TrackOcc: Camera-based 4D Panoptic Occupancy Tracking

Zhuoguang Chen[1,3], Kenan Li[1], Xiuyu Yang[1], Tao Jiang[1], Yiming Li[2] and Hang Zhao[1,3,†]

Abstract—Comprehensive and consistent dynamic scene understanding from camera input is essential for advanced autonomous systems. Traditional camera-based perception tasks like 3D object tracking and semantic occupancy prediction lack either spatial comprehensiveness or temporal consistency. In this work, we introduce a brand-new task, Camera-based 4D Panoptic Occupancy Tracking, which simultaneously addresses panoptic occupancy segmentation and object tracking from camera-only input. Furthermore, we propose TrackOcc, a cutting-edge approach that processes image inputs in a streaming, end-to-end manner with 4D panoptic queries to address the proposed task. Leveraging the localization-aware loss, TrackOcc enhances the accuracy of 4D panoptic occupancy tracking without bells and whistles. Experimental results demonstrate that our method achieves state-of-the-art performance on the Waymo dataset. The code will be released for future research.

I. INTRODUCTION

A holistic and precise understanding of dynamic environments is essential for perception systems in robotics and autonomous vehicles. A robust perception system needs to estimate the geometry, semantics, and identities of the current scene in a spatial-continuous and temporal-consistent way to interact with complex, changing 3D surroundings.

Previous efforts to achieve this goal include 3D object tracking [1], [2], semantic occupancy prediction [3], [4], [5], [6], and 4D LiDAR segmentation [7], [8], [9]. As depicted in Fig. 1, 3D object tracking (a) typically focuses on tracking objects across frames using bounding boxes. However, bounding boxes neglect fine-grained geometric details and struggle to represent general objects. On the other hand, occupancy-based tasks (b) offer a more complete representation of the 3D scene by incorporating both geometry and semantic information. Despite this advantage, existing approaches to occupancy prediction are mostly limited to spatial semantic perception, lacking a broader focus on panoramic temporal understanding. 4D Panoptic LiDAR Segmentation (4DPLS) [7] tackles both semantic and instance segmentation in a 3D space over time, but it is constrained by the high cost of LiDARs and the limited sparsity of point clouds.

Different from previous efforts, we propose a brand-new task: Camera-based 4D Panoptic Occupancy Tracking, as shown in Fig.1 (d). Unlike 4DPLS, which poses restricted semantic details, our task utilizes camera-only input, offering a cost-effective and easy-to-deploy solution that enhances spatiotemporal urban scene understanding.

*Zhuoguang Chen and Kenan Li contribute equally.

[†]Corresponding author

[1]Tsinghua University, [2]New York University, [3]Shanghai AI Lab
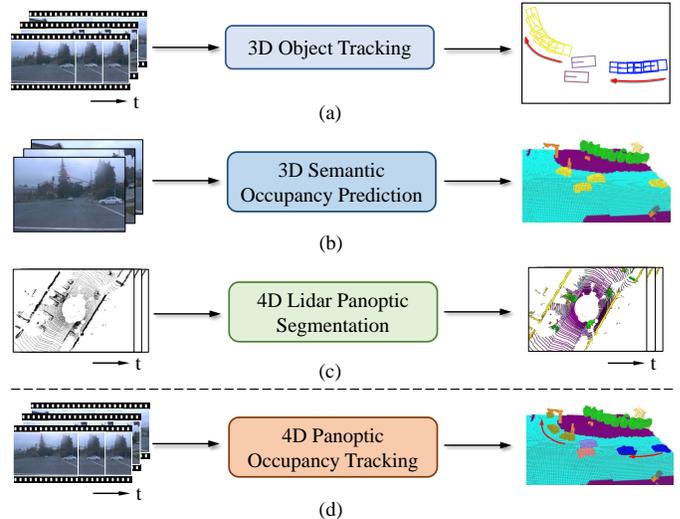
Corresponding at: hangzhao@mail.tsinghua.edu.cn

Fig. 1: **Comparison of tasks for scene understanding.** (a) Outputs bounding box tracks from multi-view image sequences. (b) Predicts semantic labels for the occupancy volume from multi-view images. (c) Applies 3D panoptic segmentation on aggregated 4D LiDAR point volumes, followed by post-processing instance matching. (d) Our proposed task: Predicts temporally consistent panoptic labels of the occupancy from multi-view image sequences.

Furthermore, as shown in Fig.1 (c), previous methods for 4D LiDAR panoptic segmentation all conduct 4D LiDAR panoptic segmentation in a (3+1)D way, i.e., segment aggregated 3D volumes, and then post-process them to link the volumes [7]. Instead, we introduce **TrackOcc**, an end-to-end learning-based tracker approach for Camera-based 4D Panoptic Occupancy Tracking.

TrackOcc processes image inputs in a streaming, end-to-end manner, eliminating the need for extensive post-processing. Specifically, we introduce 4D panoptic queries into TrackOcc, enabling the prediction of occupancy with time-consistent panoptic labels in a unified framework. To ensure that the 4D panoptic queries capture spatially-accurate 3D features, we employ localization-aware loss to guide TrackOcc towards the targeted areas, which significantly enhances the model's overall performance.

To summarize, our main contributions are three-fold:

- To the best of our knowledge, we make the first attempt to explore a camera-based 4D panoptic occupancy tracking task, which jointly tackles occupancy panoptic segmentation and object tracking with camera input.
- We propose TrackOcc, which uses *4D panoptic queries* to perform the proposed task in a streaming, end-to-end

manner. We also introduce a localization-aware loss to enhance the tracking performance.

- For fair evaluations, we propose the OccSTQ metric and build a set of baselines adapted from other domains. Experiments demonstrate that our TrackOcc achieves state-of-the-art performance on Waymo dataset.

II. RELATED WORK

A. 3D Occupancy Prediction

Occupancy representation, compared with 3D bounding boxes, offers finer geometric details and can assist in handling general, out-of-vocabulary objects [10]. The task requires to jointly estimate the occupancy state and semantic label of every voxel in the scene from images. Among the explorations, Occ3D [10] offers a 3D occupancy dataset, followed by multiple solutions [11], [3], [6]. Several recent works attempt to extend the task to panoptic prediction. SparseOcc [6] utilized the inherent sparsity property of the scene to reduce the computational cost while offering an extension to panoptic occupancy prediction. PaSCo [12] extended the Semantic Scene Completion task to the panoptic segmentation. However, they cannot guarantee temporally consistent panoptic label prediction, specifically in tracking objects. Achieving this is a critical milestone for robots and autonomous vehicles, as it enables them to perceive and plan their paths effectively in the physical world.

B. Video Panoptic Segmentation

Video Panoptic Segmentation (VPS) is proposed in [13] to unify video semantic and video instance segmentation. In the beginning, numerous methods [13], [14] like VPSNet, depend on multiple sub-networks and complex post-processing (e.g. NMS, thing-stuff fusion). To alleviate these shortcomings, Recent research efforts for end-to-end video panoptic segmentation, such as [15], [16], [17], [18] are extended from image panoptic segmentation method, incorporating specific trackers for this purpose. However, Since these methods rely on feeding video clips into the model, they can only handle short-term associations. Furthermore, these VPS methods are mainly applied to the 2D video domain and cannot be directly applied to the 3D world domain.

C. 3D Object Tracking

3D Multi-object tracking (3D MOT) involves using multi-view images from cameras or point clouds from LiDAR to track multiple objects across frames. Building upon advances in 3D object detection [19], [20], [21], [22], recent 3D trackers usually adopt the tracking-by-detection paradigm, associating trajectories with detected boxes through post-processing. However, this paradigm is heavily dependent on complex human designs and post-processing steps. To overcome these limitations, MUTR3D [23] has successfully extended the MOTR-type tracker to the 3D domain, achieving end-to-end 3D tracking with promising results. ADA-Track [24] and DQTrack [1] utilize query-based frameworks and learnable associations to detect and track objects in multi-view images. However, these methods are based on

bounding boxes rather than occupancy, neglecting the finer geometric details and general objects. And they only focus on tracking *thing* classes, ignoring recognizing *stuff* classes, where we unify them in one framework.

D. 4D Panoptic LiDAR Segmentation

The spatiotemporal interpretation in LiDAR point-based tasks has been extensively explored [9], [25], [26], [8]. Since the task was first defined [7]. 4D-PLS [7] initially proposed applying 3D panoptic segmentation to a frame of multi-scan-aggregated point clouds, aiming to reduce the need for explicit data association. Subsequently, methods such as Mask4Former [8] have emerged, focusing on spatially compact instance segmentation with an additional 6-DOF bounding box regression. However, these approaches still rely on 3D segmentation, which cannot eliminate the need for offline post-association. As a result, it does not fully address the 4D panoptic segmentation problem and is not suitable for real-time streaming applications, limiting its use in mobile robots and autonomous driving systems.

III. METHOD

A. Problem Definition

In this paper, we propose the **Camera-based 4D Panoptic Occupancy Tracking** task, which jointly addresses occupancy panoptic segmentation and object tracking in both spatial and temporal domains. Specifically, the model takes as input a sequence of RGB images, denoted by denoted by $\mathbf{I} = \{\mathbf{I}_t^i, \mathbf{I}_{t-1}^i, \ldots, \mathbf{I}_{t-L}^i\}_{i=1}^N$, where N represents the number of surrounding cameras and L denotes the sequence length. Given a predetermined set of C semantic classes, encoded as $\mathcal{C} := \{0, \ldots, C-1\}$, the task requires a neural network to map each voxel i in the grid to a pair $(c_i, z_i) \in \mathcal{C} \times \mathbb{N}$, where c_i represents the semantic class of voxel i and z_i denotes its instance ID. The semantic label set comprises both *stuff* and *thing* categories. When a voxel is labeled as *stuff*, its corresponding instance ID z_i is irrelevant. Voxels in the free space are assigned a special *free* stuff label. The instance IDs z_i group voxels of the same class into distinct segments, which should persist throughout the entire sequence. The grid dimensions are $X \times Y \times Z$, representing the height, width, and depth of the grid, respectively.

B. Overall Architecture

The overall proposed TrackOcc architecture is depicted in Fig. 2 and comprises three main components. First, an encoder (Sec. III-C) extracts volume features from multi-view images. Second, a novel query design facilitates both panoptic segmentation and instance tracking. Finally, a decoder (Sec. III-E) interacts with the volume features and queries to generate the final 4D prediction in an end-to-end, streaming process.

C. Encoder

The encoder constructs the 3D volume features from multi-view images by integrating an image feature extractor with a 2D-to-3D transformation module. The image feature
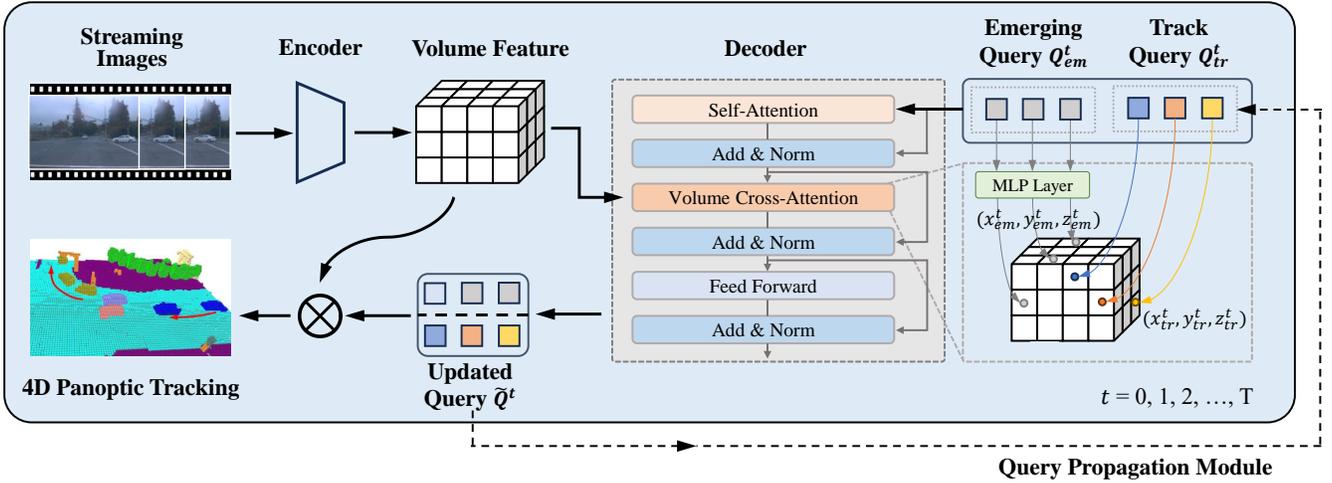
Fig. 2: **Overall Pipeline of TrackOcc.** At each timestep, the encoder extracts multi-view image features and transforms them into 3D volume features. These volume features serve as critical context for updating the 4D panoptic queries via a designed decoder. The query propagation module facilitates efficient object tracking in a streaming, end-to-end manner. The symbol $\otimes$ denotes matrix multiplication.

extractor captures multi-scale features from multi-view images, which are then transformed into 3D volume representations $\mathbf{F} \in \mathbb{R}^{X \times Y \times Z \times D}$, where D denotes the feature dimension per voxel. Extensive research has been dedicated to methods for transforming 2D image space into 3D volume space, including LSS [27], BevDepth [28], BEVFormer [29], FB-Occ, and COTR [3]. In our experiments, we utilize COTR's encoder [3] as our 2D-to-3D transformation module. Importantly, our TrackOcc is flexible and compatible with all of these approaches.

D. 4D Panoptic Query

When designing queries for 4D Panoptic Occupancy Tracking, two main challenges arise:1) how to define queries for panoptic segmentation and instance tracking. 2) how to manage newborn and terminated instances.

Query Definition. We define two types of queries: emerging queries Q_{em} and track queries Q_{tr}. Emerging queries are utilized for both *stuff* and detecting newborn instances, while track queries are responsible for predicting all existing tracked instances. For simplicity, *stuff* is assigned to emerging queries since it does not require tracking. The set of track queries is updated dynamically, and its size varies over time. In the first frame, there are no track queries; only the fixed-length, learnable emerging queries are fed into the decoder. For subsequent frames, we input Q^t, which is the concatenation of track queries Q_{tr}^t and the learnable emerging queries Q_{em}^t into the decoder. These queries interact with the volume features in the decoder to generate updated queries $\tilde{Q}^t = \tilde{Q}_{em}^t \cup \tilde{Q}_{tr}^t$ which are used to produce the final 4D predictions. The updated queries $\tilde{Q}^t$ are also passed to the Query Propagation Module to generate the track queries Q_{tr}^{t+1} for the next frame.

Query Propagation. To handle instances that may appear or disappear in intermediate frames, we introduce the Query Propagation Module. This module manages the propagation of queries for newborn and terminated instances within our method. During training, for queries in $\tilde{Q}_{em}^t$, only those queries matched to ground truth *things* are converted into track queries. Queries corresponding to ground truth *stuff* are not converted into track queries. For queries in $\tilde{Q}_{tr}^t$ are removed if they have been matched to $\varnothing$ for T_f successive frames.

During inference, we use the predicted classification scores to determine appearance of newborn instances and disappearance of tracked instances. For queries in $\tilde{Q}_{em}^t$, predictions with classification scores higher than the entrance threshold τ_1 and corresponding to *things* are retained. Queries corresponding to *stuff* are discarded even if their classification scores exceed τ_1. For queries in $\tilde{Q}_{tr}^t$, predictions with classification scores lower than the exit threshold τ_2 for consecutive T_f frames are removed.

E. Decoder

We detail the decoding process as follows. First, the queries interact with each other to enhance features using a self-attention [30] layer. Then, they look up and aggregate volume features via a Volume Cross-Attention (VCA) layer. Due to the large input scale of the volume features, the computational cost of vanilla attention [30] is high. Therefore, we adopt a VCA layer based on deformable attention [31], a resource-efficient attention mechanism where each query interacts only with its regions of interest in the volume. Specifically, each emerging query $q \in Q_{em}$ predicts a 3D reference point $p \in \mathbb{R}^3$ using an MLP layer while each track query $q \in Q_{tr}$ maintains its corresponding reference point over time, and the volume features F around these reference points are sampled. A weighted sum of the sampled features is then performed as the output. The process of volume cross-

attention (VCA) can be formulated as:

$$\text{VCA}(q, F) = \sum_{j=1}^{J} w_j \cdot F(p + \Delta p_j), \qquad (1)$$

where j indexes the sample point, and $w_j \in \mathbb{R}^1$ and $\Delta p_j \in \mathbb{R}^3$ are learnable weights and offsets.

Finally, a linear classifier, followed by a sigmoid activation function, is applied on top of the query embeddings to yield class probability predictions. Notably, the classifier predicts an additional "no object" category $\varnothing$ in cases where the query does not correspond to any region. For mask prediction, we obtain each binary volume mask prediction $m_i \in [0,1]^{X \times Y \times Z}$ via a matrix multiplication between each query and the volume feature, followed by a sigmoid activation. The binary volume mask, combined with the class prediction, generates the panoptic occupancy prediction. For *things*, a temporally consistent instance ID is further assigned for the 4D prediction.

F. Loss function

The optimization objectives of our approach focus on two main aspects: mask prediction and position prediction of all queries. We calculate the mask classification loss $\mathcal{L}_{mask-cls}$ for masks, and different types of queries are matched with the ground truth using different strategies. Track queries persist throughout the entire occupancy flow, which allows us to assign the ground truth for each of them only once across all timesteps. Once created by inheriting from emerging queries, track queries are tied to the corresponding ground truth and remain unchanged.

Emerging queries, which transition between different time frames, do not require explicit assignment at each timestep. Following Mask2Former [32], we establish a correspondence between the ground truth and the prediction via bipartite matching, solved by the Hungarian algorithm [33]. Once the matching is established, we compute the mask classification loss for queries, denoted as $\mathcal{L}_{mask-cls}$, which includes the binary mask loss $\mathcal{L}_{mask}$ and the multi-class cross-entropy loss $\mathcal{L}_{cls}$.

Localization-aware Loss. Localizing and tracking each object accurately is both crucial and challenging. Previous works that rely on mask classification for 3D panoptic prediction, such as SparseOcc [6], have limited localization capabilities. Thus, we propose the localization-aware loss:

$$\mathcal{L}_{loc} = \text{L1}(P', P), \qquad (2)$$

to improve the accuracy of query-specific 3D positions, since they are used in our VCA module to serve as the reference points for the attention. Note that the predicted points P' only come from the queries that are tied to *thing* classes due to the lack of the proper and consistent ground truth points of *stuff* regions. We can obtain P by easily calculate the centroid of each object. We observe that this enhancement significantly improves the metrics, as demonstrated in our experiments(Sec. IV).

Overall, the final training loss is:

$$\mathcal{L} = \mathcal{L}_{mask-cls} + \mathcal{L}_{loc}. \qquad (3)$$

IV. EXPERIMENTS

A. Dataset

We conduct our experiments on the Occ3D-Waymo [10] dataset, which provides dense semantic labels for 3D occupancy grids. This dataset includes 798 training scenes and 202 validation scenes, with a spatial range of $[-40m, 40m]$ for the x and y axes, and $[-1m, 5.4m]$ for the z axis. The voxel grid size is $(0.4m, 0.4m, 0.4m)$, resulting in a resolution of $(200 \times 200 \times 16)$ for (X, Y, Z). However, the dataset lacks panoptic labels, offering only semantic labels.

To generate 4D panoptic labels for our proposed 4D Panoptic Occupancy tracking task, we utilize the annotations from the Waymo Open Dataset [35]. These annotations provide exhaustive 3D bounding boxes for vehicles, pedestrians, and cyclists, each with a unique tracking ID. We classify voxels corresponding to vehicles, pedestrians, and cyclists as *thing* while all other categories are labeled as *stuff*. Furthermore, *thing* voxels are divided into instances, with voxels within a single 3D box grouped as one instance. For *thing* voxels not located inside any 3D box, we assign them to the nearest 3D box. Finally, each voxel is assigned the tracking ID from its corresponding 3D box.

B. Implement Details

We adopt ResNet-50 [36] as the image backbone with an input resolution of 256 × 704. Both Self-attention and volume cross-attention is single-layer. The number of sampling points per reference point, J, is set to 4 for VCA in the Decoder. We put 200 emerging queries into each frame and the query dimension is set to 256. The query lifecycle is controlled by $\tau_1 = 0.3$, $\tau_2 = 0.25$, and $T_f = 3$. All our experiments are conducted after training for 24 epochs with a learning rate of 2×10^{-4}, using the AdamW [37] optimizer and a batch size of 1, on 8 NVIDIA GeForce RTX 3090 GPUs.

C. Evaluation Metrics

To evaluate the methods for Camera-based 4D Panoptic Occupancy Tracking, and showcase the segmentation and tracking ability simultaneously, we propose the Occupancy

TABLE I: **Performance Comparison Between SparseOcc and COTR on 3D Panoptic Prediction.** Under a fair revision metric PQ*, COTR outperforms SparseOcc in both small objects and stuff regions. C.C represents the construction cone.

Methods	Metrics	Overall	vehicle	pole	C.C.	road
SparseOcc [6]	PQ	5.0	12.4	**0.1**	**2.6**	13.6
	PQ*	12.8	21.6	*7.0*	*13.8*	34.4
COTR [3]	PQ	13.6	13.7	**0.5**	**0.6**	76.9
	PQ*	22.4	24.3	*15.0*	*12.0*	78.2

TABLE II: **Camera-based 4D Panoptic Occupancy Tracking Performance on the Occ3D-Waymo dataset [10].** MinVIS and CTVIS are query-based association methods, while AB3DMOT refers to the tracking-by-detection method. 4D-LCA denotes 4D LiDAR cross-volume association method in [7]. E2E represents end-to-end.

Method	E2E	OccSTQ	OccSQ	OccAQ			
				Overall	Vehicle	Pedestrian	Cyclist
MinVIS [15]	✓	9.8	29.1	3.3	4.0	1.5	3.1
CTVIS [17]	✓	10.7	26.5	4.3	5.3	2.1	1.6
AB3DMOT [34]	✗	12.4	29.1	5.3	6.8	1.7	2.9
4D-LCA [7]	✗	16.2	29.1	9.0	11.4	**3.4**	4.1
TrackOcc (ours)	✓	**20.0**	**29.4**	**13.5**	**18.1**	3.1	**4.6**

TABLE III: **Ablation Studies.** Comparison of results with different settings including Localization-aware Loss and the number of training frames.

w/ Loc. Loss	#frames	OccSTQ	OccSQ	OccAQ
✗	3	16.5	28.9	9.4
✓	3	20.0	29.4	13.5
✓	5	**20.6**	**29.5**	**14.4**

Segmentation and Tracking Quality (OccSTQ) metric. Inspired by [38], [39], [7], the OccSTQ is derived from the STQ [39], which is designed for video panoptic segmentation, effectively disentangles association and classification errors, and is well-suited for long sequences. Therefore, we adapt it as OccSTQ for our task.

We define a sequence of occupancy frames as $\Omega = \{(\mathbf{v}, t) \mid \mathbf{v} \in X \times Y \times Z, t \in [0, T)\}$, where $\mathbf{v}$ represents a voxel in a 3D spatial domain, and t denotes the timestep over a total duration. The ground truth is represented as $gt(\mathbf{v}, t)$, and the prediction as $pr(\mathbf{v}, t)$. OccSTQ consists of Occupancy Segmentation Quality (OccSQ) and Occupancy Association Quality (OccAQ), which measure segmentation and association quality respectively.

OccSQ is defined as mean IoU (Intersection-over-Union), similar to [39] and [7], but with voxels as the basic elements.

For OccAQ, we define the prediction for a specific identity in *thing* class as: $pr_{id}(id) = \{(\mathbf{v}, t) \mid pr(\mathbf{v}, t) = (c, id), c \in C_{th}\}$, and the ground-truth is defined analogously as $gt_{id}(id)$. Thus the true positive associations (TPA) is: $TPA(p, g) = |pr_{id}(id') \cap gt_{id}(id)|$. It is important to note that the predicted id' and the ground-truth id can belong to different *thing* classes, as incorrect semantic classification is penalized only in OccSQ. Similarly, false positive associations (FPA) and false negative associations (FNA) can be defined. Then, the OccAQ can be summarized as:

$$OccAQ = \frac{1}{|G|} \sum_{g \in G} \frac{1}{|gt_{id}(g)|} \sum_{\substack{p \in P \\ p \cap g \neq 0}} TPA(p, g) \cdot \text{IoU}(p, g). \quad (4)$$

Finally, the OccSTQ is defined as the geometric mean of OccSQ and OccAQ:

$$OccSTQ = \sqrt{OccSQ \times OccAQ}. \quad (5)$$

According to the definition of Occ3D [10], we only calculate the metric within the visible region. Moreover, OccSTQ can be extended to evaluate scene completion capability without visible region constriction.

D. Baselines

Given the novelty of our proposed task, existing works do not fully align with our objectives. To facilitate a fair comparison, we constructed a series of baselines and demonstrated our method's superior performance. Specifically, we first adopt COTR [3] as our baseline foundational 3D panoptic occupancy prediction model, since COTR shows better performance than SparseOcc [40] in Tab. I. Then we adapt it with mainstream temporal association techniques to achieve 4D panoptic occupancy tracking.

To select the proper 3D panoptic occupancy prediction model, we test the MaskFormer-based methods, COTR [3] and SparseOcc [6], using the panoptic quality (PQ) metric [41] and PQ*, as shown in Tab. I. Since PQ's matching requirement, IoU > 0.5, is too strict for camera-based methods, resulting in the difficulty of comparing performance on small objects, we transform the matching problem in PQ to maximum weighted bipartite matching [42] and denote the metric as PQ*, intend to achieve a fairer comparison.

After the selection, we extend COTR [3] to the temporal domain with adapted temporal association methods from related fields for fair comparison.

- From video instance segmentation, we directly adopt the bipartite matching approach for cross-frame query association of MinVIS [15];
- We use the contrastive training strategy between cross-frame query from CTVIS [17] to enhance the query association ability;
- To associate instances across frames, an intuitive method is to convert predicted voxels with the same instance ID into bounding boxes and track them using a bounding box tracker, thus, AB3DMOT [34] is adapted;
- Drawing from the two-stage association approach of 4D LiDAR panoptic segmentation [7], the 4D LiDAR

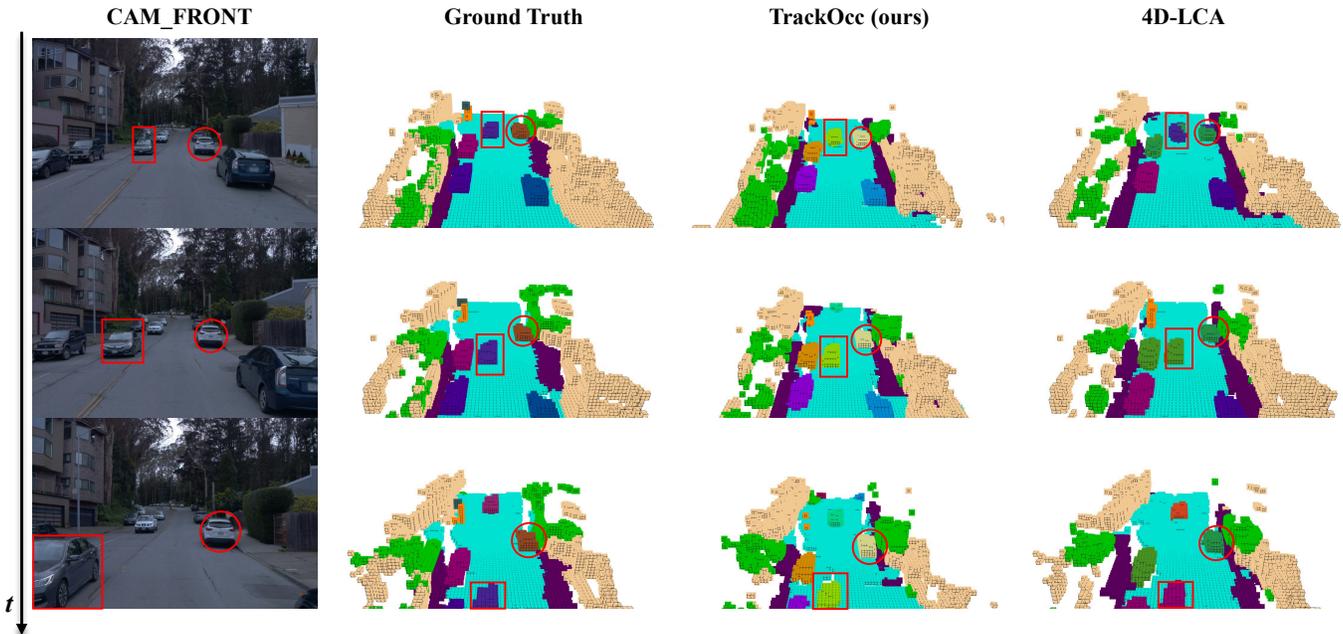| CAM_FRONT | Ground Truth | TrackOcc (ours) | 4D-LCA |

Fig. 3: **Qualitative results of our method and 4D-LCA.** The same color indicates identical instances across different time steps. Our method effectively tracks both moving and stationary vehicles, while the 4D-LCA method encounters difficulties with moving vehicles. For clarity, only images from the front camera are displayed, although both methods process data from surrounding cameras.

cross-volume association (4D-LCA) is incorporated as a baseline.

E. Main results

As shown in Table II, we present a quantitative comparison between our method and the baselines. TrackOcc outperforms all competitors, particularly in tracking quality. Compared to 4D-LCA [7], our method achieves a significant improvement of 3.8 OccSTQ and a 50% margin in OccAQ, highlighting its strong tracking ability.

Several insights can be drawn from the results. Although CTVIS [17] incorporates contrastive loss to enhance instance association across frames, leading to improved performance over MinVIS [15], these query-based association methods underperform compared to position-based approaches like AB3DMOT [34] and 4D-PLS [7]. This observation highlights the importance of emphasizing localization in our task. It also demonstrates that our method effectively perceives both appearance and location information, significantly improving segmentation and tracking performance.

F. Ablation Studies

Localization-aware loss. We introduce the localization-aware loss, which adds position supervision for the reference points of 4D panoptic queries, and observe an increase of 3.5 in OccSTQ and more than a 40% improvement margin in OccAQ, as shown in Tab. III. This demonstrates the loss benefits 4D panoptic queries by capturing more positionally accurate features, leading to more reliable tracking.

The number of training frames. Since increasing the number of training frames better simulates real-world long-sequence driving scenarios, we examine the performance

with 3 and 5 training frames, as reported in Tab. III. The results demonstrate a positive impact on performance when more training frames are added, indicating the scalability of our method in the temporal dimension.

G. Qualitative Comparison

In Fig. 3, we provide a qualitative comparison between our method and 4D-LCA [7]. TrackOcc demonstrates superior performance in tracking both moving and stationary vehicles. In contrast, 4D-LCA struggles with moving vehicles, failing to maintain consistent instance associations over time, as indicated by the red bounding boxes. These results highlight the strong tracking capability of our method.

V. CONCLUSIONS

In this paper, we introduce a new task, Camera-based 4D Panoptic Occupancy Tracking, which jointly tackles occupancy panoptic segmentation and object tracking from camera input. Besides, We propose TrackOcc, a novel camera-based method that achieves 4D panoptic occupancy tracking in a streaming, end-to-end fashion using 4D panoptic queries and localization-aware loss. Experiments demonstrate that TrackOcc outperforms other methods. We hope this work will inspire further research in vision-based perception, particularly in applications like end-to-end autonomous driving systems.

APPENDIX

For more details, please visit `https://github.com/zgchen33/TrackOcc_ICRA2025`.

REFERENCES

[1] Y. Li, Z. Yu, J. Philion, A. Anandkumar, S. Fidler, J. Jia, and J. Alvarez, "End-to-end 3d tracking with decoupled queries," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 302–18 311.

[2] T. Zhang, X. Chen, Y. Wang, Y. Wang, and H. Zhao, "MUTR3D: A Multi-camera Tracking Framework via 3D-to-2D Queries," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. New Orleans, LA, USA: IEEE, June 2022, pp. 4536–4545.

[3] Q. Ma, X. Tan, Y. Qu, L. Ma, Z. Zhang, and Y. Xie, "Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction," *arXiv preprint arXiv:2312.01919*, 2023.

[4] A.-Q. Cao and R. de Charette, "Monoscene: Monocular 3d semantic scene completion," in *CVPR*, 2022.

[5] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3d semantic occupancy prediction," *arXiv preprint arXiv:2302.07817*, 2023.

[6] H. Liu, H. Wang, Y. Chen, Z. Yang, J. Zeng, L. Chen, and L. Wang, "Fully sparse 3d panoptic occupancy prediction," *arXiv preprint arXiv:2312.17118*, 2023.

[7] M. Aygun, A. Osep, M. Weber, M. Maximov, C. Stachniss, J. Behley, and L. Leal-Taixé, "4d panoptic lidar segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5527–5537.

[8] K. Yilmaz, J. Schult, A. Nekrasov, and B. Leibe, "Mask4Former: Mask Transformer for 4D Panoptic Segmentation," in *International Conference on Robotics and Automation (ICRA)*, 2024.

[9] L. Kreuzberg, I. E. Zulfikar, S. Mahadevan, F. Engelmann, and B. Leibe, "4d-stop: Panoptic segmentation of 4d lidar using spatio-temporal object proposal generation and aggregation," in *European Conference on Computer Vision Workshop*, 2022.

[10] X. Tian, T. Jiang, L. Yun, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *arXiv preprint arXiv:2304.14365*, 2023.

[11] Z. Li, Z. Yu, D. Austin, M. Fang, S. Lan, J. Kautz, and J. M. Alvarez, "FB-OCC: 3D occupancy prediction based on forward-backward view transformation," *arXiv:2307.01492*, 2023.

[12] A.-Q. Cao, A. Dai, and R. de Charette, "Pasco: Urban 3d panoptic scene completion with uncertainty awareness," in *CVPR*, 2024.

[13] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Video panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9859–9868.

[14] P. Voigtlaender, J. Luiten, P. H. Torr, and B. Leibe, "Siam r-cnn: Visual tracking by re-detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6578–6588.

[15] D.-A. Huang, Z. Yu, and A. Anandkumar, "Minvis: A minimal video instance segmentation framework without video-based training," *Advances in Neural Information Processing Systems*, vol. 35, pp. 31 265–31 277, 2022.

[16] T. Zhang, X. Tian, Y. Wu, S. Ji, X. Wang, Y. Zhang, and P. Wan, "Dvis: Decoupled video instance segmentation framework," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1282–1291.

[17] K. Ying, Q. Zhong, W. Mao, Z. Wang, H. Chen, L. Y. Wu, Y. Liu, C. Fan, Y. Zhuge, and C. Shen, "Ctvis: Consistent training for online video instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 899–908.

[18] T. Zhang, X. Tian, Y. Zhou, S. Ji, X. Wang, X. Tao, Y. Zhang, P. Wan, Z. Wang, and Y. Wu, "Dvis++: Improved decoupled framework for universal video segmentation," *arXiv preprint arXiv:2312.13305*, 2023.

[19] 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*. PMLR, 2022, pp. 180–191.

[20] 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*. Springer, 2022, pp. 531–548.

[21] 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*, 2023, pp. 172–181.

[22] X. Lai, Y. Chen, F. Lu, J. Liu, and J. Jia, "Spherical transformer for lidar-based 3d recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 545–17 555.

[23] 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*, 2022, pp. 4537–4546.

[24] S. Ding, L. Schneider, M. Cordts, and J. Gall, "Ada-track: End-to-end multi-camera 3d multi-object tracking with alternating detection and association," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 184–15 194.

[25] R. Marcuzzi, L. Nunes, L. Wiesmann, E. Marks, J. Behley, and C. Stachniss, "Mask4D: End-to-End Mask-Based 4D Panoptic Segmentation for LiDAR Sequences," *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 7487–7494, 2023. [Online]. Available: https://www.ipb.uni-bonn.de/wp-content/papercite-data/pdf/marcuzzi2023ral-meem.pdf

[26] M. Zhu, S. Han, H. Cai, S. Borse, M. G. Jadidi, and F. Porikli, "4d panoptic segmentation as invariant and equivariant field prediction," in *International Conference on Computer Vision*, 2023.

[27] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 194–210.

[28] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 1477–1485, Jun. 2023.

[29] 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*. Springer, 2022, pp. 1–18.

[30] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[31] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4794–4803.

[32] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," *arXiv*, 2021.

[33] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[34] X. Weng, J. Wang, D. Held, and K. Kitani, "3D Multi-Object Tracking: A Baseline and New Evaluation Metrics," *IROS*, 2020.

[35] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[38] J. V. Hurtado, R. Mohan, W. Burgard, and A. Valada, "Mopt: Multi-object panoptic tracking," *arXiv preprint arXiv:2004.08189*, 2020.

[39] M. Weber, J. Xie, M. Collins, Y. Zhu, P. Voigtlaender, H. Adam, B. Green, A. Geiger, B. Leibe, D. Cremers, *et al.*, "Step: Segmenting and tracking every pixel," *arXiv preprint arXiv:2102.11859*, 2021.

[40] X. Yan, J. Gao, J. Li, R. Zhang, Z. Li, R. Huang, and S. Cui, "Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3101–3109.

[41] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9404–9413.

[42] D. B. West, "Introduction to graph theory," 2001.