2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Mask4Former: Mask Transformer for 4D Panoptic Segmentation

Kadir Yilmaz[1], Jonas Schult[1], Alexey Nekrasov[1], Bastian Leibe[1]

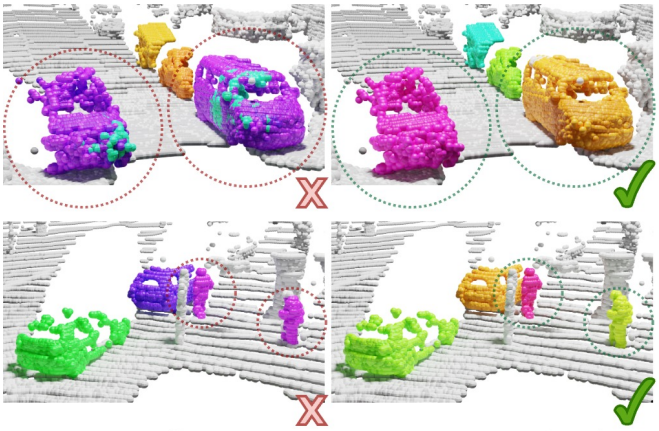
***Abstract*— Accurately perceiving and tracking instances over time is essential for the decision-making processes of autonomous agents interacting safely in dynamic environments. With this intention, we propose Mask4Former for the challenging task of 4D panoptic segmentation of LiDAR point clouds. Mask4Former is the first transformer-based approach unifying semantic instance segmentation and tracking of sparse and irregular sequences of 3D point clouds into a single joint model. Our model directly predicts semantic instances and their temporal associations without relying on hand-crafted non-learned association strategies such as probabilistic clustering or voting-based center prediction. Instead, Mask4Former introduces spatio-temporal instance queries that encode the semantic and geometric properties of each semantic tracklet in the sequence. In an in-depth study, we find that promoting spatially compact instance predictions is critical as spatio-temporal instance queries tend to merge multiple semantically similar instances, even if they are spatially distant. To this end, we regress 6-DOF bounding box parameters from spatio-temporal instance queries, which are used as an auxiliary task to foster spatially compact predictions. Mask4Former achieves a new state-of-the-art on the SemanticKITTI test set with a score of 68.4 LSTQ.**

I. INTRODUCTION

LiDAR is a popular sensor modality in the robotics community due to its ability to provide accurate 3D spatial information. It allows precise scene understanding of the 3D environment over time, which is essential for agents to safely navigate in dynamic environments by predicting traffic movements and identifying potential hazards. To achieve the full potential of LiDAR data, in this work, we address the task of 4D panoptic segmentation. That is, given a sequence of LiDAR scans, the goal is to predict the semantic class of each point while consistently tracking object instances. The research community has made remarkable progress in advancing 3D vision tasks, fueled by the rapid advancement of deep learning methods [27, 34, 42] and the availability of large-scale benchmark datasets [5, 14, 15, 39]. Powerful feature extractors [11, 41, 42, 51] that exploit the rich information offered by LiDAR sensors have been proposed, leading to remarkable improvements in object detection [23, 36, 46], segmentation [29, 42, 51], and tracking [45, 48].

To accomplish holistic 3D scene understanding, 4D panoptic segmentation [2] has recently attracted attention. Traditionally, approaches follow the tracking-by-detection paradigm [32] which decouples 4D panoptic segmentation in the subtasks of semantic segmentation [29, 42], object detection [23] and tracking [31, 45]. While this separation of segmentation, detection, and tracking allows for independent improvements in each component, it tends to neglect joint learning of temporal relationships with semantic information. Significant advances in 4D panoptic segmentation methods address this problem by introducing model architectures that approach the task as a whole and predict semantic class labels for each point and temporally consistent instances [1]. Recent methods generate instance predictions by grouping proposals in the 4D spatio-temporal volume [2, 17, 21] or learned embedding space [28]. However, all previous 4D panoptic segmentation methods fundamentally rely on non-learned clustering methods to aggregate tracklets.

At the same time, we observe a noticeable shift towards unifying tasks [20, 43, 47] and model architectures [7, 10] for holistic scene understanding. Central to this trend are mask transformers [8, 9, 34] that directly predict foreground masks and their associated semantic labels, eliminating the need for non-learned clustering strategies. Typically, these models consist of two main components: a convolutional feature extractor and a transformer decoder. The convolutional feature extractor processes the point cloud and generates multi-scale features. The transformer decoder leverages these extracted features and iteratively refines queries each of



Baseline **Mask4Former** (Ours)

Fig. 1: **Spatially non-compact instances.** Naively adapted for 4D panoptic segmentation, mask transformer approaches reveal a crucial shortcoming: instance predictions tend to be spatially non-compact. As a result, the baseline model predicts two cars as a single object *(left)*. To overcome this limitation, we introduce Mask4Former, which additionally regresses 6-DOF bounding box parameters for the instance trajectory. We find that optimizing these bounding box parameters provides a valuable loss signal that promotes spatially compact instances *(right)*.

[1]Computer Vision Group, RWTH Aachen University, Germany.
Project Page: https://vision.rwth-aachen.de/Mask4Former

which encodes the spatial and semantic features for an instance. Throughout multiple transformer decoder layers, the queries are refined sequentially. Ultimately, these refined queries directly predict the final semantic class and mask predictions, allowing mask transformers to avoid handcrafted grouping. Despite the remarkable performance of mask transformer architectures across diverse tasks, such as image segmentation [9, 10], video segmentation [8], and 3D scene segmentation [27, 34, 38], it remains open whether such a paradigm generalizes to the unique challenges of 4D panoptic segmentation of sparse point cloud sequences.

To answer this question, our goal in this paper is to extend mask transformers to 4D panoptic segmentation of point clouds. Unlike prevailing top-performing approaches for 4D panoptic segmentation [2, 21, 28, 50], we directly predict foreground masks for *thing* instances and *stuff* regions and their associated semantic labels, bypassing the need for post-processing clustering which requires hand-engineered methods and fine-tuned hyperparameters. Therefore, in an initial study, we adapt Mask3D [34] for 4D panoptic segmentation. We follow recent approaches [2, 18, 21] by superimposing consecutive LiDAR scans into spatio-temporal point clouds that are processed by a sparse convolutional feature backbone [11]. Furthermore, we introduce point-wise spatio-temporal positional encoding in the transformer decoder [8]. Our findings indicate that these modifications are already competitive with specialized 4D panoptic segmentation methods [21]. Yet, a deeper examination reveals a significant flaw in mask transformer approaches for 3D point clouds: instances are not always spatially compact [34, 35]. Specifically, an instance query may connect multiple instances in the spatio-temporal point cloud, even if they are spatially distant but share semantic similarities (Fig. 1, *left*).

Based on these findings, we introduce our novel approach called Mask4Former, which is tailored to ensure spatially compact instances, thus unleashing the full potential of mask transformer architectures for 4D panoptic segmentation. We achieve this by regressing 6-DOF bounding box parameters from the spatio-temporal queries, providing a loss signal to foster spatially compact instance predictions (Fig. 1, *right*). We evaluate our Mask4Former model on the challenging SemanticKITTI 4D panoptic segmentation benchmark and achieve state-of-the-art performance on the test set.

In summary, our contributions are fourfold: **(1)** We extend the state-of-the-art instance segmentation method Mask3D [34] to the 4D panoptic segmentation task. **(2)** In experiments, we discover a crucial shortcoming of this straightforward adaptation, namely, the tendency for spatio-temporal instance predictions to lack spatial compactness. **(3)** We propose Mask4Former which effectively addresses the aforementioned limitation by introducing a box regression branch that promotes spatially compact instance predictions in an end-to-end trainable fashion, rather than relying on a geometric grouping mechanism with hand-tuned hyperparameters. **(4)** Mask4Former achieves state-of-the-art performance on the SemanticKITTI 4D panoptic segmentation benchmark.

II. RELATED WORK

Mask Transformers. MaskFormer [10] proposes mask classification as a novel segmentation technique, showcasing its advantages over conventional pixel-based methods. Inspired by DETR [7], it combines CNNs and transformer networks in a universal segmentation architecture, eliminating the need for task-specific architectures, and streamlining development processes. Subsequently, Mask2Former [9] introduces masked attention in the transformer decoder, directing the attention only to relevant parts of the image, and incorporates high-resolution multi-scale features for segmenting smaller objects. This improves convergence and performance, achieving state-of-the-art results in 2D segmentation tasks [20, 24, 49]. The paradigm extends to the video instance segmentation [8] task, where Mask2Former effectively addresses temporal consistency, showcasing its universal applicability. Inspired by its success in 2D, Mask3D [34] applies the mask transformer architecture to the 3D domain by leveraging a sparse convolutional backbone [11], and eliminates the need for the predominantly used center-voting and clustering algorithms [13, 19, 44]. For LiDAR panoptic segmentation, MaskPLS [27] compares mask transformer architectures with adapted semantic segmentation approaches [6, 11, 12, 17, 41, 51], demonstrating the superiority of the mask transformer architecture.

4D panoptic segmentation. 4D-PLS [2] introduces the 4D panoptic segmentation task, associated evaluation metrics, and their method for solving the task. It superimposes consecutive LiDAR scans to form a spatio-temporal point cloud, performs semantic segmentation, and follows a probabilistic approach for clustering instances based on their predicted centers. Along the same lines, 4D-DS-Net [18] and 4D-StOP [21] propose to cluster instances based on spatio-temporal proximity. 4D-DS-Net [18] extends DS-Net [17] to the 4D domain by applying a dynamic shifting module to spatio-temporal point clouds which iteratively refines the estimated instance centers and clusters the points in the spatio-temporal volume. 4D-StOP [21], on the other hand, replaces the probabilistic clustering with an instance-centric voting approach. Here, initial instance proposals are generated using center votes and then aggregated using learned geometric features. Building on the success of 4D-StOP, the concurrent work Eq-4D-StOP [50] predicts equivariant fields and incorporates the necessary layers into the models. This reinforcement of rotation equivariance ensures that the models account for rotational symmetries in the data, resulting in a more robust feature learning. Contrastingly, CA-Net [28] clusters instances in the feature space. It leverages an off-the-shelf 3D panoptic segmentation network [17] and uses extracted point features in a contrastive learning framework [16] to generate instance-wise consistent features, resulting in robust instance associations over time. Bypassing the need for non-learned clustering approaches, the concurrent work Mask4D [26] adopts the mask transformer-based paradigm but opts for queries that encode single frame instances, and re-uses these queries in subsequent frames to facilitate tracking. Unlike

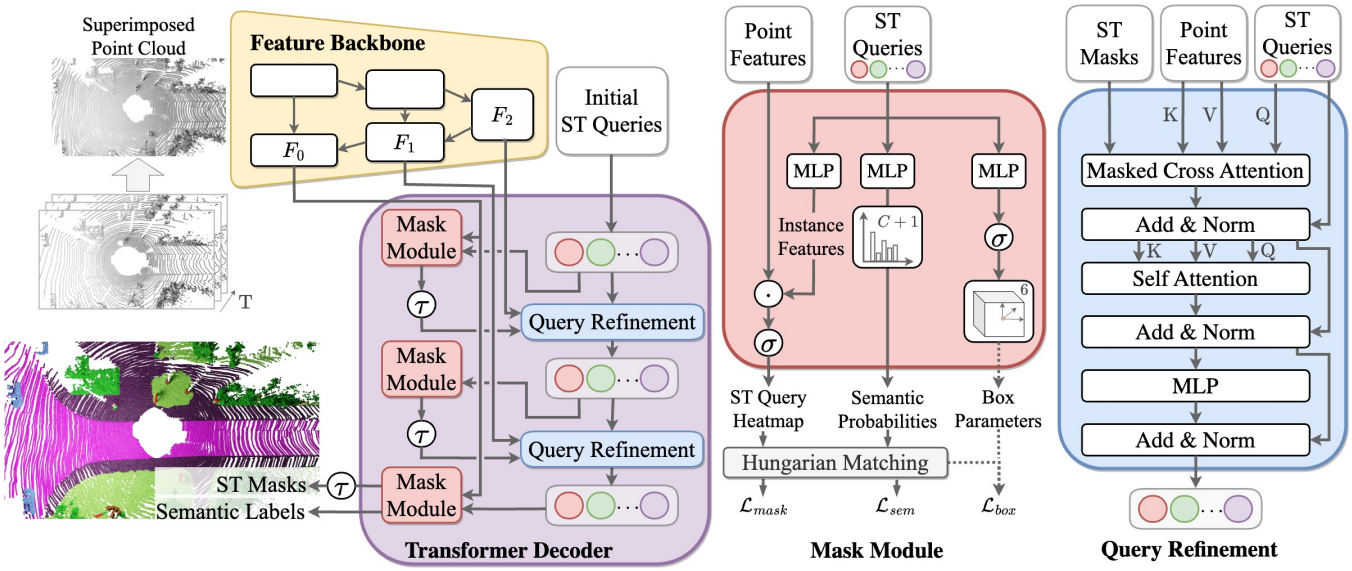


Fig. 2: **Illustration of the Mask4Former model.** We superimpose a sequence of T point clouds into a spatio-temporal representation which is subsequently processed by a sparse convolutional feature backbone. Given a multi-scale feature representation extracted from the feature backbone, the transformer decoder iteratively refines spatio-temporal (ST) instance queries. A mask module consumes ST queries and point features at various scales and predicts semantic class probabilities, instance heatmaps, and a 6-DOF bounding box for each ST query.

previous approaches, Mask4Former unifies segmentation and tracking by directly predicting the spatio-temporal instance masks and their corresponding semantic labels.

III. Method

Inspired by the success of mask transformer approaches for 3D instance segmentation [27, 34, 38] and 2D video instance segmentation [8], we propose Mask4Former – the first mask transformer-based approach for 4D panoptic segmentation. Building on Mask3D [34] for 3D instance segmentation, we introduce technical components that are key to enabling 4D panoptic segmentation of point clouds, *i.e.*, predicting the semantic class of each point and consistently tracking instances over time.

Overview. (Fig. 2) As the input to our model, we use a single voxelized point cloud consisting of superimposed consecutive LiDAR scans. We process the point cloud with a sparse convolutional *feature extractor* (Fig.2,), which generates a multi-resolution voxel representation for the *transformer decoder* . At the core of the model are spatio-temporal (ST) queries that encode geometric and semantic attributes of all instances in a sequence. To learn ST query features, we use a transformer decoder that encompasses consecutive query refinement and mask modules. A *mask module* takes the ST queries and predicts instance heatmaps, semantic class probabilities, and also regresses a bounding box for each instance trajectory. A *query refinement module* updates the ST queries by cross-attending to multi-scale voxel representations. In the following, we provide a detailed description of each component involved.

Input Spatio-Temporal Point Cloud. We represent a temporal sequence of point clouds as a single superimposed and voxelized point cloud. Similar to other approaches [2, 21], we use pose estimates of the ego vehicle [3, 4] to create a single scene containing points from multiple LiDAR scans in a global coordinate frame. Subsequently, this superimposed point cloud represents a spatio-temporal volume, denoted as $\mathcal{P} \in \mathbb{R}^{M \times 3}$, which captures the temporal evolution of the scene. We partition this point cloud into equally sized cubic voxels, thus yielding the representation $\mathcal{V} \in \mathbb{Z}^{K_0 \times 3}$. This voxelization process not only keeps memory constraints in bounds but also allows for efficient processing of the resulting point cloud by sparse convolutional extractors [11].

Feature Backbone. (Fig. 2,) The sparse convolutional feature extractor processes the voxelized point cloud $\mathcal{V} \in \mathbb{Z}^{K_0 \times 3}$ and extracts multi-scale features $F_r \in \mathbb{R}^{K_r \times D_r}$ at various resolutions r. This design allows the network to capture both local geometry and global context while ensuring the preservation of fine-grained spatial details.

Mask Module. (Fig. 2,) Each of the N_q ST queries $\mathbf{X} \in \mathbb{R}^{N_q \times D}$ represents a distinct instance over a time period. The mask module predicts the foreground mask of an instance throughout the sequence and the semantic class of the mask, as well as estimating the 6-DOF bounding box parameters of its trajectory. To generate this binary foreground mask, ST queries are processed by an MLP, and aligned with the feature space of the backbone's output. To obtain spatio-temporal masks at the finest resolution, we compute the dot product with the finest backbone features $\mathbf{F}_0$, which – after sigmoid activation and thresholding – yields the final binary ST mask. In addition to these masks, we predict

semantic class probabilities for each ST query via a linear projection layer to $C+1$ dimensions, followed by a softmax normalization. A critical element for consistent tracking of instances over time is the bounding box regression branch. We feed the ST queries to an MLP followed by sigmoid activation to map the features to a 6-dimensional bounding box parameter space that encodes the normalized bounding box center coordinates (x, y, z) as well as the box dimensions (w, h, d).

Query Refinement Module. (Fig. 2, ❑) Following Cheng *et al.* [9], the query refinement blocks refine the ST queries $\mathbf{X}$ by using the voxel features $\mathbf{F}_r$ at various resolutions r. First, a masked cross-attention layer [9] transforms voxel features $\mathbf{F}_r$ into keys K and values V, while ST queries are mapped to queries Q. Here, ST queries attend only to the foreground voxels predicted by the previous mask module. We then apply self-attention between queries to ensure that multiple queries do not converge on a single instance.

We use spatio-temporal Fourier positional encodings [40] to incorporate both spatial and temporal information into our transformer blocks. To do this, we sum spatial positional encodings based on the voxel positions and temporal positional encodings based on the LiDAR scan time frame [8].

Hungarian Matching. (Fig. 2, ❑) In a single forward pass, Mask4Former determines N_q foreground masks along with their associated semantic class labels. Since both these predictions and the ground truth targets are not in any particular order, it is necessary to establish optimal one-to-one correspondences between them for model optimization. Typically, Mask transformer methods [7, 9, 10] rely on the Hungarian Algorithm [22] for this purpose. The assignment cost for a predicted semantic mask, *i.e.*, *thing* instances and *stuff* regions, and a target mask is defined as follows:

$$\mathcal{C} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{sem}} \tag{1}$$

where $\mathcal{L}_{\text{mask}}{=}\lambda_{\text{dice}}\mathcal{L}_{\text{dice}}{+}\lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}}$ is a weighted combination of the binary cross-entropy loss and the dice loss [30] for supervising foreground mask predictions and $\mathcal{L}_{\text{sem}}{=}\lambda_{\text{CE}}\mathcal{L}_{\text{CE}}$ is the multi-class cross-entropy loss $\mathcal{L}_{\text{CE}}$ for supervising mask semantics. The Hungarian algorithm is applied to solve the assignment problem and to find the globally optimal matches that minimize the total cost while ensuring that each target mask is assigned only once. The unmatched predicted masks are assigned to a "no-object" mask.

Training the model. After establishing one-to-one correspondences, we can directly optimize each predicted mask. Our resulting loss consists of three loss functions: We keep the same binary mask loss $\mathcal{L}_{\text{mask}}$ and the multi-class cross-entropy loss $\mathcal{L}_{\text{sem}}$ from the Hungarian matching as referenced in Eq. 1. Observing that the $\mathcal{L}_{\text{mask}}$ loss does not consider the distance of incorrectly added points to the mask, we introduce a new auxiliary bounding box regression loss $\mathcal{L}_{\text{box}}$ which promotes spatially compact instances. We implement the bounding box loss as an L1 loss on the normalized axis-aligned box parameters. By optimizing the bounding box parameters from ST queries, the spatial location of their corresponding masks is supervised. Consequently, this helps to distinguish similar instances of the same class that are spatially separated. The overall loss is:

$$\mathcal{L} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{box}} \tag{2}$$

Extracting 4D panoptic segmentations. Mask4Former predicts N_q instance tracks as semantic heatmaps which are not necessarily non-overlapping. To assign a single semantic class label and instance ID to every point within the spatio-temporal point cloud, we proceed in the following manner: First, for each spatio-temporal query, we obtain semantic confidence by selecting the semantic class with the maximum probability. Second, this semantic confidence is multiplied with the corresponding instance heatmap, resulting in an overall confidence heatmap. We then assign each point to the query with the maximum confidence.

Tracking over long sequences. To track instances across long LiDAR sequences that exceed memory limits, it is critical to associate instances across successive spatio-temporal point clouds. Therefore, we follow Aygün *et al.* [2] and construct long sequences from short sequences in a way that ensures seamless associations. We establish a one-to-one match between predicted instances in the last and first frames between short sequences.

IV. EXPERIMENTS

A. Comparing with State-of-the-Art Methods.

Dataset. We evaluate Mask4Former on the well-established SemanticKITTI dataset [3], which is derived from the KITTI odometry dataset [15]. The dataset is split into training, validation, and test sets, and consists of over $43,000$ LiDAR scans recorded with a Velodyne-64 laser scanner capturing various urban driving scenarios. Each point in the LiDAR point clouds is densely annotated with one of $C{=}19$ semantic labels, *e.g.*, *car*, *road*, *cyclist*, as well as a unique instance ID that is consistent over time. For every time step, the dataset includes precise pose estimates of the ego vehicle, which is critical for the 4D panoptic segmentation task.

Metric. The LiDAR Segmentation and Tracking Quality Metric (LSTQ) [2] is designed to evaluate the performance of 4D panoptic segmentation algorithms. It consists of two main components: classification and association scores. The classification score S_{cls} evaluates how well the algorithm performs in assigning correct semantic labels to the LiDAR points. It is calculated as the instance-agnostic mean intersection over union (mIoU) over all classes. The association score S_{assoc} evaluates the quality of point-to-instance associations considering the entire LiDAR sequence. It measures how well the algorithm tracks object instances over time without considering the semantic predictions. The overall LSTQ metric is computed as the geometric mean of the classification score and the association score: $LSTQ = \sqrt{S_{cls} \times S_{assoc}}$. The geometric mean ensures that a high score can only be obtained if the approach performs well in both the classification and the association task.

TABLE I: **Scores on SemanticKITTI validation.** SemanticKITTI 4D panoptic segmentation validation set results. *Abbreviations:* PP: PointPillars [23], MOT: Multi-Object Tracking [45], SFP: Scene flow based propagation [31]. * denote concurrent work.

Method	LSTQ	S_{assoc}	S_{cls}	IoU^{St}	IoU^{Th}
KPConv [42]+PP+MOT	46.3	37.6	57.0	64.2	54.1
RangeNet++ [29]+PP+SFP	43.4	35.7	52.8	60.5	42.2
KPConv [42]+PP+SFP	46.0	37.1	57.0	64.2	54.1
4D-PLS [2]	62.7	65.1	60.5	65.4	61.3
4D-StOP [21]	67.0	74.4	60.3	65.3	60.9
4D-DS-Net [18]	68.0	71.3	64.8	64.5	65.3
Eq-4D-StOP* [50]	70.1	**77.6**	63.4	66.4	67.1
Mask4D* [26]	**71.4**	75.4	**67.5**	65.8	**69.9**
Mask4Former (Ours)	70.5	74.3	66.9	**67.1**	66.6

TABLE II: **Scores on SemanticKITTI test.** * denote concurrent work.

Method	LSTQ	S_{assoc}	S_{cls}	IoU^{St}	IoU^{Th}
KPConv [42]+PP+MOT	38.0	25.9	55.9	66.9	47.7
RangeNet++ [29]+PP+SFP	34.9	23.3	52.4	64.5	35.8
KPConv [42]+PP+SFP	38.5	26.6	55.9	66.9	47.7
4D-PLS [2]	56.9	56.4	57.4	66.9	51.6
4D-DS-Net [18]	62.3	65.8	58.9	65.6	49.8
CIA [28]	63.1	65.7	60.6	66.9	52.0
4D-StOP [21]	63.9	69.5	58.8	67.7	53.8
Eq-4D-StOP* [50]	67.8	**72.3**	63.5	70.4	61.9
Mask4D* [26]	64.3	66.4	62.2	69.9	52.2
Mask4Former (Ours)	**68.4**	67.3	**69.6**	**72.7**	**65.3**

Implementation Details. In all experiments, we use N_q=100 ST queries which are initialized with Farthest Point Sampled (FPS) point positions [33, 34]. Each spatio-temporal point cloud is formed by superimposing 2 consecutive LiDAR scans which are voxelized with a voxel size of 5 cm. The sparse feature backbone is a Minkowski Res16UNet34C [11]. We train the model for 30 epochs with a batch size of 4 using the AdamW optimizer [25] and the one-cycle learning rate scheduler [37] with a maximum learning rate of $2 \cdot 10^{-4}$. We perform standard data augmentation techniques including random rotation, translation, scaling, and instance population [46]. For the test set submission, we employ random rotation and translation as test time augmentations to enhance the semantic predictions.

Results. In Tables I and II, we report the scores on the SemanticKITTI 4D panoptic segmentation validation and test set, respectively. Mask4Former outperforms previous approaches by at least +2.5 LSTQ on the validation set and +4.5 LSTQ on the test set. Notably, Mask4Former demonstrates strong semantic understanding by achieving at least +9.0 S_{cls} improvement over previous methods on the test set.

B. Analysis Experiments.

Spatio-Temporal Formation. We achieve a globally consistent sequence of LiDAR scans by leveraging the precise pose estimates from the LiDAR sensor [4]. Considering that the sparse convolutional feature backbone (Fig. 2, ▢) can process 3- and 4-dimensional inputs [11], we investigate which representation is best for extracting meaningful spatio-temporal features from a sequence. In Table III, we explore 3 different strategies for representing spatio-temporal feature volumes. Similar to Cheng *et al.* [8], in the first option ①, we process each LiDAR frame individually and then concatenate them along the spatial dimension before passing them to the Transformer decoder. In the second option ②, we represent a LiDAR sequence as a 4D feature volume, which is fed into a 4D sparse convolutional feature backbone [11], facilitating the learning of both spatial and temporal relationships directly within the backbone. Incorporating temporal data early in the backbone shows significant improvements in association quality, yielding an increase of +6.8 S_{assoc}. Given the inherent sparsity of point clouds, the third approach ③ superimposes, *i.e.* concatenates, several point clouds into a single 3D volume [2, 21]. We suspect that superimposing LiDAR scans leads to a denser representation, that is less susceptible to noise, yielding the best performance (Tab. III).

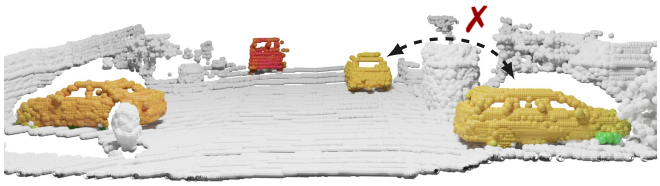
TABLE III: **Spatio-Temporal Formation.** We compare 3 different strategies for representing LiDAR point cloud sequences. We observe that it is key to enable the feature backbone to incorporate temporal information in the feature representation by creating a 4D spatio-temporal representation or superimposing 3D scans, leading to association improvements of up to +7.8 S_{assoc}.

Feature Extraction	LSTQ	S_{assoc}	S_{cls}	IoU^{St}	IoU^{Th}
① Sequential 3D	64.3	65.8	62.8	64.0	61.2
② Spatio-temporal 4D	68.8	72.6	65.2	66.0	64.1
③ Superimposed 3D	**70.2**	**73.6**	**66.9**	**67.2**	**66.5**

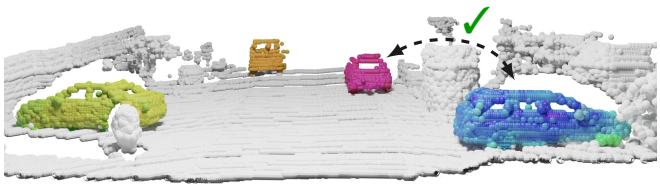
TABLE IV: **Ablation study on bounding box regression.** We observe that optimizing Mask4Former using the regressed bounding box parameters leads to substantially better association scores compared to the baseline (+3.5 S_{assoc}).

	$\mathcal{L}_{box}$	DBS	LSTQ	S_{cls}	S_{assoc}
①	✗	✗	68.6	**67.3**	70.1
②	✗	✓	70.1	**67.3**	72.8 (+2.7)
③	✓	✗	70.2	66.9	73.6 (+3.5)
④	✓	✓	**70.5**	66.9	**74.3** (+4.2)

Spatially non-compact instance predictions. Achieving consistent tracking of multiple instances over time in LiDAR sequences is particularly challenging. This is due to the sparsity of the point clouds, as well as the occlusions and deformations that instances undergo over time, requiring robust temporal feature learning. In an initial study, we analyze our baseline method without the bounding box regression branch in the mask module (Fig. 2, ▢ and Tab. IV, ①), which reveals a crucial shortcoming of applying mask-transformer approaches directly to the task of 4D panoptic segmentation: Instance predictions tend to lack spatial compactness, *i.e.*, the spatio-temporal queries group multiple instances with similar semantics together, even if they are spatially distant (Fig. 1, *left*). To validate this observation, we apply the density-



(a) Mask4Former without box loss



(b) **Mask4Former** (Ours)

Fig. 3: **Visualization of learned point representations.** We use PCA to project the learned point representation of instances into RGB space. Our model trained without bounding box supervision, exhibits reduced variance in its feature representation for instances. In contrast, Mask4Former effectively separates distinct instances in the feature space.

based clustering method, DBSCAN [6], to each foreground mask prediction. This separates the instance mask predictions into spatially compact instances. The impact was noticeable: applying DBSCAN ② to the instance predictions results in a significant improvement of +2.7 S_{assoc}, confirming our initial findings and supporting our hypothesis. Anticipating further improvements by replacing DBSCAN with a learned component, we introduce a specialized box regression branch ③ which promotes spatial awareness to better separate instances. This approach outperforms the baseline, both with and without DBSCAN, by a margin of up to $+3.5\,S_{assoc}$. Combining the box regression branch with DBSCAN yields our proposed method Mask4Former ④, which not only ensures a strong association between instances ($+4.2\,S_{assoc}$) but also achieves strong semantic scene understanding, scoring $66.9\,S_{cls}$ on the SemanticKITTI validation.

Visualization of point features learned by Mask4Former. In Fig. 3, we show examples of PCA projected features F_0 extracted from the finest resolution of Mask4Former's feature backbone (Fig. 2, □). When trained without our suggested box loss, Mask4Former shows less distinct separation of instance point features within the feature space (Fig. 3a). Conversely, the model optimized with the auxiliary task of 6-DOF bounding box regression for each instance trajectory shows a distinct separation of instance point features in the feature space (Fig. 3b). This indicates that Mask4Former learns a more semantically meaningful feature space for the task of 4D panoptic segmentation leading to its superior association score S_{assoc}, as highlighted in Tab. IV.

Qualitative results. In Fig. 4a, we show qualitative results. We observe that Mask4Former not only produces sharp instance masks but also reliably tracks the moving bicyclist ● throughout the entire sequence. We also demonstrate a failure case of our tracking approach. As we process long sequences by stitching short sequences with overlaps, we incorrectly split tracks when an instance is not present in the overlapping

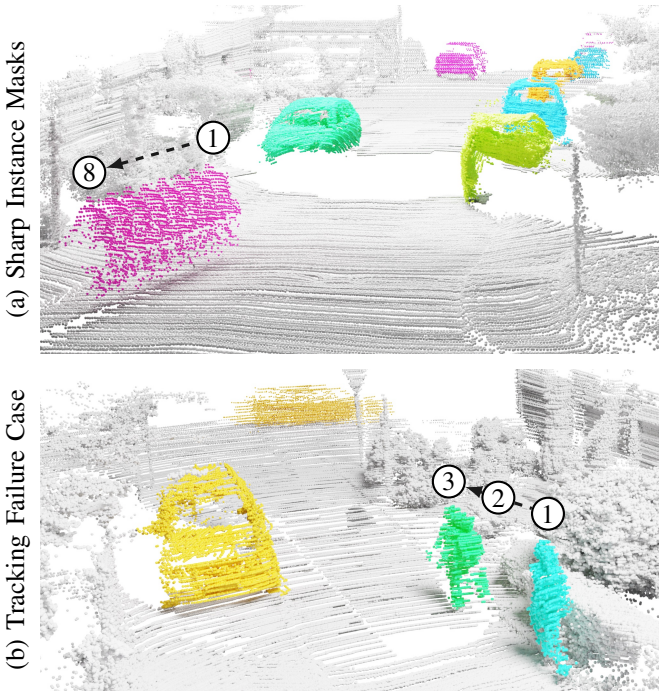


Fig. 4: **Qualitative Results.** We show color-coded instance tracks over 8 superimposed frames in a spatio-temporal point cloud and a failure where a pedestrian track is split due to an observation being outside of the LiDAR's field of view.

LiDAR scan. For example, in Fig. 4b, a pedestrian near the ego vehicle falls below the LiDAR's field of view. As a result, when the pedestrian becomes visible again, our tracking approach fails and predicts it as a new instance.

V. CONCLUSION

Inspired by the success of recent mask transformer-based approaches, we have extended Mask3D to the task of 4D panoptic segmentation and have achieved promising results. In an in-depth analysis, we have found that Mask3D for 4D panoptic segmentation tends to produce spatially non-compact instances, resulting in poor association quality. To overcome this limitation, we have introduced Mask4Former, the first transformer-based approach, that unifies segmentation and tracking of 3D point cloud sequences and is tailored to ensure spatially compact instances. To this end, Mask4Former regresses 6-DOF bounding box parameters that are optimized to provide a loss signal to encourage spatially compact instance predictions. Through extensive experimental evaluations, we have demonstrated the effectiveness of Mask4Former, achieving state-of-the-art performance on the SemanticKITTI 4D panoptic segmentation benchmark. We anticipate follow-up work along the lines of direct prediction of instance and semantic labels.

Acknowledgments: This project is partially funded by the Bosch-RWTH LHC project "Context Understanding for Autonomous Systems", the BMBF project 6GEM (16KISK036K) and the NRW project WestAI (01IS22094D). Compute resources were granted by RWTH Aachen under project supp0003. This work is part of the first author's master thesis.

REFERENCES

[1] Ali Athar, Enxu Li, Sergio Casas, and Raquel Urtasun. 4D-Former: Multimodal 4D Panoptic Segmentation. In *Conference on Robot Learning*, 2023.

[2] Mehmet Aygun, Aljosa Osep, Mark Weber, Maxim Maximov, Cyrill Stachniss, Jens Behley, and Laura Leal-Taixé. 4D Panoptic Lidar Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *International Conference on Computer Vision*, 2019.

[4] Jens Behley and Cyrill Stachniss. Efficient Surfel-Based SLAM using 3D Laser Range Data in Urban Environments. In *Robotics: Science and Systems*, 2018.

[5] 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. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[6] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, 2013.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.

[8] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv:2112.10764*, 2021.

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[10] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Neural Information Processing Systems*, 2021.

[11] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[12] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[13] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3D-MPA: Multi-proposal aggregation for 3d semantic instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[14] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. In *IEEE Robotics and Automation Letters (RA-L)*, 2022.

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[17] Fangzhou Hong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Lidar-based panoptic segmentation via dynamic shifting network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[18] Fangzhou Hong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. LiDAR-based 4D Panoptic Segmentation via Dynamic Shifting Network. In *arXiv:2203.07186*, 2022.

[19] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[20] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[21] Lars Kreuzberg, Idil Esen Zulfikar, Sabarinath Mahadevan, Francis Engelmann, and Bastian Leibe. 4D-StOP: Panoptic Segmentation of 4D LiDAR using Spatio-temporal Object Proposal Generation and Aggregation. In *European Conference on Computer Vision Workshop*, 2022.

[22] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52, 1955.

[23] 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*, 2018.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

[26] 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*, 2023.

[27] Rodrigo Marcuzzi, Lucas Nunes, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. Mask-based panoptic lidar segmentation for autonomous driving. *IEEE Robotics and Automation Letters*, 2023.

[28] Rodrigo Marcuzzi, Lucas Nunes, Louis Wiesmann, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Contrastive instance association for 4d panoptic segmentation using sequences of 3d lidar scans. In *IEEE Robotics and Automation Letters (RA-L)*, 2022.

[29] Andres Milioto, Ignacio Vizzo, Jens Behley, and C. Stachniss. RangeNet++: Fast and Accurate LiDAR Semantic Segmentation. In *International Conference on Intelligent Robots and Systems*, 2019.

[30] F Milletari, N Navab, SA Ahmadi, and V-net. Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*, 2016.

[31] Himangi Mittal, Brian Okorn, and David Held. Just Go With the Flow: Self-Supervised Scene Flow Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[32] Kenji Okuma, Ali Taleghani, Nando De Freitas, James J Little, and David G Lowe. A boosted particle filter: Multitarget detection and tracking. In *European Conference on Computer Vision*, 2004.

[33] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[34] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D for 3D Semantic Instance Segmentation. In *IEEE International Conference on Robotics and Automation*, 2023.

[35] Yichao Shen, Zigang Geng, Yuhui Yuan, Yutong Lin, Ze Liu, Chunyu Wang, Han Hu, Nanning Zheng, and Baining Guo. V-DETR: DETR with Vertex Relative Position Encoding for 3D Object Detection. In *International Conference on Learning Representations*, 2024.

[36] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[37] Leslie N. Smith and Nicholay Topin. Super-convergence: very fast training of neural networks using large learning rates. In *Defense + Commercial Sensing*, 2017.

[38] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint Transformer for 3D Scene Instance Segmentation. In *Conference on Artificial Intelligence*, 2022.

[39] 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 M. 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. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[40] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. In *Neural Information Processing Systems*, 2020.

[41] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European Conference on Computer Vision*, 2020.

[42] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *International Conference on Computer Vision*, 2019.

[43] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and B. Leibe. MOTS: Multi-Object Tracking and Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[44] Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. SoftGroup for 3D Instance Segmentation on 3D Point Clouds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[45] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. In *International Conference on Intelligent Robots and Systems*, 2019.

[46] Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely Embedded Convolutional Detection. *Sensors*, 2018.

[47] Linjie Yang, Yuchen Fan, and Ning Xu. Video Instance Segmentation. In *International Conference on Computer Vision*, 2019.

[48] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3D Object Detection and Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[49] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[50] Minghan Zhu, Shizong Han, Hong Cai, Shubhankar Borse, Maani Ghaffari Jadidi, and Fatih Porikli. 4D Panoptic Segmentation as Invariant and Equivariant Field Prediction. *International Conference on Computer Vision*, 2023.

[51] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and Asymmetrical 3D Convolution Networks for LiDAR Segmentation. *arXiv:2011.10033*, 2020.