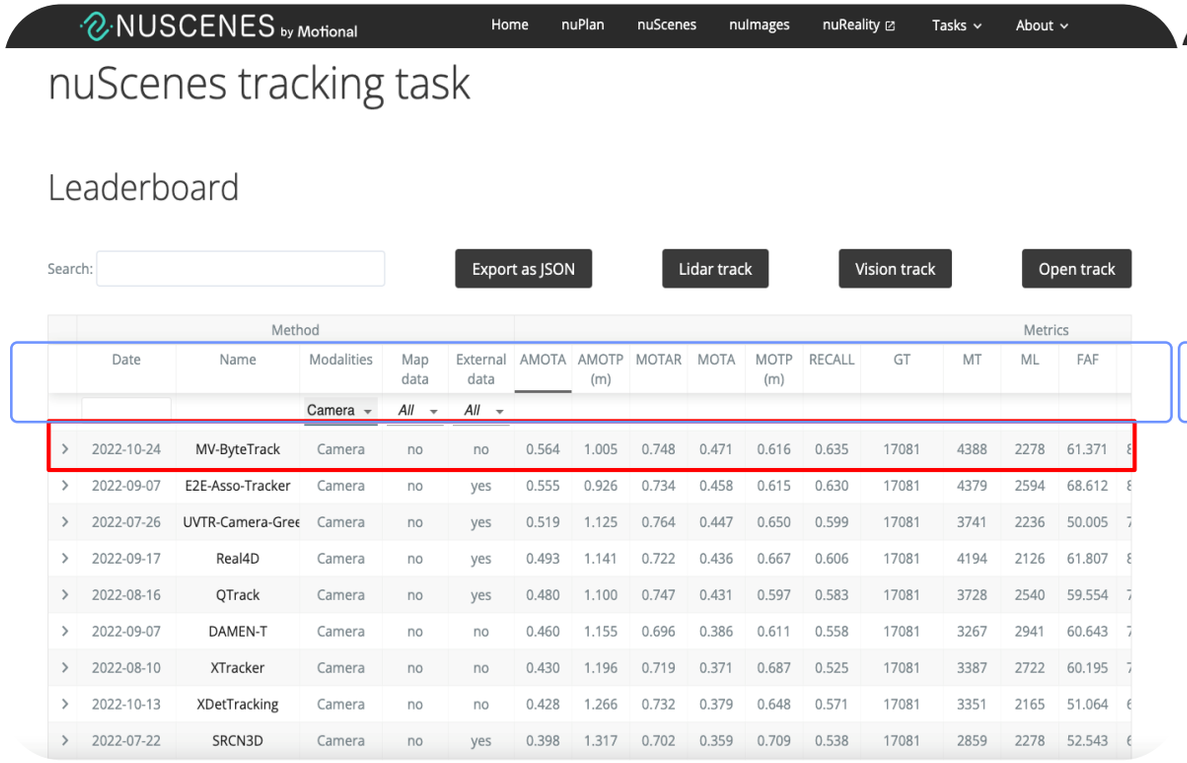
# UniBEV-CAPE for Tracking




Zhang, Yifu, et al. "ByteTrackV2: 2D and 3D Multi-Object Tracking by Associating Every Detection Box." arXiv:2303.15334.

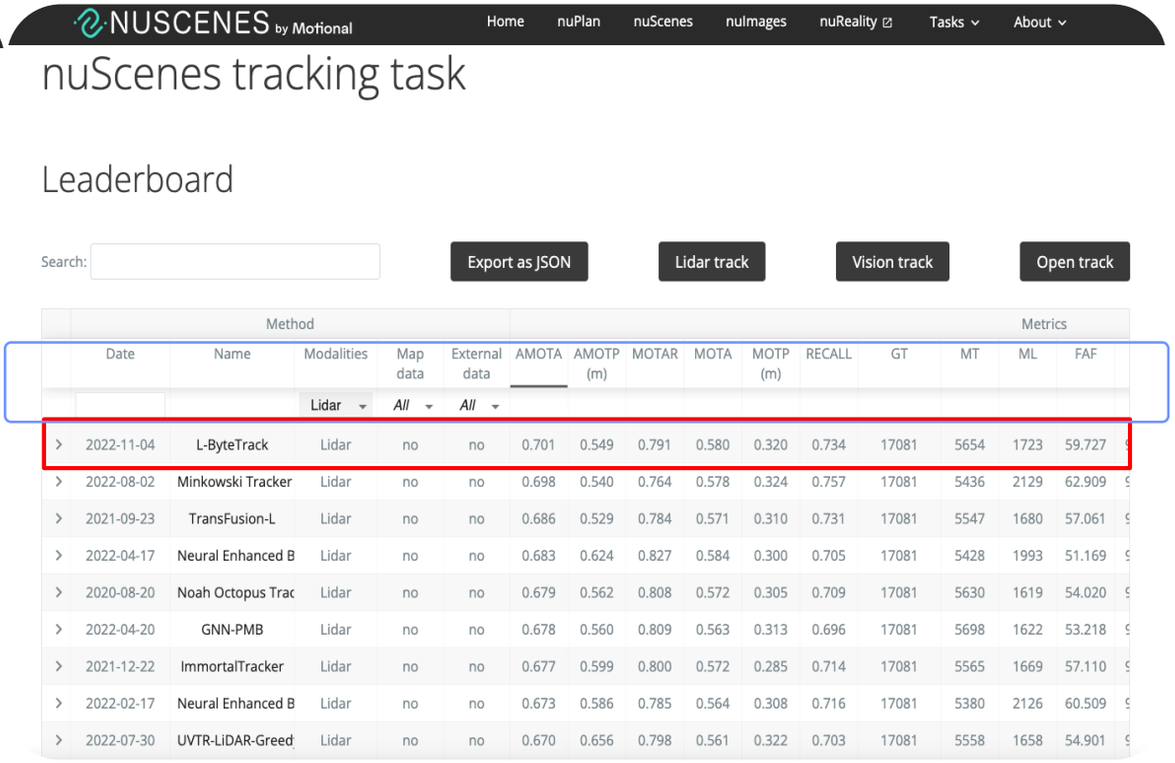
# UniBEV-CAPE for Tracking

## Camera modality




| Method | | | | | Metrics | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Date | Name | Modalities | Map data | External data | AMOTA | AMOTP (m) | MOTAR | MOTA | MOTP (m) | RECALL | GT | MT | ML | FAF |
| | | Camera | All | All | | | | | | | | | | |
| 2022-10-24 | MV-ByteTrack | Camera | no | no | 0.564 | 1.005 | 0.748 | 0.471 | 0.616 | 0.635 | 17081 | 4388 | 2278 | 61.371 |
| 2022-09-07 | E2E-Asso-Tracker | Camera | no | yes | 0.555 | 0.926 | 0.734 | 0.458 | 0.615 | 0.630 | 17081 | 4379 | 2594 | 68.612 |
| 2022-07-26 | UVTR-Camera-Gree | Camera | no | yes | 0.519 | 1.125 | 0.764 | 0.447 | 0.650 | 0.599 | 17081 | 3741 | 2236 | 50.005 |
| 2022-09-17 | Real4D | Camera | no | yes | 0.493 | 1.141 | 0.722 | 0.436 | 0.667 | 0.606 | 17081 | 4194 | 2126 | 61.807 |
| 2022-08-16 | QTrack | Camera | no | yes | 0.480 | 1.100 | 0.747 | 0.431 | 0.597 | 0.583 | 17081 | 3728 | 2540 | 59.554 |
| 2022-09-07 | DAMEN-T | Camera | no | no | 0.460 | 1.155 | 0.696 | 0.386 | 0.611 | 0.558 | 17081 | 3267 | 2941 | 60.643 |
| 2022-08-10 | XTracker | Camera | no | no | 0.430 | 1.196 | 0.719 | 0.371 | 0.687 | 0.525 | 17081 | 3387 | 2722 | 60.195 |
| 2022-10-13 | XDetTracking | Camera | no | no | 0.428 | 1.266 | 0.732 | 0.379 | 0.648 | 0.571 | 17081 | 3351 | 2165 | 51.064 |
| 2022-07-22 | SRCN3D | Camera | no | yes | 0.398 | 1.317 | 0.702 | 0.359 | 0.709 | 0.538 | 17081 | 2859 | 2278 | 52.543 |

## LiDAR modality




| Method | | | | | Metrics | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Date | Name | Modalities | Map data | External data | AMOTA | AMOTP (m) | MOTAR | MOTA | MOTP (m) | RECALL | GT | MT | ML | FAF |
| | | Lidar | All | All | | | | | | | | | | |
| 2022-11-04 | L-ByteTrack | Lidar | no | no | 0.701 | 0.549 | 0.791 | 0.580 | 0.320 | 0.734 | 17081 | 5654 | 1723 | 59.727 |
| 2022-08-02 | Minkowski Tracker | Lidar | no | no | 0.698 | 0.540 | 0.764 | 0.578 | 0.324 | 0.757 | 17081 | 5436 | 2129 | 62.909 |
| 2021-09-23 | TransFusion-L | Lidar | no | no | 0.686 | 0.529 | 0.784 | 0.571 | 0.310 | 0.731 | 17081 | 5547 | 1680 | 57.061 |
| 2022-04-17 | Neural Enhanced B | Lidar | no | no | 0.683 | 0.624 | 0.827 | 0.584 | 0.300 | 0.705 | 17081 | 5428 | 1993 | 51.169 |
| 2020-08-20 | Noah Octopus Trac | Lidar | no | no | 0.679 | 0.562 | 0.808 | 0.572 | 0.305 | 0.709 | 17081 | 5630 | 1619 | 54.020 |
| 2022-04-20 | GNN-PMB | Lidar | no | no | 0.678 | 0.560 | 0.809 | 0.563 | 0.313 | 0.696 | 17081 | 5698 | 1622 | 53.218 |
| 2021-12-22 | ImmortalTracker | Lidar | no | no | 0.677 | 0.599 | 0.800 | 0.572 | 0.285 | 0.714 | 17081 | 5565 | 1669 | 57.110 |
| 2022-02-17 | Neural Enhanced B | Lidar | no | no | 0.673 | 0.586 | 0.785 | 0.564 | 0.308 | 0.716 | 17081 | 5380 | 2126 | 60.509 |
| 2022-07-30 | UVTR-LiDAR-Greed | Lidar | no | no | 0.670 | 0.656 | 0.798 | 0.561 | 0.322 | 0.703 | 17081 | 5558 | 1658 | 54.901 |

Uploaded on 2022.11

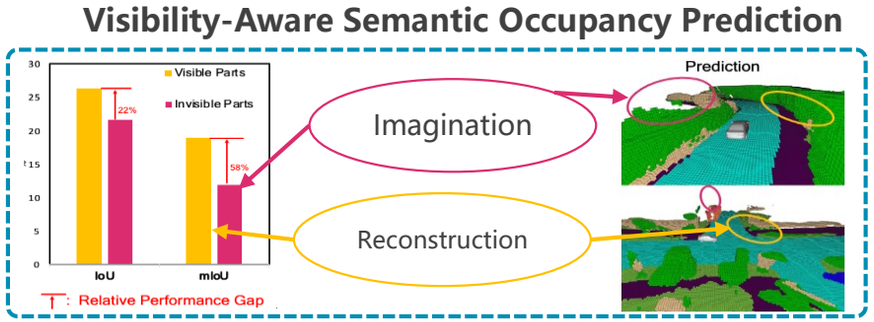
Zhang, Yifu, et al. "ByteTrackV2: 2D and 3D Multi-Object Tracking by Associating Every Detection Box." arXiv:2303.15334.

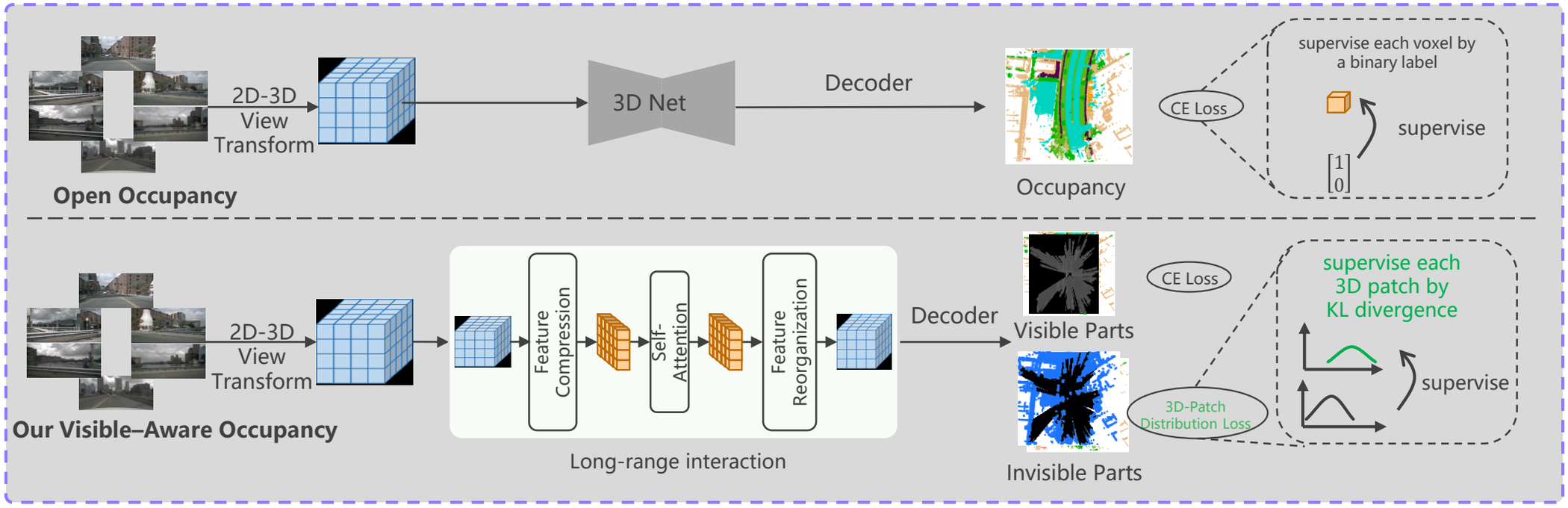
# UniBEV-3D-Occupancy

***Motivation***

Visible results **vs** Invisible results

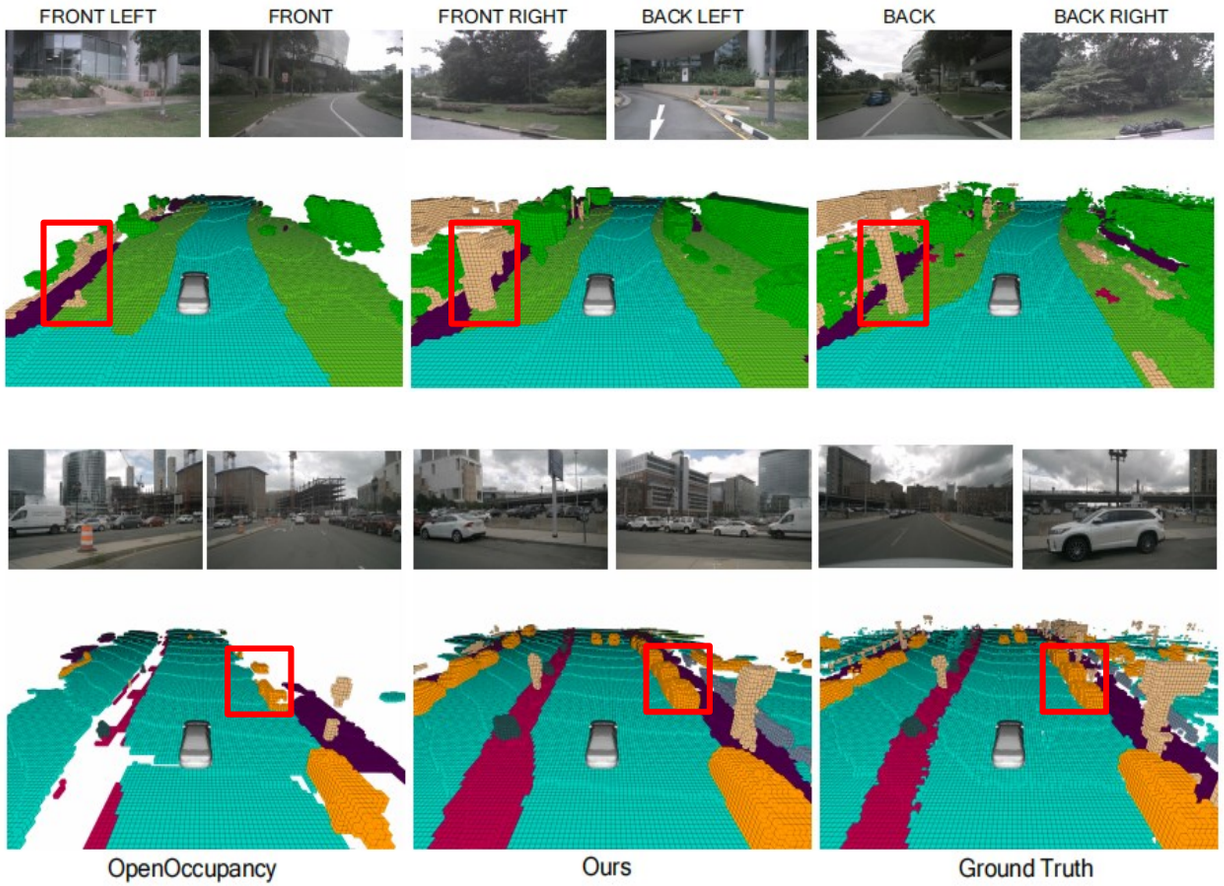
Plain occupancy -> Visibility-aware occupancy



# Results Comparison and Analysis

- More accurate predictions in the visible parts
- Denser predictions in the invisible parts




- Video results on nuScenes dataset

**Demo 1**

**Demo 2**

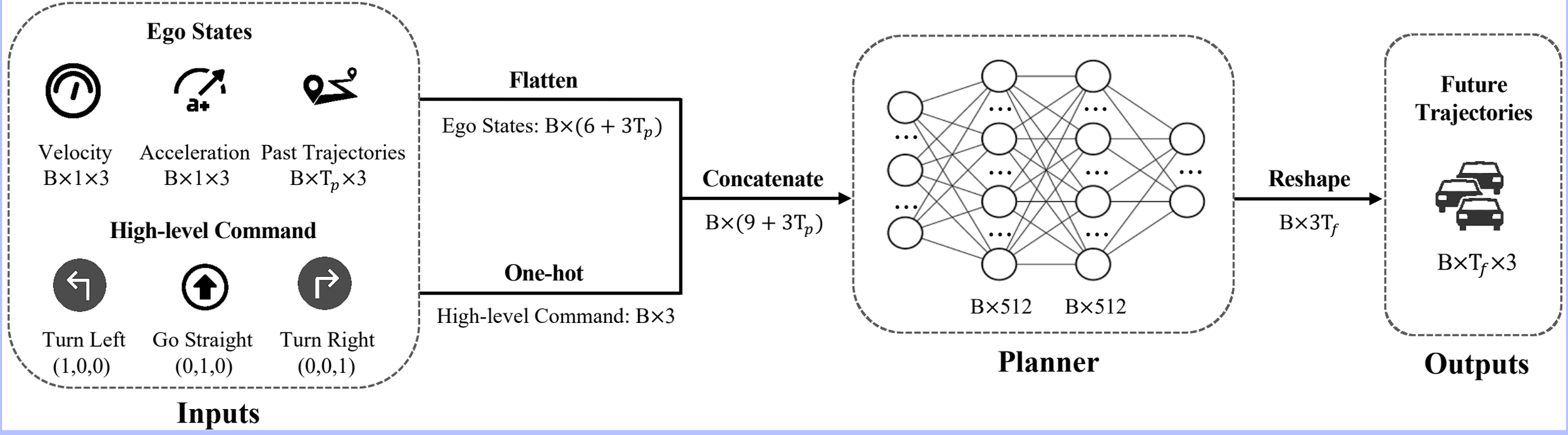
Wang, Xiaofeng, et al. "OpenOccupancy: A Large Scale Benchmark for Surrounding Semantic Occupancy Perception." arXiv:2303.03991.

Aoyan Li, et.al.. Visibility-Aware Semantic Occupancy Prediction.【Confidential】

# End-to-End Autonomous Driving Evaluation

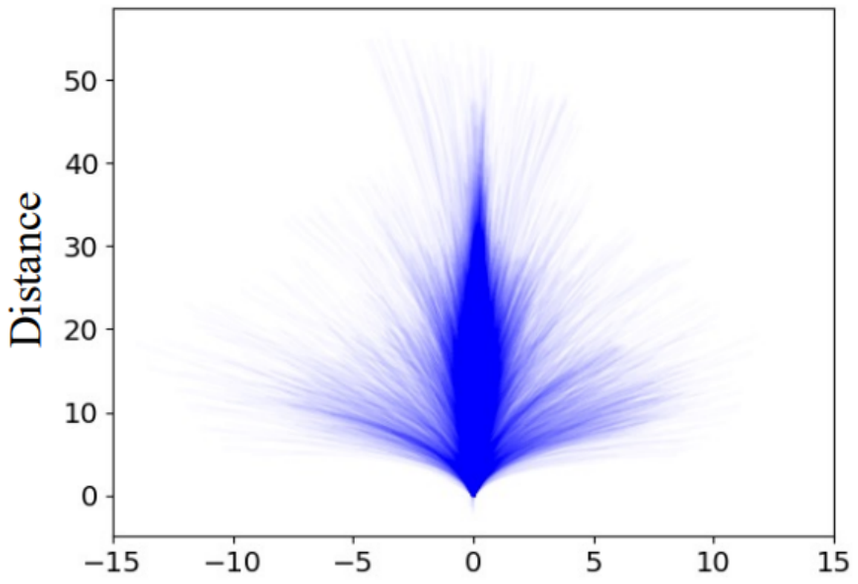
# Rethinking the Open-Loop Evaluation of End-to-End Autonomous Driving in nuScenes

- Input: Ego States (velocity, acceleration, past trajectories) w/o High-level Command (go straight, turn left, turn right)
- Output: Future Trajectories ([x, y, yaw] for future 6 frames corresponding to 3s)
- Model: simple MLP with two hidden layers

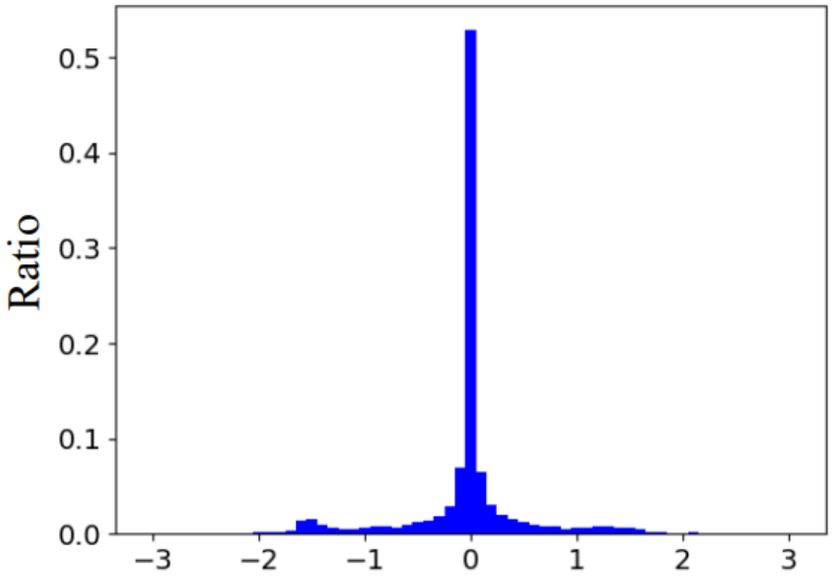


# nuScenes: Ego vehicles move along straight lines and at small angles

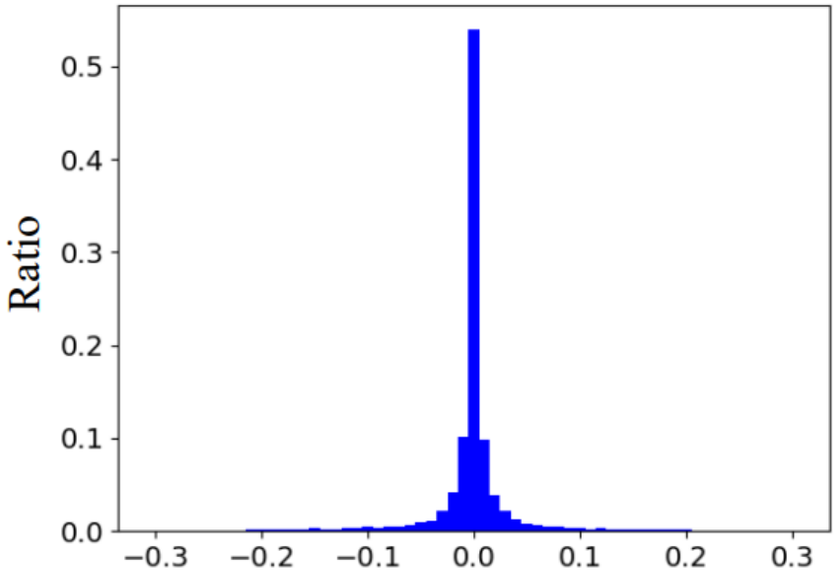
- The offset, angle, and their change rates in the lateral direction of the sample trajectories are relatively small.
- Ego vehicles tend to move forward along straight lines and at small angles during driving in short-time horizons (3s).




(a) Trajectory Points




(b) Heading Angle

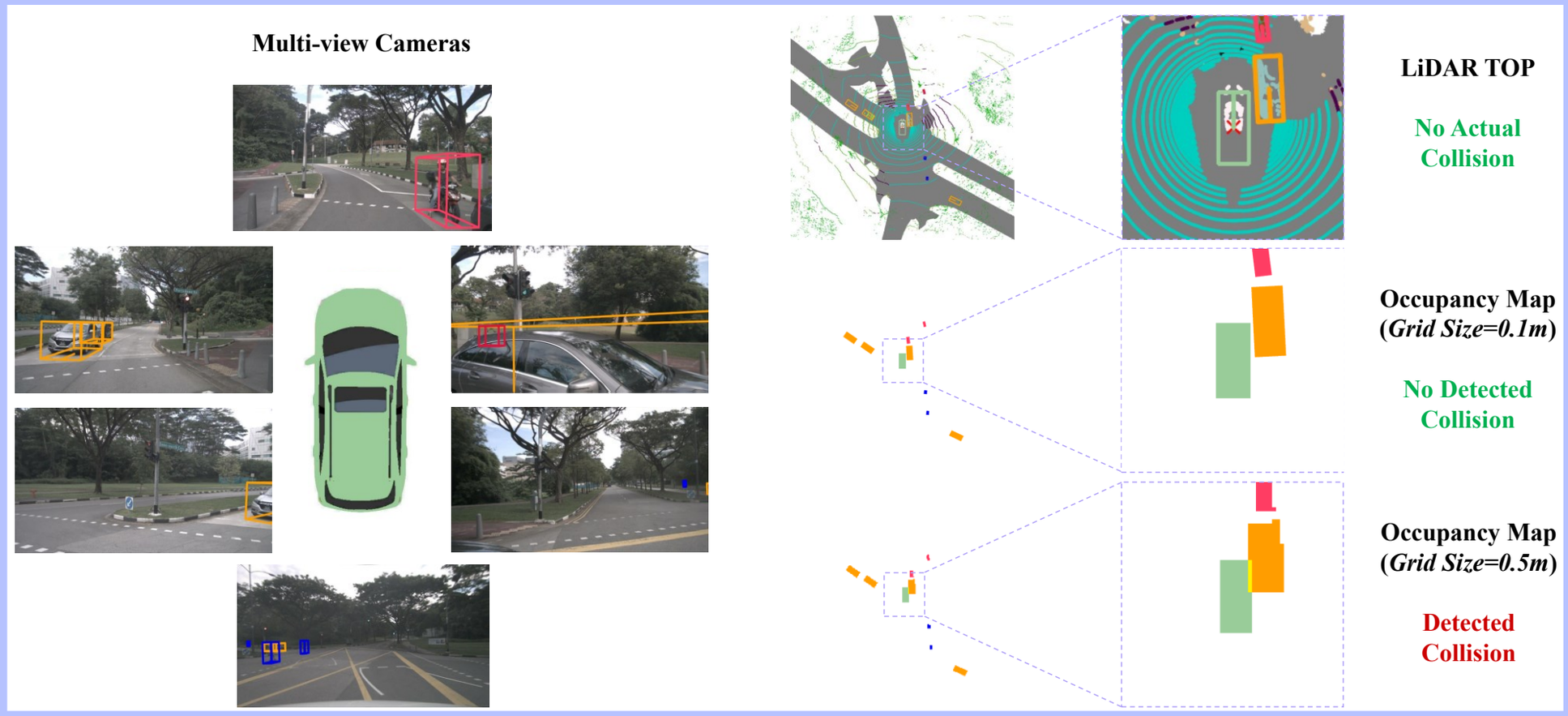



(c) Curvature Angle

Jiang-Tian Zhai, et.al., Rethinking the Open-Loop Evaluation of End-to-End Autonomous Driving in nuScenes. CoRR abs/2305.10430 (2023).

# Using Occupancy Map to Calculate Collision: Grid Size Matters

- Using occupancy map to calculate collision may introduce errors.
- When we set grid size to 0.5m, a collision is detected, however the figure is from GT and actually no collision happens.




Jiang-Tian Zhai, et.al., Rethinking the Open-Loop Evaluation of End-to-End Autonomous Driving in nuScenes. CoRR abs/2305.10430 (2023).

# Results on nuScenes Validation Set

| Method | Perception Information | Ego State | High- level Command | L2(m) ↓ | Collision(%) ↓ |
|---|---|---|---|---|---|
| FF | ✓ | ✗ | ✗ | 1.43 | 0.43 |
| EO | ✓ | ✗ | ✗ | 1.60 | 0.33 |
| ST-P3 | ✓ | ✗ | ✓ | 2.11 | 0.71 |
| UniAD | ✓ | ✗ | ✓ | 1.03 | 0.31 |
| VAD | ✓ | ✓ | ✓ | 0.37 | 0.14 |
| **Ours** | ✗ | ✓ | ✗ | 0.25 | 0.16 |
| | | | ✓ | **0.23** | **0.12** |

- *Despite its simplicity and the absence of perceptual information, the simple model achieves remarkable performance on the nuScenes dataset.*

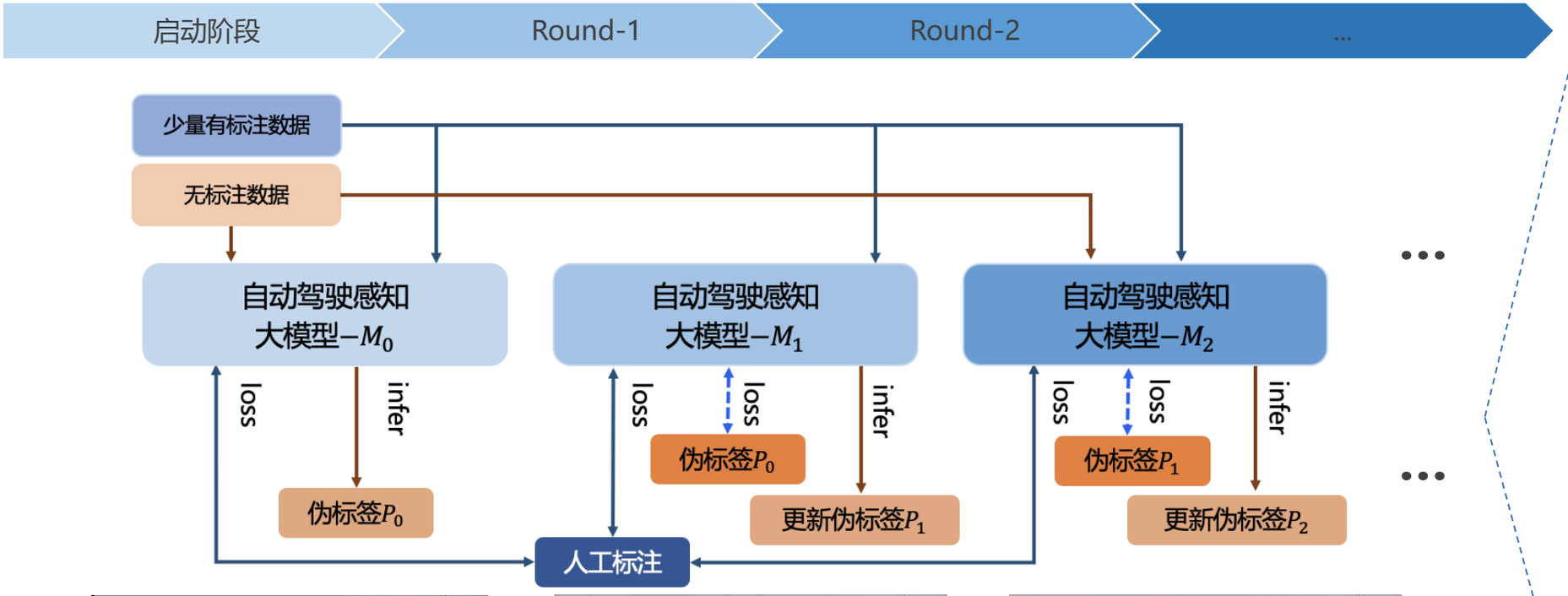
- ***Notes: The current evaluation on nuScenes may not adequately capture the superiority of different methods.***

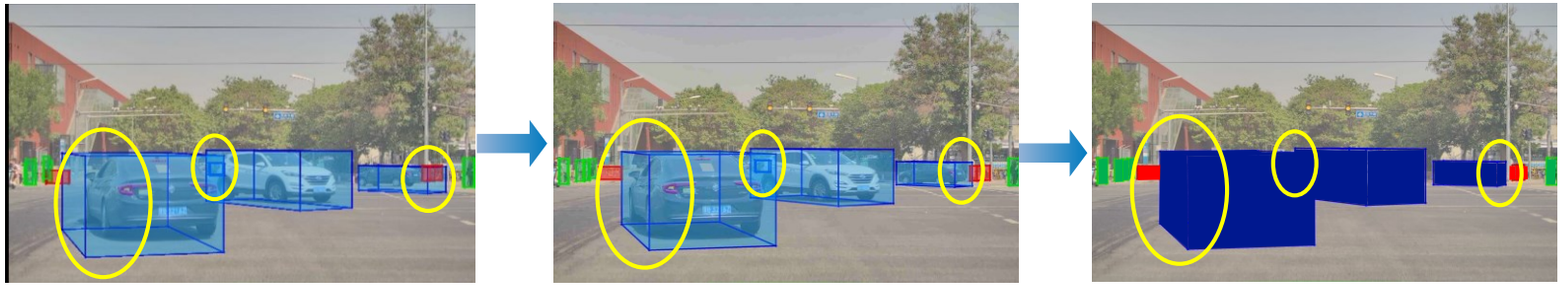
Jiang-Tian Zhai, et.al., Rethinking the Open-Loop Evaluation of End-to-End Autonomous Driving in nuScenes. CoRR abs/2305.10430 (2023).
Ours code：https://github.com/E2E-AD/AD-MLP

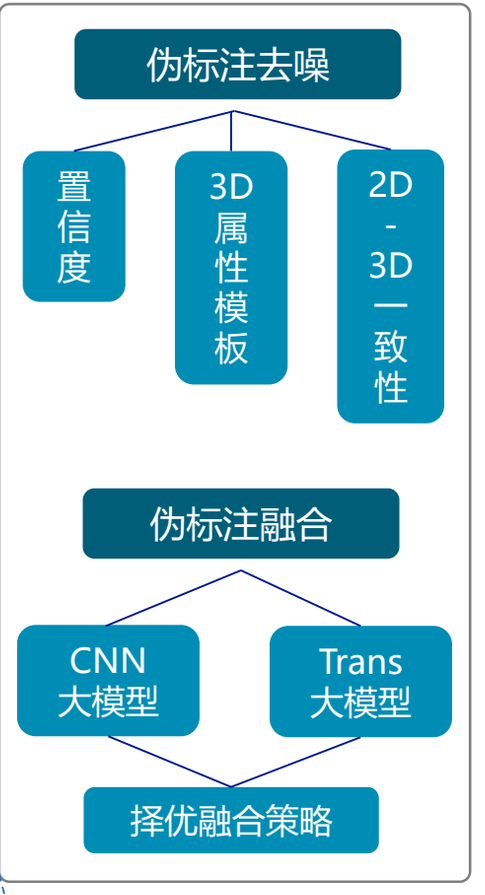
# Large Models for Autonomous Driving

# 自动驾驶感知大模型：Iterated Self-Training

海量未标注数据 ⟶ 迭代自训练：不断优化更新伪标签，扩大数据势能



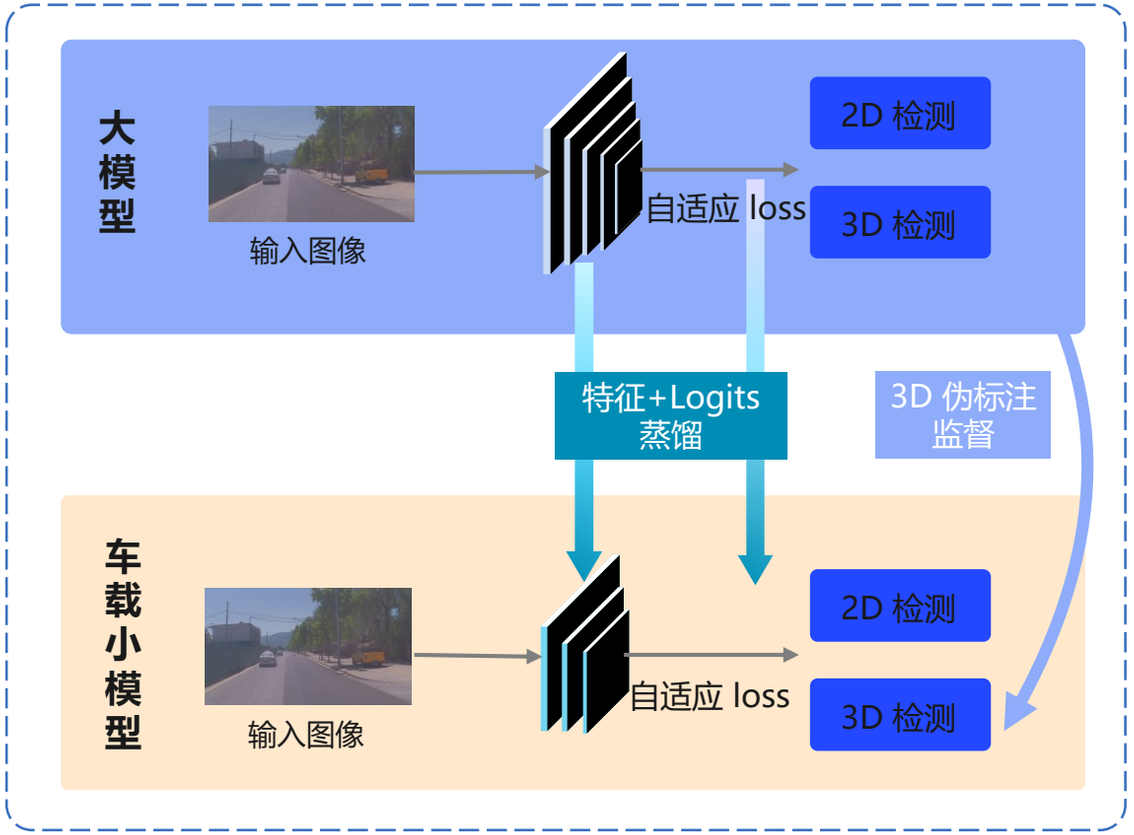



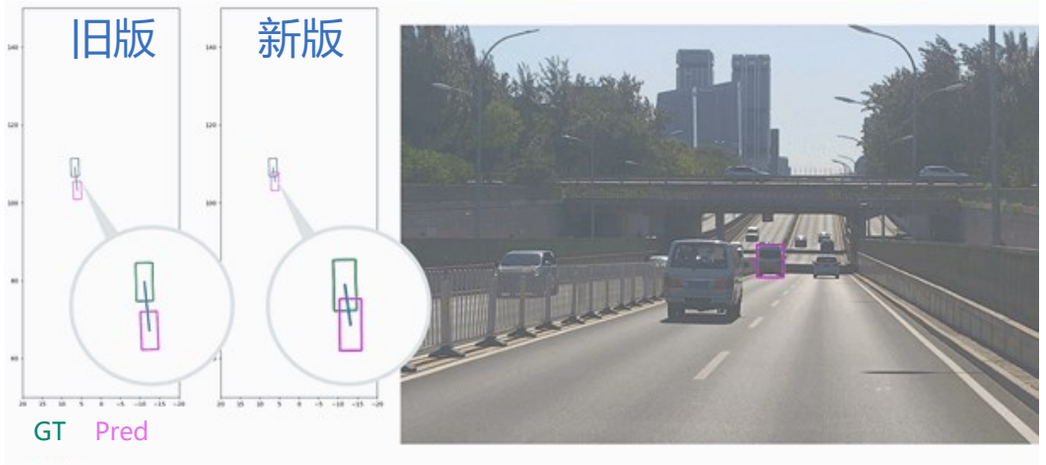


# 自动驾驶感知大模型：大模型帮助小模型

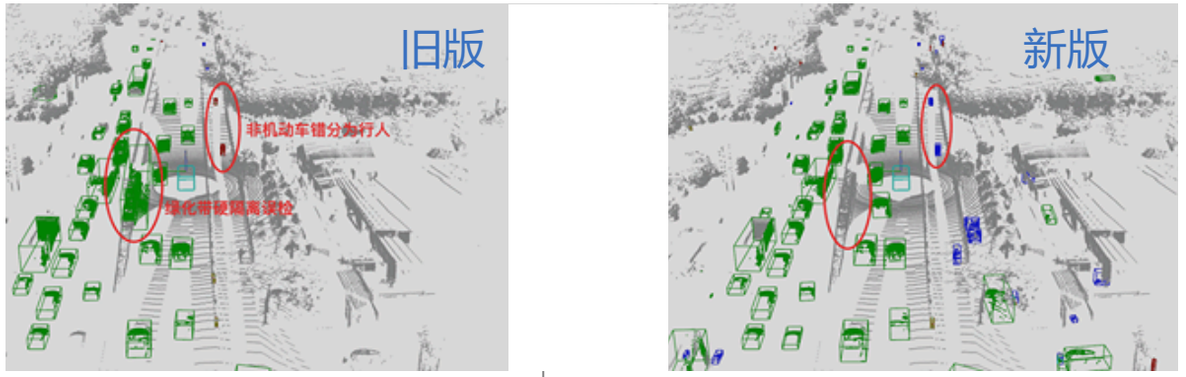
车端算力有限，大模型
无法直接部署上车

「半监督+大模型蒸馏」
进行小型化



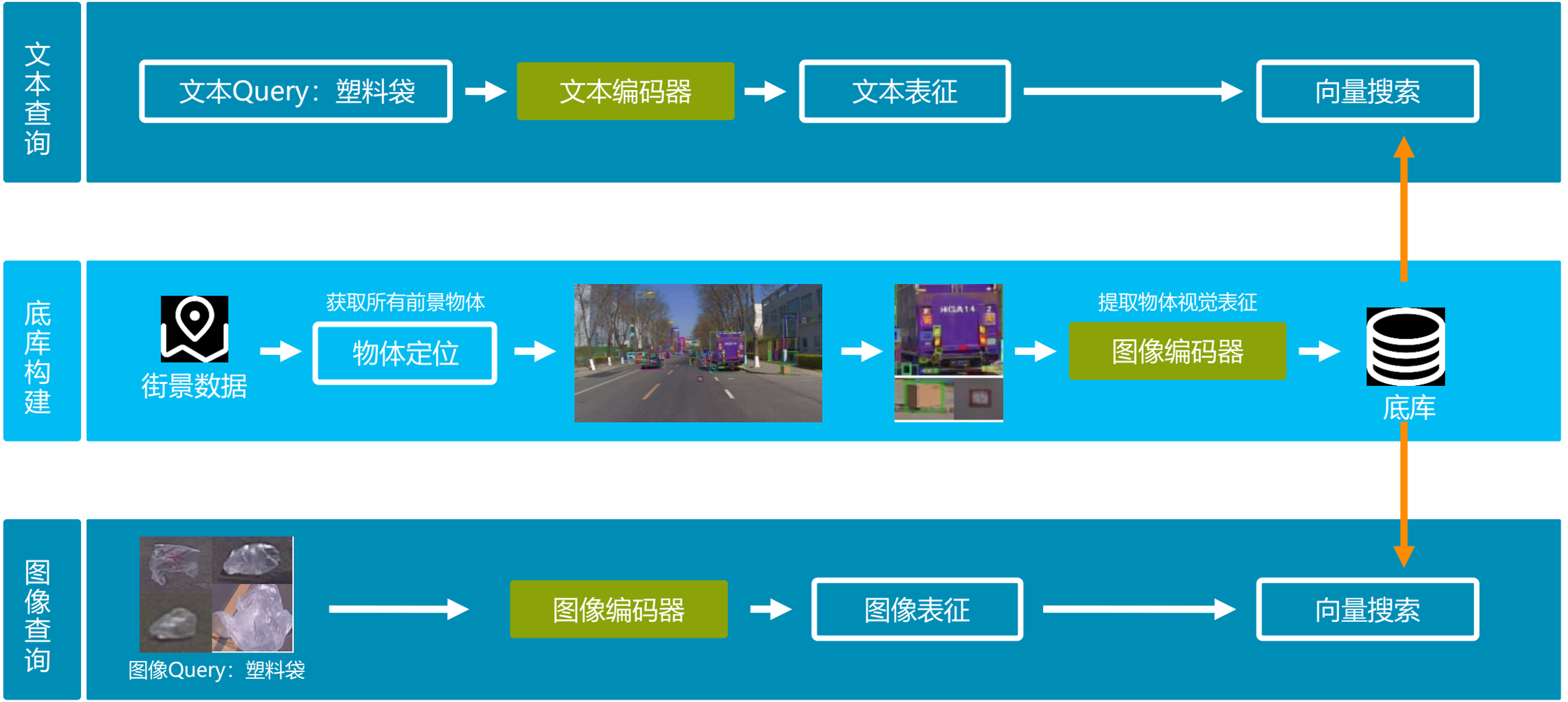

**图像：路面起伏的远视距效果显著提升**




**点云：隔离带误检优化**

# 自动驾驶感知长尾数据挖掘

- 自动驾驶可以利用自动驾驶以外的数据
- 通用图文数据预训练文本、图像编码器，帮助自动驾驶数据挖掘



# 数据挖掘效果示例

**定向挖掘驱动问题解决**

儿童
塑料袋
快递车
…

非刚体异形车
行李箱
…

**定向挖掘驱动能力扩展**

消防车
救护车
猫狗
…

施工人员
…

# Discussions

Transformer:
Dominant in vision?

Planning:
Joint with prediction?

Planning:
Learning vs rule, evaluation

Large model:
Vision is dead?

End-to-End:
L2 and L4?

Large models:
AD Future?

# Codes are Available

OCRNet

https://github.com/HRNet

Conditional/Group DETR, DWNet, CAE

https://github.com/Atten4Vis

VIMER: Vision Foundation model

https://github.com/PaddlePaddle/VIMER

CAPE

https://github.com/PaddlePaddle/Paddle3D

# Thanks!