Accelerate Multi-Camera Perception in Bird's-Eye-View Space by Input Patchification and Content-Aware Pruning

Anonymous ICCV submission

Paper ID ****

Abstract

The Bird's-Eye-View (BEV) perception paradigm has recently garnered increasing attention among autonomous driving practitioners due to its superior performance in downstream tasks when compared to traditional late fusion approaches. This paradigm involves learning a shared representation in the top-down view from surrounding cameras, but its computational cost is considerably higher than that of monocular-camera algorithms, as the backbone model must attend to all side-view cameras. To address this challenge and facilitate the practical application of BEV-paradigm methods in real-world vehicles, we propose a method to accelerate inference speed and reduce model complexity by patchifying input and pruning out task non-relevant and less-confident regions. With a simple yet efficient design, our pruning technique achieves comparable performance to state-of-the-art methods on the nuScenes [2] leaderboard. To the best of our knowledge, this is the first work to accelerate BEV perception tasks from the backbone optimization perspective.

1. Introduction

Pure vision multi-camera perception in Bird's-Eye-View (BEV) space [8] [7] [17] [44] [20] [38] [15] [28] has garnered significant attention among autonomous driving practitioners. Learning a shared representation in the top-down view from surrounding cameras and leveraging it to perform downstream tasks has demonstrated superior performance compared to traditional late fusion-based approaches [13] [41], as evidenced by their placement atop various public leaderboards [2] [37] [45] [3]. As such, BEV-based methods are widely regarded as the next-generation perception paradigm for autonomous driving. By projecting 2D input from the camera domain into the 3D domain, BEV space offers a smooth and robust representation to fuse multi-camera inputs and undertake downstream tasks. This approach effectively leverages the spatial relationships between different camera views, while capitalizing on the strengths of top-down representation for efficient feature extraction and downstream task execution. The utilization of multi-camera BEV-paradigm methods is a promising avenue for enhancing the perceptual capabilities of autonomous vehicles.

However, the adoption of such methods is impeded by their significant computational overhead, which is typically beyond the capabilities of on-board computing systems. These methods [22] rely on inputs from six or more surrounding cameras, which are processed by a backbone model for semantic feature extraction, and then projected onto bird's-eye-space using a view projector. Subsequently, a BEV feature encoder is applied to fuse the spatial and temporal information before the task heads generate the desired output. Notably, the backbone model utilized in multi-camera BEV-paradigm methods generates at least six times more computational complexity than monocular-camera algorithms. Although the use of multiple camera views in the feature extraction stage may be advantageous, it is imperative to recognize that not all pixels carry equal significance for downstream tasks. Thus, the challenge lies in developing efficient algorithms that can effectively leverage the spatial and temporal information offered by multiple camera views, while minimizing the computational burden associated with such approaches.

Inspired by the attention mechanism of the human visual system, we propose a novel approach for accelerating the inference speed and reducing the computational cost of BEV-based methods by selectively cropping out task-irrelevant regions from the input. While perception algorithms are expected to detect all visible stuff and objects in the driving scene, motion and behaviour planning modules [42] [30] of autonomous vehicles prioritize the detection of stuff and objects that may impact the function and safety of the vehicle, while disregarding redundant stuff and objects that have minimal interactions. In this regard, non-interest regions, such as the sky, building facades, and static objects from afar, can be eliminated to reduce computational cost without compromising performance or interfering with the

behavior of autonomous vehicles. By prioritizing the detection of regions that are critical for the safety and efficacy of autonomous systems, our proposed approach allows for a more streamlined and efficient perception pipeline, providing a promising avenue for enhancing the performance of autonomous vehicles.

Traditional Region-of-Interest (ROI) cropping methods are often designed heuristically by engineers, rather than being derived statistically from data. Such methods employ one or more geometry-based region selectors to identify regions of images that are deemed more important than others. However, the hard cropping of ROI can lead to imprecise boundary delineation between objects and their surroundings, thereby negatively affecting the performance of detection algorithms. To circumvent this challenge, we replace the hard ROI cropping method with a novel task-aware ROI cropping method, which leverages a confidence-driven filtering module to expunge non-interest regions. Specifically, we partition the input frames into tokens and remove tokens whose confidence scores fall below a predefined percentage threshold. The confidence score of each token is learned end-to-end, concurrently with the training process of 3D detection tasks. The proposed method is demonstrated to be effective in a series of ablative experiments, as even with a 40% token dropout, we achieve comparable performance with state-of-the-art methods.

We summarise our contributions below:

- We proposed a domain-specific pruning method that can reduce the model complexity by 40%. To the best of our knowledge, this is the first work to investigate token prunning technique for perception tasks in bird's-eye-space. Token pruning has been a popular topic of reducing the computational cost of ViT [4] model; even though these methods claim to be a universal solution to all vision task, domain-specific acceleration methods for BEV has yet been proposed.

- We empirically studied the impact of various state-of-the-art pruning methods for 3D detection task on nuScenes dataset. We show that without sophisticated design, a topk algorithm is sufficiently good to remove redundant features and accelerate model inference speed with marginal performance drop.

2. Related Work

2.1. Perception in BEV Space

Accurate and robust surrounding understanding has been a long-standing task for autonomous driving. Multi-modality algorithms with cameras and LiDAR have been widely adopted to acquire reliable perception results for autonomous vehicles. However, recent advancement in BEV perception has embarked on a new fashion in pure vision solutions.

Camera to Bird's-Eye-View Projection The fundamental component of BEV-paradigm methods is the camera to BEV view projector. As summarized in [22], view projectors can be divided into two major categories, geometry-based and network-based. Geometry-based methods [23] [11] [33] leverage the geometric transformation relationship to project perspective view input into bird's eye view space. Some recent methods [29] [34] [32] also proposed to use depth information to enhance the transformation, as the ground plane may not present in an ideal flat form as assumed in previous methods. Network-based methods [21] [26] [36] utilise the magic of big data to learn an implicit mapping function between the camera input and feature map in BEV space.

3D Detection in BEV 3D detection has been the long-standing fundamental task to autonomous driving since the advent of this subject. Many methods have been developed to improve the performance of 3D detection on top of shared BEV features. BEVFormer [17] proposed to exploit spatial-temporal information on BEV feature map to improve the performance of 3D detection task. With the input of CAN-BUS and proper alignment of cross-camera and historic feature map, BEVFormer proved that pure-vision 3D detection algorithm can outperform LiDAR-based baselines. BEVDet [8] and BEVDet4D [7] further pushed the boundary of 3D detection performance on nuScenes leaderboard by 3D-NMS and adding temporal cue with caching BEV feature of last frame. By eliminating ego-motion and time factor in the learning target, BEVDet4D reduced more than 60% of velocity error and effectively improved NDS on nuScenes dataset by a large margin. BEVDepth [15] is another line of work in 3D detection tasks; by introducing a novel depth acquisition network, BEVDepth reached 60 NDS for the first time on the nuScenes test set.

Trajectory Predictions in BEV Future motion prediction is an essential building block of autonomous driving. With the correct and in-time prediction of objects' trajectory on the road, the planning module generates the routes and behaviour of autonomous vehicles for the next few seconds. FIERY [6] proposed to carry prediction task in BEV space with the combination of probabilistic model and feature map in BEV space; for the first time, trajectory prediction, perception and sensor fusion are achieved in an end-to-end manner.

Segmentation in BEV In the field of map segmentation, HDMapNet [14] proposed a semantic map learning method to predict vectorized map elements in the bird's-eye view. Surrounding camera inputs are first encoded by vision backbone and then projected to BEV space, with a semantic-level and instance-level learning metric devised to enhance the learning performance, HDMapNet outperforms

existing baseline models by more than 50%. VectorMapNet [19] is another concurrent literature that explores the viability of predicting vectorized maps from BEV feature map. Different from HDMapNet, VectorMapNet predicts a sparse set of polylines instead of densed segmentation mask. BEVSegFormer [28] proposed to acquire real-time segmentation map with a transformer model. Camera-to-segmentation mask mapping is implicitly learned in an end-to-end manner, making the BEVSegFormer model robust to camera noises, making reliable road structure acquisition available.

Multi-Task Learning in BEV Apart from task-specific models, BEV can also serve as a smooth representational space for multi-task learning. BEVerse [44] and M2BEV [38] both focus on 3D perception and prediction tasks. While BEVerse produces spatiotemporal Birds-Eye-View representations from multi-camera videos and outperforms existing single-task methods, M2BEV presents a unified framework for joint 3D object detection and map segmentation in the BEV space with multi-camera image inputs, achieving state-of-the-art results.

Sensor Fusion in BEV Sensor fusion in Bird's Eye View (BEV) space is a rapidly evolving area of research, characterized by a diverse range of investigations aimed at optimizing multi-sensor fusion techniques. BEVFusion [20] presents an efficient multi-task multi-sensor fusion framework that seamlessly integrates multimodal features within the shared bird's-eye view representation space. SimpleBEV [5] , on the other hand, explores the viability of fusing radar and camera input in BEV space, employing a reduction methodology to investigate the impact of various factors in the design and training protocols of BEV perception models. Although radar input leads to a non-marginal drop in performance, SimpleBEV contributes a novel approach to exploiting BEV space as an intermediate representation for conducting sensor fusion. These investigations demonstrate a growing interest in advancing sensor fusion techniques in BEV space, which will likely result in further innovations in the field of autonomous systems.

2.2. ViT Compression Methods

The Vision-Transformer (ViT), introduced in the literature [4], has demonstrated remarkable efficacy in various visual tasks. However, the quadratic complexity arising from multi-head attention calculation poses a significant challenge when deploying ViT on resource-limited platforms. To tackle this challenge, a plethora of model compression techniques have been extensively explored.

Magnitude Based Token Pruning Many researchers have attempted to compress ViT models by utilizing token magnitudes, which refer to attention scores. It is widely accepted that patchified tokens from natural images contain redundancies due to the inherent redundancy in images. To address this, researchers have processed tokens in various ways. Some researchers have chosen to delete tokens based on their attention value; if the value is below a hard-defined threshold, they will delete these attention nodes. Other researchers choose to set the value of the token to zero instead of deletion. The selection criteria also vary between studies. Some researchers select tokens directly based on the attention score, while others dynamically select tokens using a specially designed algorithm or a simple neural network to predict their importance. For example, LTP [10] proposed adaptively removing tokens whose attention scores fall below a threshold, and TokenLearner [35] proposed using a mining algorithm to keep important tokens. SPViT [12] proposed learning a selector to divide instance-wise tokens into two groups: frequently attended tokens are processed using the original ViT model, while less important tokens are grouped together and processed by a lightweight model. AViT [40] proposed a halting mechanism to reduce the number of tokens in vision transformers. The halting algorithm is designed with a distributional prior regularization method that stabilizes the training process. DynamicVIT [31] proposed a dynamic token sparsification framework to identify and progressively prune redundant tokens based on the input. All of these works utilize the magnitude of tokens in the attention feature map.

Occurance Based Compression A multitude of techniques have emerged to alleviate the complexity of ViT models through the exploitation of token occurrence or the frequency of token usage in downstream tasks. ToME [1] introduces a compression mechanism that merges tokens with similar values via a similarity-matching algorithm, circumventing the need for pruning. In the event of several pairs of tokens with comparable values, the matching algorithm selects the tokens that are most frequently utilized. EVit [18] proposes a reorganization of image tokens according to their attention scores, splitting frequently attended and less attended tokens into two groups within the multi-head self-attention and feed-forward network, and processing them with distinct branches. MiniViT [43] proposes multiplexing weights between successive transformer blocks, implemented through module parameter sharing and weight distillation over the self-attention layer. TokenPooling [24] presents an innovative token downsampling technique that leverages redundancy in intermediate token representations of images to reduce complexity.

Pipeline Based Compression Several investigations have prioritized a comprehensive optimization strategy for the ViT pipeline, which transcends token-level compression and addresses embedding complexity, attention calculation, dimension size, and MLP layers complexity altogether. For example, EfficientFormer [16] proposes to optimize patch embedding with large kernel sizes and strides and underscores the significance of preserving consistent

feature dimensions. Moreover, the study reveals that Conv-BN is a more efficient option than Layer Normalization. NViT [39] identifies the potential for structural pruning on the embedding dimension, number of heads, MLP hidden dimension, and QKV, providing a mechanism for more effective optimization. ASTAR [27] introduces a novel pruning algorithm leveraging an A* search algorithm to compress ViT models by eliminating redundant attention heads. In addition, IARED2 [25] proposes the utilization of an interpretability-aware redundancy reduction module to compress ViT models, dynamically dropping redundant patches by employing a multi-head interpreter and hierarchical training scheme. These studies all adopt a comprehensive approach, focusing on all submodules of ViT, as opposed to token-level optimizations alone, in order to promote a more effective and efficient Vision Transformer pipeline.

3. Method

In this section Method: Expand DynamicVIT method to 3D domain. Hypothesis: Cross-Camera Correlation has extra redundancy to reduce. Method: Extand DynamicVIT methods to 3D domain. First concat frames of surrounding cameras into a $N \times 6 \times 224 \times 224$ matrix. Then use 3D convolution to learn and fuse feature from the concated matrix, output of this model is $N \times 6 \times M \times M$, the output is deliberately set to be 6, because we need to calcualte spatial-temporal attention later. 3D dynamicVIT reduced 50% of GFlops while keeping marginal drop of performance. The rest keep the same with DynamicVIT.

Patchify shared feature matrix after resnet3D and convert to tokes. Apply dynamicVIT on patchified input to get token features.

Spatial-temporal attention is calculated based on the tokens. Temporal token use history of last 4 seconds. Finally is the 3D detection head.

3D CNN in the first step use resnet3D [9]. The second step use naive vit [4].

During training phase use gumble softmax to calculate the gradients of prediction and pruning layer. During infenrence phase we fix the number of tokens based on the dropout ratio, use topk method to find tokens with highest value and keep. Abalation study: no significant improvement of dropping connection of tokens and directly setting to 0, so we decideto set to zero directly.

In this section, we first review the pipeline of BEV-based percpetion methods on a modular basis and discuss it's advantage and drawbacks.

Next we analyse the pros and cons of token prunning methods for vision transformers.

finally we present the detailed design of our pruning method.

3.1. BEV modules

3.2. Setup

We adopt BEVFormer [17] and BEVDet4D [7] as our baselines.

Backbone model is set to be 3D resnet18. View projector is set to be LSS. SwinT is used to do

4. Experiments

4.1. Results

Comparison to state-of-the-art methods

Performance VS drop out ratio graph

Performance change in terms of dropout ratio w.r.t. different object classes.

4.2. Ablation Study

2D dynamicVIT 3D dynamicVIT Topk pruning Channel pruning. Post training pruning. Hard threshold pruning.

Impact to different c

Different task pruning:

BEVDet, BEVFormer

As shown by figure xx, TODO: Calculate attention token distribution Out pruning methods: dataset level statistics. We prove that without sophiscated pruning method, simple threshold level over dataset is good enough to accelerate the inference speed and maintain a good performance.

We prune out small tokens and retrain the model.

Without training directly use is very good. Training once and run everywhere.

5. Conclusion

We proposed a domain-specific pruning method that can reduce the model complexity by 40%. To the best of our knowledge, this is the first work to inves- tigate token prunning technique for perception tasks in bird's-eye-space. Token pruning has been a popu- lar topic of reducing the computational cost of ViT [4] model; even though these methods claim to be a uni- versal solution to all vision task, domain-specific ac- celeration methods for BEV has yet been proposed

We empirically studied the impact of various state- of-the-art pruning methods for perception task on nuScenes dataset. We show that without sophisticated design, a topk algorithm is sufficiently good to re- move redundant features and accelerate model infer- ence speed with marginal performance drop.

In this work, we systematically analysed the latency and complexity of BEV-based methods, we find that backbone model is the bottleneck of the We proposed a effective method to accelerate the inference speed and reduce the resource loading of the model. We proved that this method is

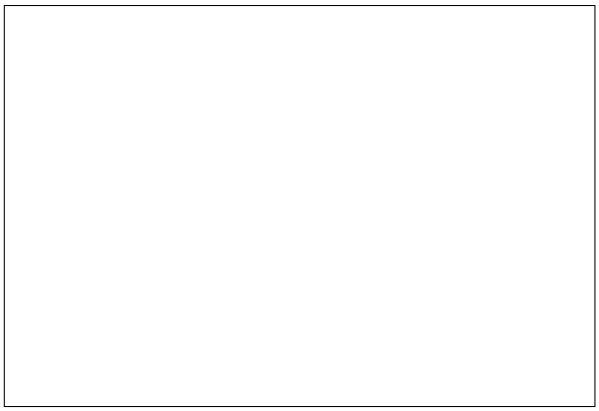

Figure 1. Example of a caption. It is set in Roman so mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 1. Results. Ours is better.

awesome. For future work we will have more experiments on pruning methods, our on the fly pruning threshold estimation algorithm make the model take 2x the time to fully converge, for future work we aim to develop a faster training pipeline

In this work we present a novel approach for compressing BEV-paradigm 3D detection models. By simple baseline for efficient 3D detection in Bird's-Eye-Space, with we find that the bottleneck of algorithm speed is the backbone module by systematical analysis and extensive experiments.

References

[1] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *International Conference on Learning Representations*, 2023. 3

[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. 2020. 1

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020. 1

[4] 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, 3, 4

[5] Adam W. Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception?, 2022. 3

[6] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. FIERY: future instance prediction in bird's-eye view from surround monocular cameras. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 15253–15262. IEEE, 2021. 2

[7] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *CoRR*, abs/2203.17054, 2022. 1, 2, 4

[8] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *CoRR*, abs/2112.11790, 2021. 1, 2

[9] Hirokatsu Kataoka, Tenga Wakamiya, Kensho Hara, and Yutaka Satoh. Would mega-scale datasets further enhance spatiotemporal 3d cnns? *CoRR*, abs/2004.04968, 2020. 4

[10] Sehoon Kim, Sheng Shen, David Thorsley, Amir Gholami, Woosuk Kwon, Joseph Hassoun, and Kurt Keutzer. Learned token pruning for transformers. In Aidong Zhang and Huzefa Rangwala, editors, *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 784–794. ACM, 2022. 3

[11] Youngseok Kim and Dongsuk Kum. Deep learning based vehicle position and orientation estimation via inverse perspective mapping image. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, page 317–323. IEEE Press, 2019. 2

[12] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, Minghai Qin, and Yanzhi Wang. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *ECCV (11)*, pages 620–640, 2022. 3

[13] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12697–12705. Computer Vision Foundation / IEEE, 2019. 1

[14] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online HD map construction and evaluation framework. In *2022 International Conference on Robotics and Automation, ICRA 2022, Philadelphia, PA, USA, May 23-27, 2022*, pages 4628–4634. IEEE, 2022. 2

[15] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *CoRR*, abs/2206.10092, 2022. 1, 2

[16] Yanyu Li, Geng Yuan, Yang Wen, Eric Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Effi-

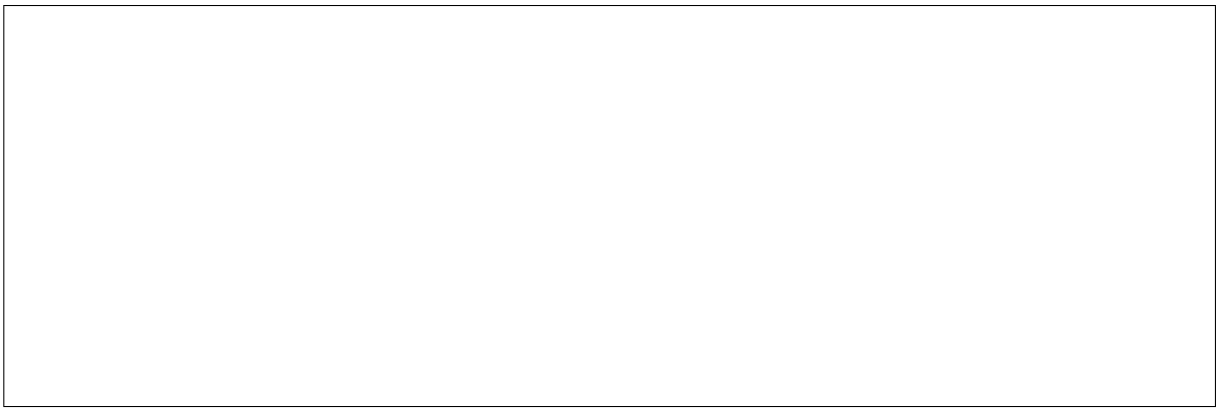

Figure 2. Example of a short caption, which should be centered.

cientformer: Vision transformers at mobilenet speed. *CoRR*, abs/2206.01191, 2022. 3

[17] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX*, volume 13669 of *Lecture Notes in Computer Science*, pages 1–18. Springer, 2022. 1, 2, 4

[18] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *CoRR*, abs/2202.07800, 2022. 3

[19] Yicheng Liu, Yuantian Tuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning, 2022. 3

[20] Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 1, 3

[21] Chenyang Lu, Marinus Jacobus Gerardus van de Molengraft, and Gijs Dubbelman. Monocular semantic occupancy grid mapping with convolutional variational encoder-decoder networks. *IEEE Robotics Autom. Lett.*, 4(2):445–452, 2019. 2

[22] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, Dinesh Manocha, and Xinge Zhu. Vision-centric bev perception: A survey, 2022. 1, 2

[23] Hanspeter A. Mallot, H. H. Bülthoff, J. J. Little, and S. Bohrer. Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biol. Cybern.*, 64(3):177–185, jan 1991. 2

[24] Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers for image classification. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 12–21. IEEE, 2023. 3

[25] Bowen Pan, Rameswar Panda, Yifan Jiang, Zhangyang Wang, Rogério Feris, and Aude Oliva. Ia-red$ˆ2$: Interpretability-aware redundancy reduction for vision transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 24898–24911, 2021. 4

[26] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics Autom. Lett.*, 5(3):4867–4873, 2020. 2

[27] Archit Parnami, Rahul Singh, and Tarun Joshi. Pruning attention heads of transformer models using a* search: A novel approach to compress big NLP architectures. *CoRR*, abs/2110.15225, 2021. 4

[28] Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, and Erkang Cheng. Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 5924–5932. IEEE, 2023. 1, 3

[29] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Proceedings of the European Conference on Computer Vision*, 2020. 2

[30] Rui Qian, Xin Lai, and Xirong Li. 3d object detection for autonomous driving: A survey. *Pattern Recognit.*, 130:108796, 2022. 1

[31] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 3

[32] Cody Reading, Ali Harakeh, Julia Chae, and Steven L. Waslander. Categorical depth distribution network for

monocular 3d object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8555–8564. Computer Vision Foundation / IEEE, 2021. 2

[33] Lennart Reiher, Bastian Lampe, and Lutz Eckstein. A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, page 1–7. IEEE Press, 2020. 2

[34] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 285. BMVA Press, 2019. 2

[35] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12786–12797. Curran Associates, Inc., 2021. 3

[36] Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. Translating images into maps. In *2022 International Conference on Robotics and Automation, ICRA 2022, Philadelphia, PA, USA, May 23-27, 2022*, pages 9200–9206. IEEE, 2022. 2

[37] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. June 2020. 1

[38] Enze Xie, Zhiding Yu, Daquan Zhou, Jonah Philion, Anima Anandkumar, Sanja Fidler, Ping Luo, and Jose M. Alvarez. M^2bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation, 2022. 1, 3

[39] Huanrui Yang, Hongxu Yin, Pavlo Molchanov, Hai Li, and Jan Kautz. Nvit: Vision transformer compression and parameter redistribution. *CoRR*, abs/2110.04869, 2021. 4

[40] Hongxu Yin, Arash Vahdat, Jose Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-ViT: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[41] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11784–11793. Computer Vision Foundation / IEEE, 2021. 1

[42] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2020. 1

[43] Jinnian Zhang, Houwen Peng, Kan Wu, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Minivit: Compressing vision transformers with weight multiplexing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 12135–12144. IEEE, 2022. 3

[44] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving, 2022. 1, 3

[45] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1