A Survey of Vision Transformers in Autonomous Driving: Current Trends and Future Directions

Quoc-Vinh Lai-Dang [1]

Abstract

This survey explores the adaptation of visual transformer models in Autonomous Driving, a transition inspired by their success in Natural Language Processing. Surpassing traditional Recurrent Neural Networks in tasks like sequential image processing and outperforming Convolutional Neural Networks in global context capture, as evidenced in complex scene recognition, Transformers are gaining traction in computer vision. These capabilities are crucial in Autonomous Driving for real-time, dynamic visual scene processing. Our survey provides a comprehensive overview of Vision Transformer applications in Autonomous Driving, focusing on foundational concepts such as self-attention, multi-head attention, and encoder-decoder architecture. We cover applications in object detection, segmentation, pedestrian detection, lane detection, and more, comparing their architectural merits and limitations. The survey concludes with future research directions, highlighting the growing role of Vision Transformers in Autonomous Driving.

Keywords: Autonomous Driving, Vision Transformers, Machine Learning

1. Introduction

Transformers [1] have revolutionized Natural Language Processing (NLP), with models like BERT, GPT, and T5 setting new standards in language understanding [2, 3, 4]. Their impact extends beyond NLP, as the Computer Vision (CV) community adopts Transformers for visual data processing. This shift from traditional Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to Transformers in CV signifies their growing influence, with early implementations in image recognition and object detection [5, 6, 7, 8] showing promising outcomes.

[1] Cho Chun Shik Graduate School of Mobility, Daejeon, Korea. Correspondence to: Quoc-Vinh Lai-Dang <ldqvinh@kaist.ac.kr>.

In Autonomous Driving (AD), Transformers are transforming a range of critical tasks, including object detection [9], lane detection [10], and segmentation [11, 12], and can be combined with reinforcement learning [13, 14] to execute complex path finding. They excel in processing spatial and temporal data, outperforming traditional CNNs and RNNs in complex functions like scene graph [15] generation and tracking [16]. The self-attention mechanism of Transformers provides a more comprehensive understanding of dynamic driving environments, essential for the safe navigation of autonomous vehicles.

This survey offers an extensive overview of Vision Transformers in AD, exploring their development, taxonomy, and varied applications. Starting with foundational aspects of Transformer architecture, the paper progresses to examine their roles in AD, highlighting improvements in 3D and 2D perception tasks. Concluding with future research directions, it emphasizes the potential in advancing AD, aiming to inspire further exploration and application in this field.

2. Exploring the Transformer: Structural and Functional Insights

2.1. The Transformer Architecture: A Structural Overview

The Transformer architecture, a groundbreaking innovation by [1], marks a departure from traditional recurrent layers by utilizing Attention mechanisms for sequence processing. It comprises two primary components: the Encoder and the Decoder. The Encoder processes input embeddings through Multi-head Attention and Feed-Forward Networks, both enhanced by Layer Normalization and residual connections. The Decoder, similar in structure to the Encoder, also focuses on the Encoder output, producing the final output sequence. Positional encodings are crucial in this architecture, as they imbue the model with the ability to recognize sequence order, a critical feature since Transformers do not inherently discern word order. This functionality is crucial for grasping language contexts, making positional encodings a fundamental component of the Transformer design. In the following, we describe each component of the Transformer in detail.
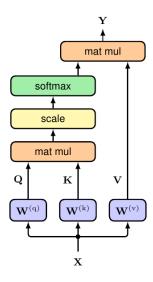
Figure 1. Self-attention process.

2.2. Self-Attention Mechanism: The Heart of Transformers

Central to the Transformer model is the Self-Attention mechanism (Figure 1), which assesses how various segments of the input sequence are related to one another. In this process, each input element is converted into three vectors: queries ($\mathbf{q}$), keys ($\mathbf{k}$), and values ($\mathbf{v}$), typically of dimension $d = 512$, and compiled into matrices Q, K, and V. The attention function then calculates interaction scores through a dot product between queries and keys, followed by normalization (dividing by $\sqrt{d}$) to stabilize training. These scores are converted into probabilities via a softmax function, indicating the degree of attention each element warrants. The final output (Y) is computed as:

$$Y = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right) V. \tag{1}$$

This is a weighted sum of the value vectors, encapsulating the context of the entire sequence. Additionally, the Encoder-Decoder Attention mechanism allows the decoder to concentrate on pertinent segments of the input sequence, informed by its present state and the output from the encoder. This mechanism, coupled with Positional encodings that add unique position information to input embeddings, ensures a comprehensive understanding of sequence ordering.

2.3. Multi-Head Attention: Enhancing Dimensional Analysis

The Multi-head Attention mechanism (Figure 2) enhances the ability to analyze various dimensions of the input data. Initially, the input vector is partitioned into three distinct sets for each head: the query set Q', the key set K', and the value set V', with each subset having a dimension of $\frac{d}{h}$.
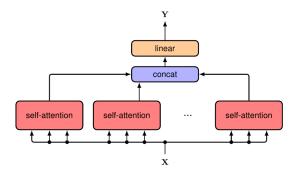

Figure 2. Multi-Head attention process.

These sets are composed of smaller vectors—specifically, h vectors per set, each with a dimension of 64 when d is 512. These vectors are then grouped to form the matrices Q', K', and V' for the subsequent attention calculations. The Multi-Head Attention process is formalized as follows:

$$\text{MultiHead}(Q', K', V') = \text{Concat}(head_1, \ldots, head_h)W^O, \tag{2}$$

where each $head_i$ is defined as Y. In this context, Q', K' and V' represent the collective matrices formed by the concatenation of their respective vectors, and W^O is a matrix of learned weights that combines the individual output of attention heads into a single output vector.

2.4. Other Core Mechanics of Transformer Models

The Feed-Forward Network (FFN) is a crucial component of Transformer models, positioned after Self-Attention computations in each unit. It consists of a two-stage linear operation with a nonlinear activation function, typically the Gaussian Error Linear Unit (GELU)[17]. This is mathematically represented as:

$$\text{FFN}(X) = W_2\sigma(W_1X), \tag{3}$$

where W_1 and W_2 are matrices of learnable parameters, and σ represents the nonlinear function. The role of FFN is to augment the capability to process complex data patterns, with its intermediate layer usually housing about 2048 units.

Skip connections are integral to each layer of Transformer models, enhancing information flow and addressing vanishing gradient issues. These connections add the input directly to the output of sub-layers:

$$\text{LayerNorm}(X + Z(X)), \tag{4}$$

where X is the input and $Z(X)$ is the output. Skip connections, combined with Layer Normalization [18], ensure stable learning. Some variants use Pre-Layer Normalization [19, 20, 21] for optimization, applying normalization before each sub-layer.

The output layer in Transformers is vital for translating vector sequences into interpretable outputs. It involves linearly mapping vectors to a logits space matching the vocabulary size, followed by a softmax function that converts logits to a probability distribution. This layer is key to transforming processed data into final, understandable results, crucial in various data processing tasks.

Transformers in Autonomous Driving function as advanced feature extractors, differing from CNNs by integrating information across larger visual fields for global scene understanding. Their capability to process data in parallel offers significant computational efficiency, essential for real-time processing in autonomous vehicles. The global perspective and efficiency make the Transformer highly advantageous for Autonomous Driving technology, enhancing system capabilities.

3. Vision Transformers in Autonomous Driving

Building on the foundational concepts of vanilla Transformers primarily used in NLP, this section ventures into the dynamic world of Vision Transformers (ViTs) and their impactful role in AD. ViTs have significantly evolved, showcasing their versatility and effectiveness in vehicular technologies. The upcoming subsections will detail how ViTs are employed across various dimensions of autonomous driving. We start by exploring their involvement in 3D tasks, including essential functions like object detection, tracking, and 3D segmentation, which are fundamental for comprehensive environmental perception. The narrative then transitions to 2D tasks, highlighting their capabilities in lane detection, sophisticated segmentation, and high-definition map creation — all crucial for interpreting two-dimensional spatial data. Finally, we delve into other pivotal roles of ViTs, such as trajectory and behavior prediction and their integration within end-to-end Autonomous Driving systems. This journey through the applications of ViTs in Autonomous Driving not only demonstrates their adaptability but also emphasizes their growing importance in enhancing the capabilities of autonomous vehicles.

3.1. The Rise of Vision Transformers

The ViT [5] (Figure 3) has brought a paradigm shift in image processing within Autonomous Driving, replacing conventional convolutional layers with Self-Attention layers. This transformative approach segments images into distinct patches for analysis using a Transformer encoder, comprised of Self-Attention and Feed-Forward layers. This enables focused analysis on essential image segments, substantially improving perception in driving scenarios. For larger images, ViT adopts a hybrid model, combining convolutional and Self-Attention layers. This innovative strategy is crucial for efficiently processing complex visual data, a key requirement for the sophisticated decision-making necessary in autonomous vehicles.

Advancing the concepts introduced by ViT, the Swin-Transformer [22] presents a novel hierarchical structure, specifically designed for image processing in Autonomous Driving systems. It effectively addresses the scalability challenges of Vision Transformers, primarily due to the high computational demands of Self-Attention mechanisms. The introduction of shifted windows in the Swin-Transformer facilitates efficient attention to adjacent patches without overlap, significantly reducing computational load and enabling the processing of larger images. Additionally, its unique tokenization method, which segments images into fixed-size patches and groups them hierarchically, maintains critical spatial information and captures both local and global scene contexts. The Swin-Transformer's proficiency in processing image features has led to its widespread use in various Autonomous Driving perception models, such as BEVFusion [23, 24] and BEVerse [25], highlighting its impact in advancing autonomous driving technology.

3.2. 3D Perception Tasks

The application of Vision Transformer models has led to significant progress in 3D and general perception tasks within autonomous driving. Initial models such as DETR [6] adopted an innovative method to object detection by framing it as a set prediction issue, employing pre-defined boxes and utilizing the Hungarian algorithm to predict sets of objects. This methodology was further refined in Deformable DETR ([7], which incorporated deformable attention for improved query clarity and faster convergence. DETR3D [26] extended these principles to 3D object detection, transforming LiDAR data into 3D voxel representations. Additionally, Vision Transformers like FUTR [27] and FUTR3D [28] have broadened their scope to include multimodal fusion, effectively processing inputs from various sensors to enhance the overall perception capabilities.

Vision Transformers have brought about significant innovations in 3D object detection, with models like PETR [29, 30], CrossDTR [31], BEVFormer [32, 33], and UVTR [34] leading the way. PETR notably uses position embedding transformations to enhance image features with 3D coordinate information, offering a more detailed spatial understanding. CrossDTR integrates the strengths of DETR3D and PETR to create a unified framework for detection that is informed by cross-view analysis and depth guidance. BEVFormer utilizes a spatio-temporal Vision Transformer architecture, achieving unified BEV representations by integrating spatial and temporal data seamlessly. UVTR, on the other hand, specializes in depth inference, employing cross-modal interactions to form distinct voxel spaces, thus enabling an
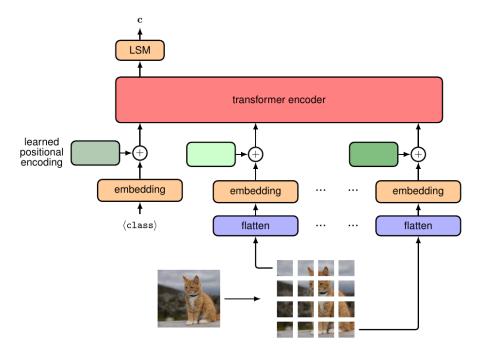
Figure 3. Vision Transformer Architecture.

extensive multimodal analysis crucial for accurate 3D object detection.

The field of 3D segmentation in autonomous driving has seen significant improvements with the integration of Vision Transformers. Models like TPVFormer [35], VoxFormer [36], and SurroundOcc [37] are notable examples. TPV-Former reduces the computational load by converting volumes into BEV planes, maintaining high accuracy in semantic occupancy predictions. VoxFormer uses 2D images to create 3D voxel query proposals, enhancing segmentation through deformable cross-attention queries. SurroundOcc utilizes a distinctive approach to extract 3D BEV features from 2D images of varying views and scales, merging these features proficiently to map out densely occupied spaces.

Vision Transformer models have brought about transformative changes in 3D object tracking for autonomous vehicles. Models like MOTR [38] and MUTR3D [39] have extended the capabilities of traditional tracking methods. MOTR, building on the DETR model, introduces a "track query" mechanism for modeling temporal variations across video sequences, avoiding reliance on conventional heuristics. MUTR3D introduces an innovative method that allows for concurrent detection and tracking. It employs associations across different cameras and frames to comprehend the three-dimensional state and appearance of objects over time, thereby greatly improving tracking accuracy and efficiency in autonomous driving systems.

3.3. 2D Perception Tasks

In Autonomous Driving, tasks related to 2D perception include crucial functions like detecting lanes, segmenting various elements, and creating high-definition maps. These tasks focus on processing and understanding two-dimensional spatial data, a critical aspect of autonomous vehicle technology. Unlike 3D tasks, which deal with depth and volume, 2D tasks require a precise interpretation of flat surfaces and plane elements, crucial for the accurate navigation and safety of autonomous vehicles.

Lane detection is a primary area where Transformer models have been effectively utilized, categorized into two distinct groups. The first group includes models like BEVSeg-Former [40], which employs cross-attention mechanisms for multi-view 2D image feature extraction and CNN-based semantic segmentation for accurate lane marking detection. Another example, PersFormer [41], combines CNNs for 2D lane detection with Transformers for enhancing BEV features. The second group, featuring models like LSTR [42] and CurveFormer [43], focuses on directly generating road structures from 2D images. These models use Transformer queries to refine road markings and implement curve queries for effective lane line generation, demonstrating the versatility and precision of Transformers in lane detection tasks.

Beyond lane detection, Transformer models are increasingly applied to segmentation tasks within autonomous driving. TIiM [44] exemplifies this application with its sequence-to-

4

sequence model that efficiently converts images and videos into overhead BEV maps, linking vertical scan lines in images to corresponding rays in maps for data-efficient and spatially aware processing. Panoptic SegFormer [45] provides an all-encompassing approach to panoptic segmentation, integrating both semantic and instance segmentation. Utilizing a supervised mask decoder and a strategy for query decoupling, it improves the segmentation efficiency. This model exemplifies the flexibility of Transformer architectures in handling intricate segmentation tasks.

In the realm of high-definition map generation, Transformer architectures like STSU [46], VectorMapNet [47], and MapTR [48] are bringing significant advancements. STSU treats lanes as directed graphs, focusing on learning Bezier control points and graph connectivity to convert front-view camera images into detailed BEV road structures. On the other hand, VectorMapNet leads the way in end-to-end vectorization of high-precision maps, utilizing sparse polyline primitives to model geometric shapes. MapTR offers an online framework for vectorized map generation, treating map elements as point sets and employing a hierarchical query embedding scheme. These models underscore the progress in merging multi-view features into a cohesive BEV perspective, crucial for creating accurate and detailed maps for autonomous driving.

3.4. Prediction, Planning and Decision-Making Tasks

Transformers are increasingly pivotal in autonomous driving, notably in prediction, planning, and decision-making. This progression marks a significant shift towards end-to-end deep neural network models that integrate the entire autonomous driving pipeline, encompassing perception, planning, and control into a unified system. This holistic approach reflects a substantial evolution from traditional models, indicating a move towards more comprehensive and integrated solutions in autonomous vehicle technology.

In trajectory and behavior prediction, Transformer-based models like VectorNet [49], TNT [50], DenseTNT [51], mm-Transformer [52], and AgentFormer [53] have addressed the limitations of standard CNN models, particularly in long-range interaction modeling and feature extraction. VectorNet enhances the depiction of spatial relationships by employing a hierarchical graph neural network, which is used for high-definition maps and agent trajectory representation. TNT and DenseTNT refine trajectory prediction, with DenseTNT introducing anchor-free prediction capabilities. The mmTransformer leverages a stacked architecture for simplified, multimodal motion prediction. AgentFormer uniquely allows direct inter-agent state influence over time, preserving crucial temporal and interactional information. WayFormer [54] further addresses the complexities of static and dynamic data processing with its innovative fusion

strategies, enhancing both efficiency and quality in data handling.

End-to-end models in autonomous driving have evolved significantly, particularly in planning and decision-making. TransFuser [55, 56] exemplifies this evolution with its use of multiple Transformer modules for comprehensive data processing and fusion. NEAT [57] introduces a novel mapping function for BEV coordinates, compressing 2D image features into streamlined representations. Building upon this, InterFuser [58] proposes a unified architecture for multimodal sensor data fusion, enhancing safety and decision-making accuracy. MMFN [59] expands the range of data types to include HD maps and radar, exploring diverse fusion techniques. STP3 [60] and UniAD [61] further contribute to this field, with STP3 focusing on temporal data integration and UniAD reorganizing tasks for more effective planning. These models collectively mark a significant stride towards integrated, efficient, and safer autonomous driving systems, demonstrating the transformative impact of Transformer technology in this domain.

4. Open Challenges and Future Directions

Vision Transformers offer promise for Autonomous Driving but face hurdles like data collection, safety, and interpretability. While they excel in perception and prediction, trends like multimodal fusion and explainability are emerging. Future focus areas include real-time processing optimization and end-to-end model development, necessitating continued research to overcome these challenges.

Challenges in Implementing Transformer Models. Transformers, evolving from their initial focus on 3D obstacle perception to a range of perception tasks in autonomous driving, encounter new challenges as they move towards integrating multimodal fusion. This transition necessitates multimodal models as well as efficient wireless connection [62, 63] to enhance efficiency gains, crucial for advanced autonomous driving. However, it also brings complexities in training and requires advancements in algorithms and system integration.

Hardware Acceleration and Model Complexity. As Transformer models grow in complexity, they demand innovative hardware acceleration solutions for efficient deployment. Integrating various models into an end-to-end system poses challenges in hardware optimization, especially with operators like Deformable Attention that complicate parallel processing. This necessitates specialized hardware designs to meet the diverse requirements of these advanced models.

Future Directions: Algorithm and Hardware Advancements. Future advancements in Transformer models for autonomous driving will focus on algorithm enhancements and hardware innovations. Key areas include mixed-precision

quantization for balancing computational demands and interpretability techniques like attention-based saliency maps. These developments aim to improve model compression and offer insights into decision-making processes, building trust in autonomous systems.

Enhancing Model Efficiency and Interpretability. Efficiency and interpretability are pivotal for the future of Transformer models in autonomous driving. The need for models that process multi-view data effectively and offer improved generalization while being optimized for performance is critical. Developing interpretable models with techniques to visually highlight crucial data will enhance system reliability and user trust in autonomous driving technology.

5. Conclusion

This paper has offered a comprehensive survey of Transformer models, especially Vision Transformers, in autonomous driving (AD), demonstrating their significance beyond traditional Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). We explored their foundational architecture, attention-based processing advantages in natural language processing and computer vision, and their superior performance in various AD tasks, including 3D object detection, 2D lane detection, and advanced scene analysis. Additionally, we highlighted challenges, trends, and future perspectives of Vision Transformers in AD, aiming to spur further interest and research in this dynamic field. The potential of Vision Transformers in transforming AD, with their nuanced data processing capabilities, promises exciting advancements in vehicular technologies.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] S. Alaparthi and M. Mishra, "Bidirectional encoder representations from transformers (bert): A sentiment analysis odyssey," *arXiv preprint arXiv:2007.01127*, 2020.

[3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI*, 2018.

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[7] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[8] T. Kim, L. F. Vecchietti, K. Choi, S. Lee, and D. Har, "Machine learning for advanced wireless sensor networks: A review," *IEEE Sensors Journal*, vol. 21, no. 11, pp. 12379–12397, 2020.

[9] J. Mao, S. Shi, X. Wang, and H. Li, "3d object detection for autonomous driving: A comprehensive survey," *International Journal of Computer Vision*, pp. 1–55, 2023.

[10] J. Han, X. Deng, X. Cai, Z. Yang, H. Xu, C. Xu, and X. Liang, "Laneformer: Object-aware row-column transformers for lane detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 799–807, 2022.

[11] A. Ando, S. Gidaris, A. Bursuc, G. Puy, A. Boulch, and R. Marlet, "Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5240–5250, 2023.

[12] S. Cakir, M. Gauss, K. Happeler, Y. Ounajjar, F. Heinle, and R. Marchthaler, "Semantic segmentation for autonomous driving: Model evaluation, dataset generation, perspective comparison, and real-time capability," *arXiv preprint arXiv:2207.12939*, 2022.

[13] M. Seo, L. F. Vecchietti, S. Lee, and D. Har, "Rewards prediction-based credit assignment for reinforcement learning with sparse binary rewards," *IEEE Access*, vol. 7, pp. 118776–118791, 2019.

[14] L. F. Vecchietti, M. Seo, and D. Har, "Sampling rate decay in hindsight experience replay for robot control," *IEEE Transactions on Cybernetics*, vol. 52, no. 3, pp. 1515–1526, 2020.

[15] H. Liu, Z. Huang, X. Mo, and C. Lv, "Augmenting reinforcement learning with transformer-based scene representation learning for decision-making of autonomous driving," *arXiv preprint arXiv:2208.12263*, 2022.

[16] Y. Zhang, S. Zhang, D. Xin, D. Chen, *et al.*, "A small target pedestrian detection model based on autonomous driving," *Journal of Advanced Transportation*, vol. 2023, 2023.

[17] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[18] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[19] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the transformer architecture," in *International Conference on Machine Learning*, pp. 10524–10533, PMLR, 2020.

[20] S. Takase, S. Kiyono, S. Kobayashi, and J. Suzuki, "On layer normalizations and residual connections in transformers," *arXiv preprint arXiv:2206.00330*, 2022.

[21] S. Shleifer, J. Weston, and M. Ott, "Normformer: Improved transformer pretraining with extra normalization," *arXiv preprint arXiv:2110.09456*, 2021.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[23] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "Bevfusion: A simple and robust lidar-camera fusion framework," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10421–10434, 2022.

[24] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 2774–2781, IEEE, 2023.

[25] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu, "Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving," *arXiv preprint arXiv:2205.09743*, 2022.

[26] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on Robot Learning*, pp. 180–191, PMLR, 2022.

[27] D. Gong, J. Lee, M. Kim, S. J. Ha, and M. Cho, "Future transformer for long-term action anticipation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3052–3061, 2022.

[28] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 172–181, 2023.

[29] Y. Liu, T. Wang, X. Zhang, and J. Sun, "Petr: Position embedding transformation for multi-view 3d object detection," in *European Conference on Computer Vision*, pp. 531–548, Springer, 2022.

[30] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, and X. Zhang, "Petrv2: A unified framework for 3d perception from multi-camera images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3262–3272, 2023.

[31] C.-Y. Tseng, Y.-R. Chen, H.-Y. Lee, T.-H. Wu, W.-C. Chen, and W. H. Hsu, "Crossdtr: Cross-view and depth-guided transformers for 3d object detection," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4850–4857, IEEE, 2023.

[32] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*, pp. 1–18, Springer, 2022.

[33] C. Yang, Y. Chen, H. Tian, C. Tao, X. Zhu, Z. Zhang, G. Huang, H. Li, Y. Qiao, L. 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*, pp. 17830–17839, 2023.

[34] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18442–18455, 2022.

[35] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9223–9232, 2023.

[36] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "Voxformer: Sparse voxel transformer for camera-based 3d semantic scene

completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9087–9098, 2023.

[37] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21729–21740, 2023.

[38] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "Motr: End-to-end multiple-object tracking with transformer," in *European Conference on Computer Vision*, pp. 659–675, Springer, 2022.

[39] T. Zhang, X. Chen, Y. Wang, Y. Wang, and H. Zhao, "Mutr3d: A multi-camera tracking framework via 3d-to-2d queries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4537–4546, 2022.

[40] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5935–5943, 2023.

[41] L. Chen, C. Sima, Y. Li, Z. Zheng, J. Xu, X. Geng, H. Li, C. He, J. Shi, Y. Qiao, *et al.*, "Persformer: 3d lane detection via perspective transformer and the openlane benchmark," in *European Conference on Computer Vision*, pp. 550–567, Springer, 2022.

[42] R. Liu, Z. Yuan, T. Liu, and Z. Xiong, "End-to-end lane shape prediction with transformers," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3694–3702, 2021.

[43] Y. Bai, Z. Chen, Z. Fu, L. Peng, P. Liang, and E. Cheng, "Curveformer: 3d lane detection by curve propagation with curve queries and attention," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7062–7068, IEEE, 2023.

[44] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Translating images into maps," in *2022 International conference on robotics and automation (ICRA)*, pp. 9200–9206, IEEE, 2022.

[45] Z. Li, W. Wang, E. Xie, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, and T. Lu, "Panoptic segformer: Delving deeper into panoptic segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1280–1289, 2022.

[46] Y. B. Can, A. Liniger, D. P. Paudel, and L. Van Gool, "Structured bird's-eye-view traffic scene understanding from onboard images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15661–15670, 2021.

[47] Y. Liu, T. Yuan, Y. Wang, Y. Wang, and H. Zhao, "Vectormapnet: End-to-end vectorized hd map learning," in *International Conference on Machine Learning*, pp. 22352–22369, PMLR, 2023.

[48] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, and C. Huang, "Maptr: Structured modeling and learning for online vectorized hd map construction," *arXiv preprint arXiv:2208.14437*, 2022.

[49] J. Gao, C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, and C. Schmid, "Vectornet: Encoding hd maps and agent dynamics from vectorized representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11525–11533, 2020.

[50] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid, *et al.*, "Tnt: Target-driven trajectory prediction," in *Conference on Robot Learning*, pp. 895–904, PMLR, 2021.

[51] J. Gu, C. Sun, and H. Zhao, "Densetnt: End-to-end trajectory prediction from dense goal sets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15303–15312, 2021.

[52] Y. Liu, J. Zhang, L. Fang, Q. Jiang, and B. Zhou, "Multimodal motion prediction with stacked transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7577–7586, 2021.

[53] Y. Yuan, X. Weng, Y. Ou, and K. M. Kitani, "Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9813–9823, 2021.

[54] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion forecasting via simple & efficient attention networks," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2980–2987, IEEE, 2023.

[55] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[56] Q.-V. Lai-Dang, J. Lee, B. Park, and D. Har, "Sensor fusion by spatial encoding for autonomous driving," *arXiv preprint arXiv:2308.10707*, 2023.

[57] K. Chitta, A. Prakash, and A. Geiger, "Neat: Neural attention fields for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15793–15803, 2021.

[58] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," in *Conference on Robot Learning*, pp. 726–737, PMLR, 2023.

[59] Q. Zhang, M. Tang, R. Geng, F. Chen, R. Xin, and L. Wang, "Mmfn: Multi-modal-fusion-net for end-to-end driving," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8638–8643, IEEE, 2022.

[60] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "Stp3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *European Conference on Computer Vision*, pp. 533–549, Springer, 2022.

[61] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17853–17862, 2023.

[62] I. Park, D. Kim, and D. Har, "Mac achieving low latency and energy efficiency in hierarchical m2m networks with clustered nodes," *IEEE Sensors Journal*, vol. 15, no. 3, pp. 1657–1661, 2015.

[63] J. Park, E. Hong, and D. Har, "Low complexity data decoding for slm-based ofdm systems without side information," *IEEE Communications Letters*, vol. 15, no. 6, pp. 611–613, 2011.