JUNE 18-22, 2023

CVPR

VANCOUVER, CANADA

MonoATT: Online Monocular 3D Object Detection with Adaptive Token Transformer

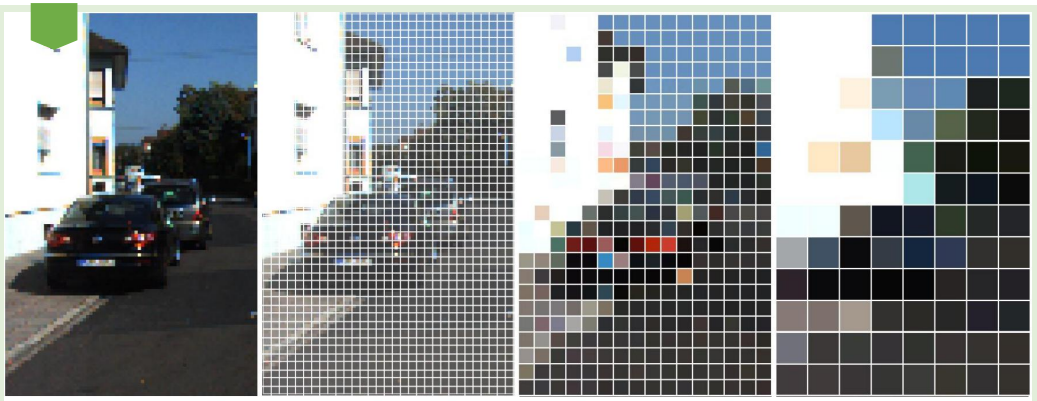
Yunsong Zhou[1], Hongzi Zhu[1], Quan Liu[1], Shan Chang[2], Minyi Guo[1]

[1]Shanghai Jiao Tong University, [2]Donghua University

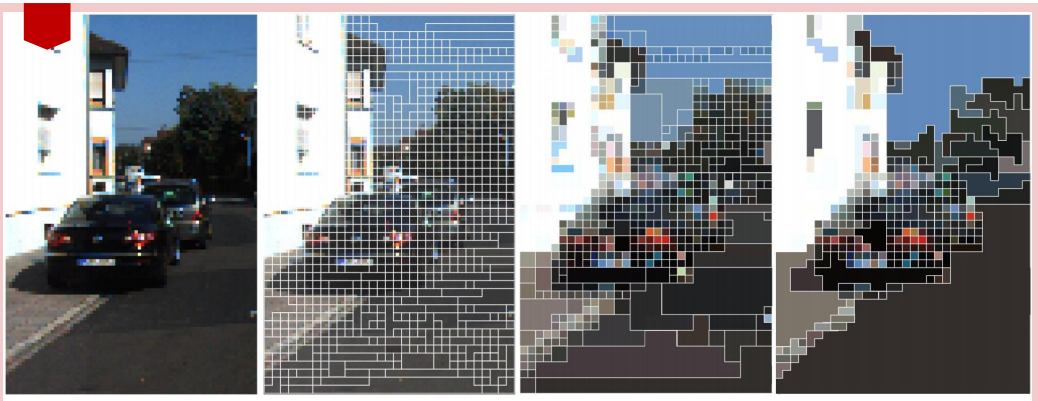
THU-AM-097

Hightlights

- Existing transformer-based offline Mono3D models adopt **grid-based** vision tokens, which are suboptimal when using coarse tokens due to the limited available computational power.
- MonoATT leverages a novel transformer with **heterogeneous** tokens of varying sizes and shapes to boost mobile Mono3D.



(a) Grid-based tokens in multiple stages



(b) Heterogeneous tokens in multiple stages

Background

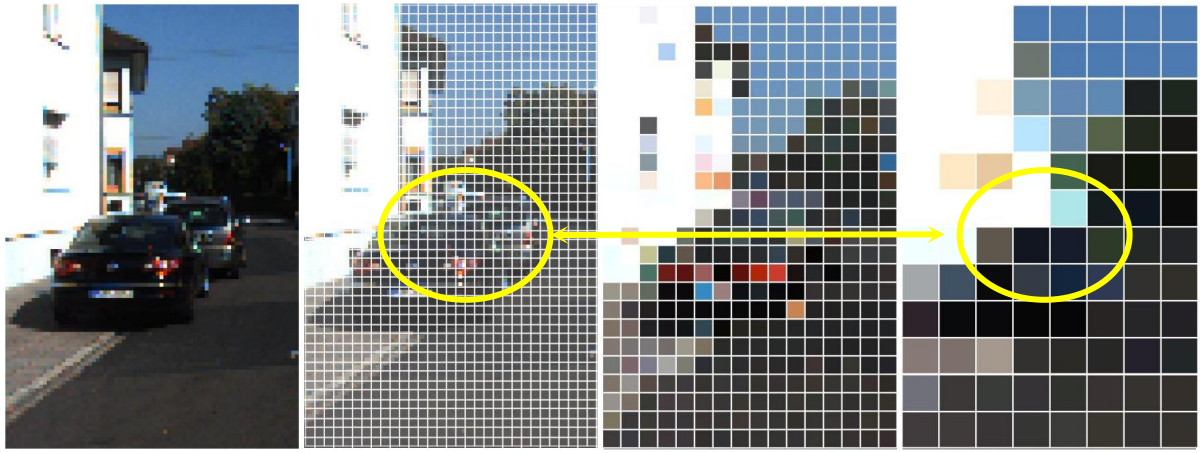
- Monocular 3D object detection (Mono3D) in mobile settings (e.g., on a vehicle, a drone, or a robot) is an important yet challenging task.
- Requirments:
  - Wide-area perception.
  - Low response time.

Background

- CenterNet based methods: conceptually straightforward, low computational overhead, insufficient to understand the scene-level geometric cues.
- Transformer based methods: long-range attention, SOTA performance, grid-based token generation.

Background

- Using grid-based tokens is sub-optimal for Mono3D applications such as autonomous driving because of the following two reasons:
  - 1) ***Far*** objects have smaller size and less image information, which makes them hard to detect with coarse grid-based tokens;
  - 2) Using fine grid-based tokens is prohibitive due to the limited computational power and the stringent ***latency*** requirement.



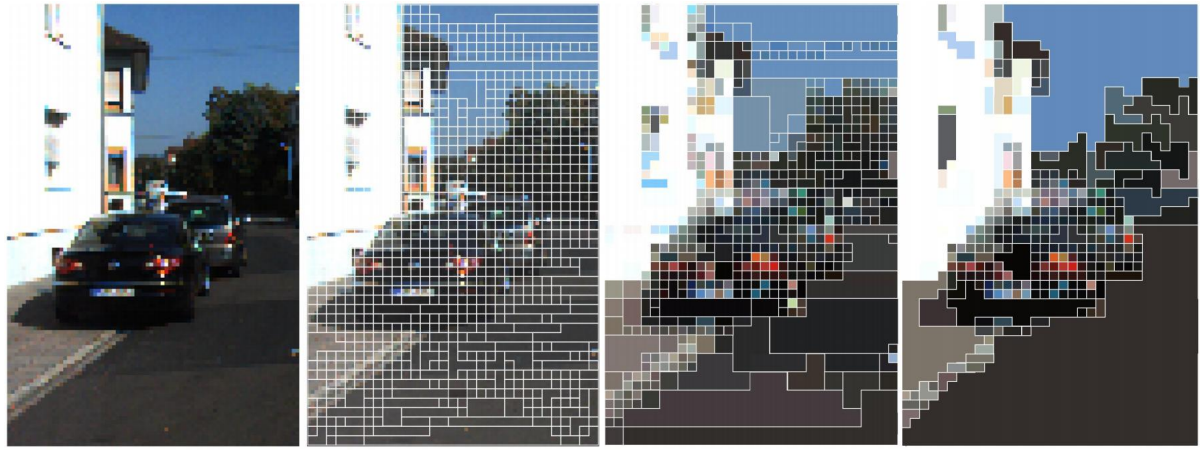
(a) Grid-based tokens in multiple stages

Key Insight

- Observation: not all image pixels of an object have equivalent significance with respect to Mono3D.
  - Depth knowledge: distant objects should be paid more attention to.
  - Semantic Knowledge: features of targets are more valuable than those of backgrounds, and outline features are more crucial than inner features.

Core Idea

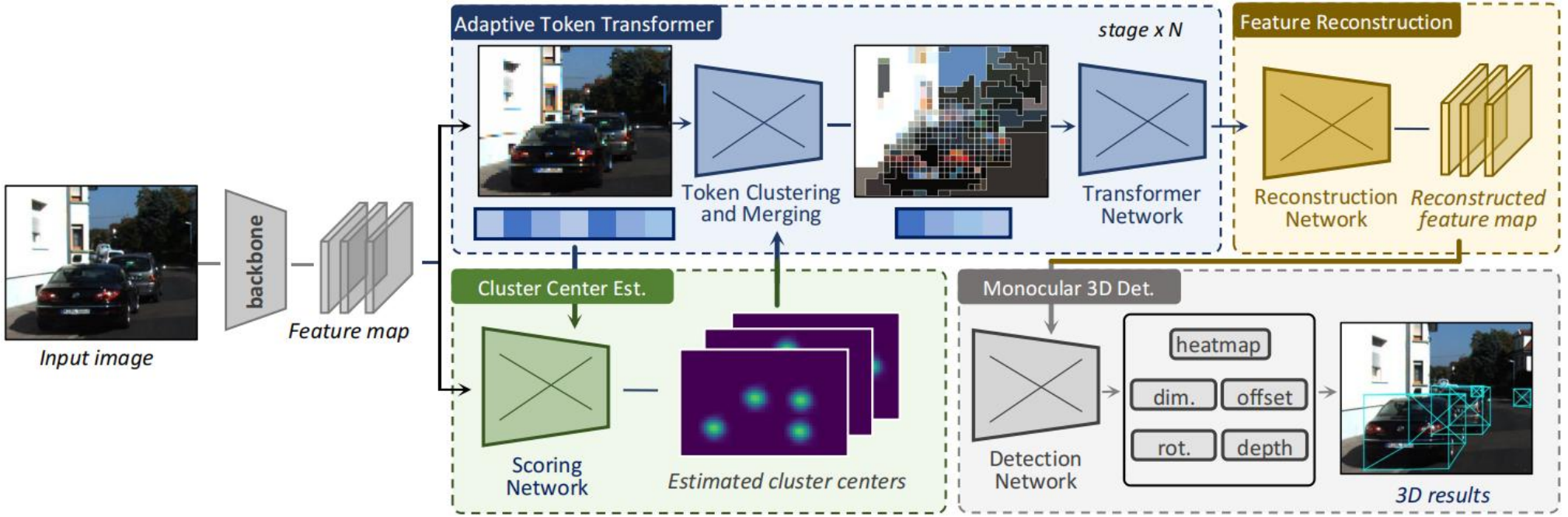
- Core idea of MonoATT:
  - 1) Automatically assign fine tokens to pixels of more significance before utilizing a transformer to enhance Mono3D detection.
  - 2) Dynamically cluster and aggregate image patches with similar features into heterogeneous tokens in multiple stages.



(b) Heterogeneous tokens in multiple stages

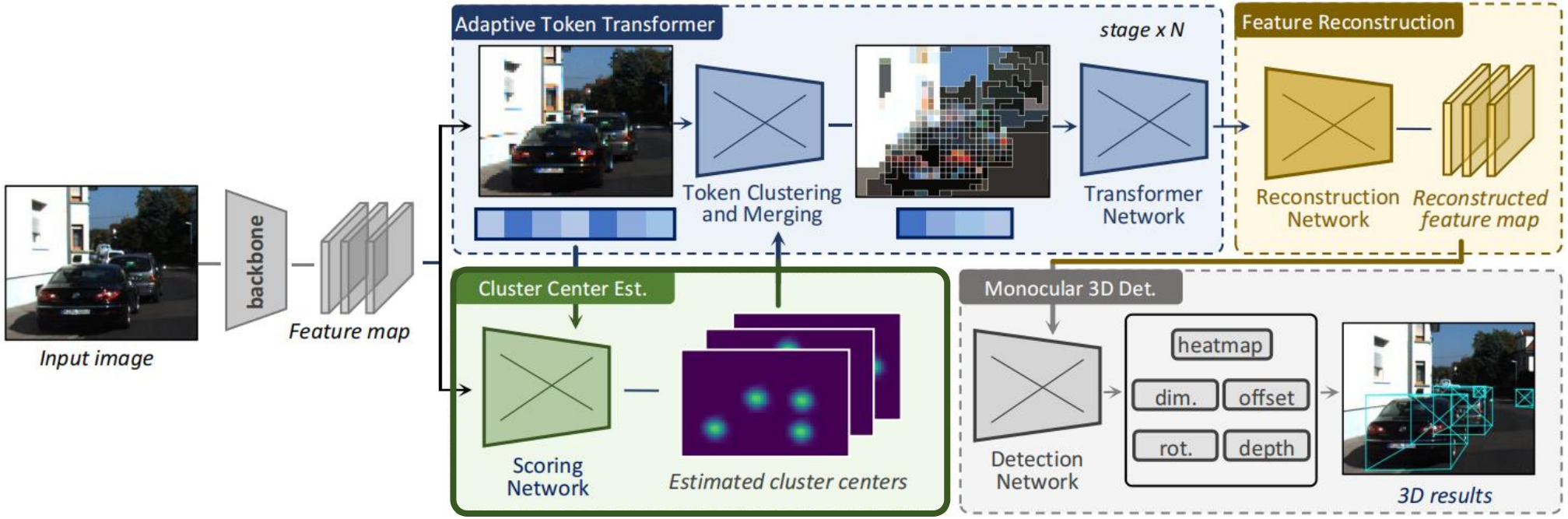
Overview

- Goals: 1) superior image features are obtained from coarse to fine to increase Mono3D accuracy for both near and far objects; 2) irrelevant information (e.g., background) is cut to reduce the number of tokens to improve the timeliness of the vision transformer.



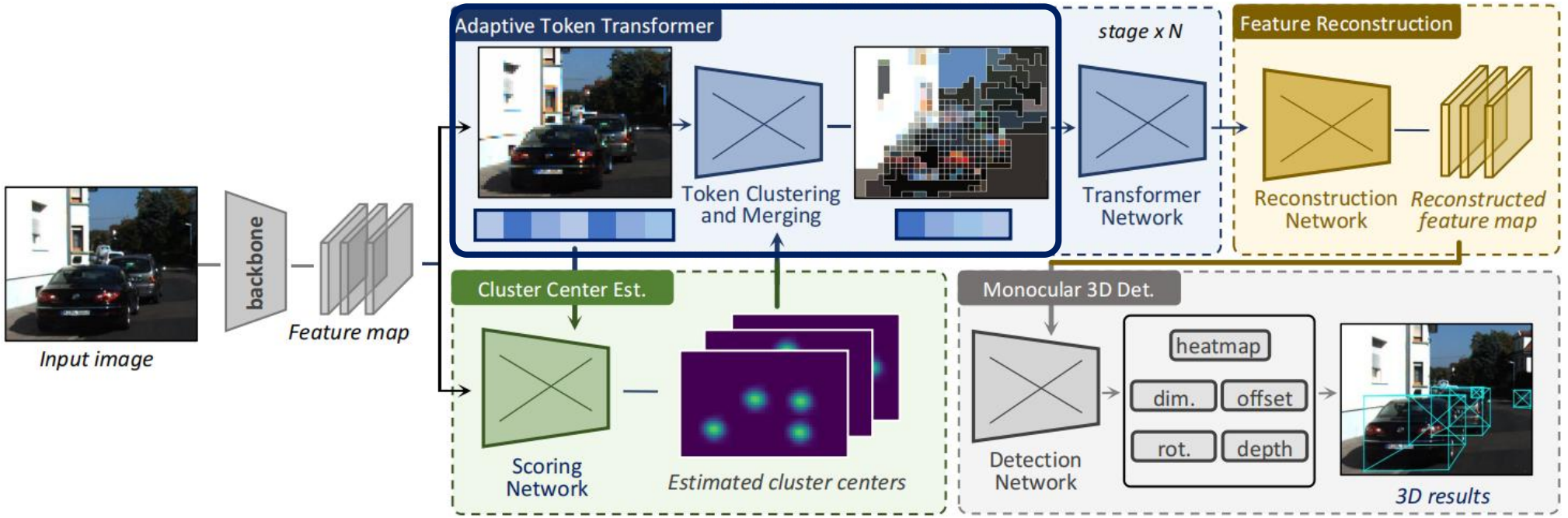
Overview

- **Cluster Center Estimation (CCE).** CCE leverages a scoring network to pick out the most crucial coordinate point locations from monocular images that are worthy of being used as cluster centers based on the ranking of scores and quantitative requirements in each stage.



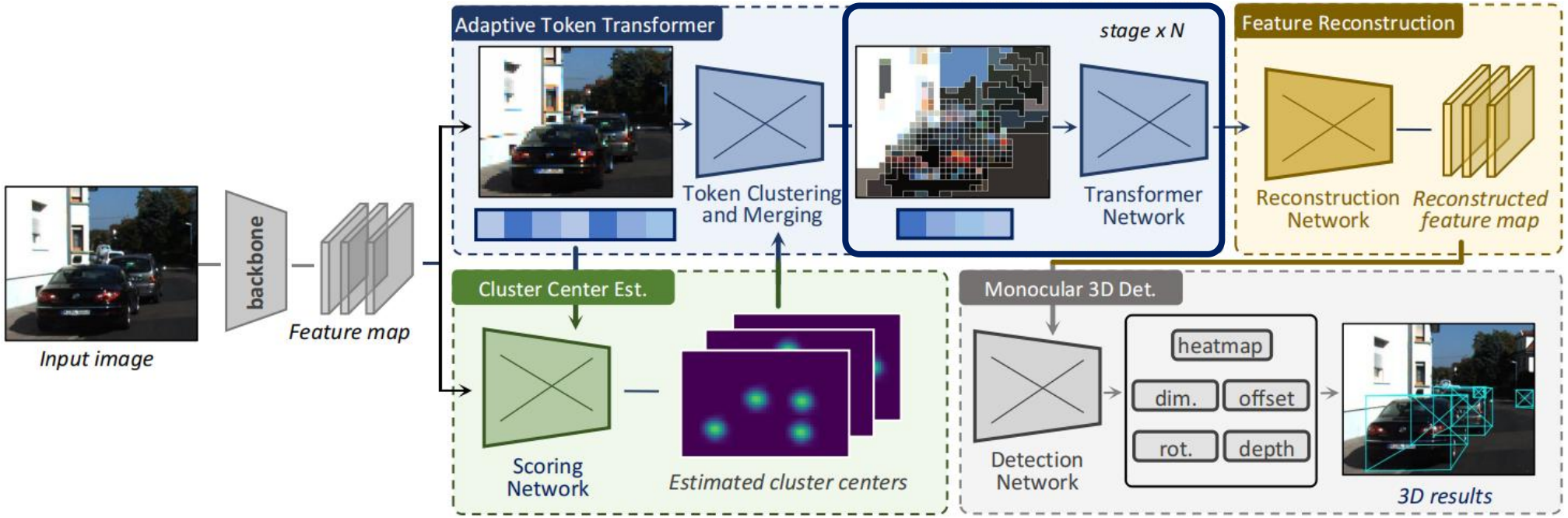
Overview

- **Adaptive Token Transformer (ATT).** Starting from the initial fine grid-based tokens obtained by slicing the feature map and the selected cluster centers, ATT groups tokens into clusters and merges all tokens within each cluster into one single token in each stage.



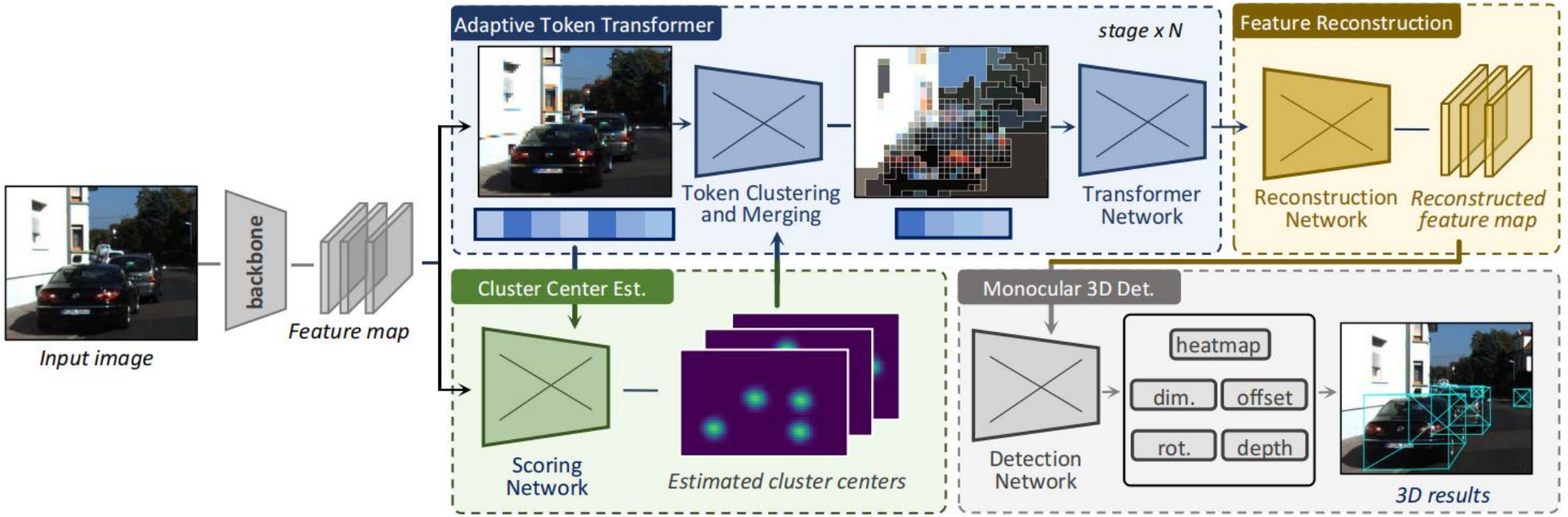
Overview

- After that, a transformer network is utilized to establish a long-range attention relationship between adaptive tokens to enhance image features for Mono3D. The ATT process is composed of *N* stages.



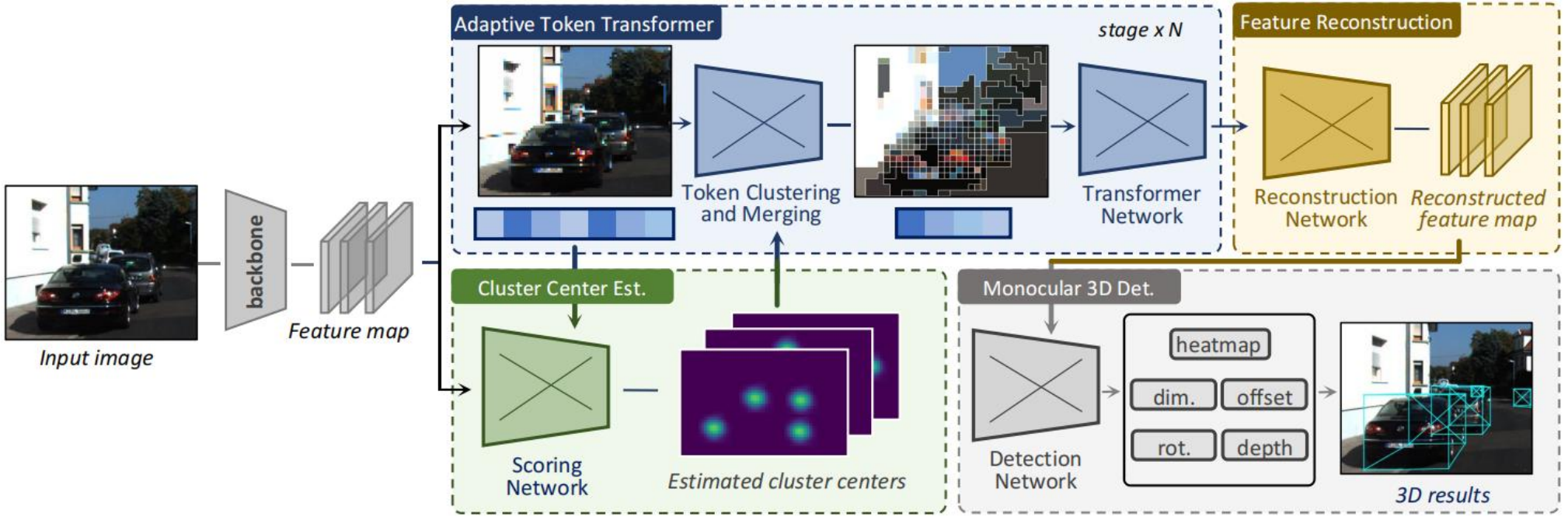
Overview

- **Multi-stage Feature Reconstruction (MFR).** MFR restores and aggregates all *N* stages of irregularly shaped and differently sized tokens into an enhanced feature map.



Overview

- **Monocular 3D Detection (M3D).** MoGDE employs GUPNet, a SOTA CenterNet based monocular 3D object detector as its underlying detection core.



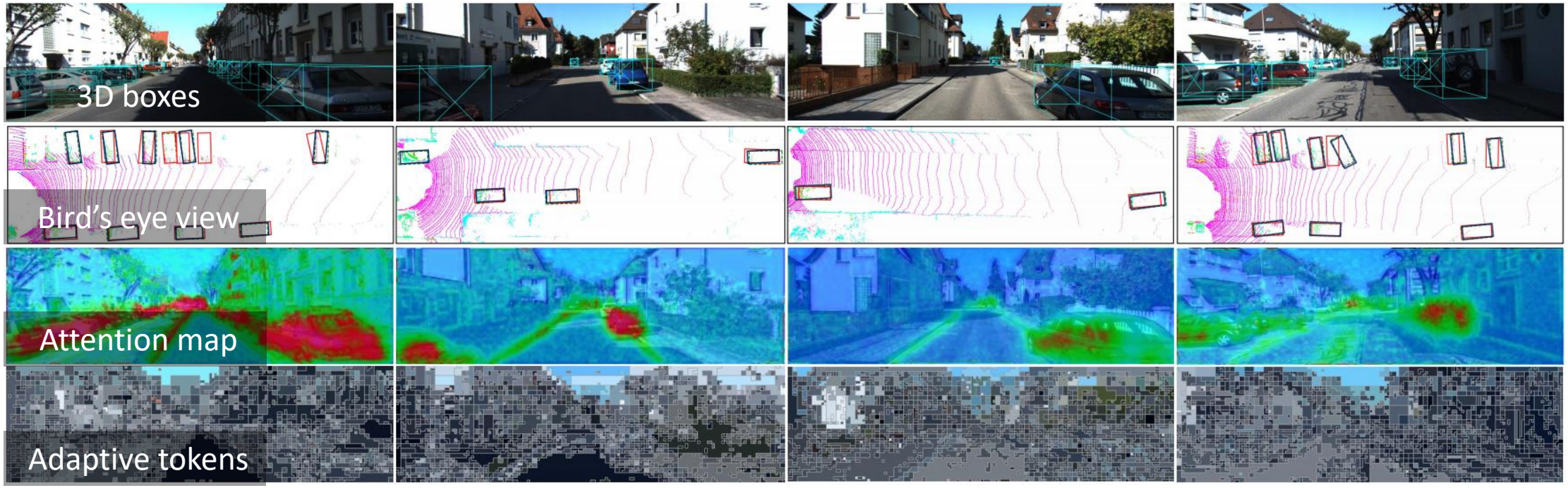
Experiments

- For the official *test* set of KITTI 3D benchmark, we achieve the highest score for all kinds of samples.

Method	Extra data	Time (ms)	Test, AP_{3D}			Test, AP_{BEV}		
			Easy	Mod.	Hard	Easy	Mod.	Hard
PatchNet [29]	Depth	400	15.68	11.12	10.17	22.97	16.86	14.97
D4LCN [15]		200	16.65	11.72	9.51	22.51	16.02	12.55
Kinematic3D [3]	Multi-frames	120	19.07	12.72	9.17	26.69	17.52	13.10
MonoRUn [7]	Lidar	70	19.65	12.30	10.58	27.94	17.34	15.24
CaDDN [40]		630	19.17	13.41	11.46	27.94	18.91	17.19
AutoShape [27]	CAD	-	22.47	14.17	11.36	30.66	20.08	15.59
SMOKE [26]	None	30	14.03	9.76	7.84	20.83	14.49	12.75
MonoFlex [65]		30	19.94	13.89	12.07	28.23	19.75	16.89
GUPNet [28]		40	20.11	14.20	11.77	-	-	-
MonoDTR [19]		37	21.99	15.39	12.73	28.59	20.38	17.14
MonoDETR [64]		43	23.65	15.92	12.99	32.08	21.44	17.85
MonoATT (Ours)	None	56	**24.72**	**17.37**	**15.00**	**36.87**	**24.42**	**21.88**
Improvement	*v.s. second-best*	-	+1.07	+1.45	+2.01	+4.79	+2.98	+4.03

Experiments

- Compared with the baseline, the predictions from MonoATT are much closer to the ground truth, especially for distinct objects.



JUNE 18-22, 2023

CVPR

VANCOUVER, CANADA

THANK YOU!

Yunsong Zhou[1], Hongzi Zhu[1], Quan Liu[1], Shan Chang[2], Minyi Guo[1]

[1]Shanghai Jiao Tong University, [2]Donghua University

上海交通大學

東華大學