CoSA-3D: Vision-Only Automatic 3D Annotation for Cooperative Perception

Chenyang Lu[1], Jiahao Wang[1*], Mingda Yu[1], Dianqiao Lei[1], Mengyang Zhong[1],
Chuang Zhang[1], Lei He[1], Shaobing Xu[1†], Jianqiang Wang[1†]
[1]Tsinghua University
lucy23@mails.tsinghua.edu.cn, {shaobxu, wjqlws}@tsinghua.edu.cn

Abstract

The creation of large-scale 3D datasets for cooperative perception is hindered by the high cost of LiDAR sensors and manual annotation. While vision-only methods offer a low-cost alternative, they suffer from inherent scale ambiguity. This paper introduces CoSA-3D, a novel vision-only,training-free offline framework for automatic 3D annotation in cooperative scenarios. Our method utilizes easily obtained multi-agent images and builds upon the latest zero-shot Structure-from-Motion (SfM) foundation models. A core contribution is our geometric fiducial alignment module, which leverages inter-agent relative poses to rectify SfM-generated pose inaccuracies and recover the true metric scale. This approach, combined with robust multi-agent fusion, effectively handles asynchronous data and overcomes occlusion. We evaluated CoSA-3D on the Griffin dataset, demonstrating the best label quality that significantly surpasses existing methods, particularly in challenging long-range (50-100m) scenes. The framework's generalization is further validated on a custom-collected real-world cooperative dataset. Ablation studies validate that our geometric alignment and data fusion mechanisms are fundamental to the framework's high accuracy. CoSA-3D provides a scalable, accurate, and LiDAR-free solution for 3D cooperative annotation.

1. Introduction

As autonomous vehicles have advanced significantly, single-vehicle perception struggles to meet the growing demands of broader scenarios and more complex targets. A single agent's perspective is fundamentally constrained by occlusion and a restricted sensing range. Consequently, cooperative perception has emerged as one of the key technologies for improving autonomous driving systems, enabling multiple intelligent agents to share perceptual information and thereby improve their understanding and responsiveness to the surrounding environment. This paradigm has spurred the development of numerous cooperative perception datasets [14, 30, 36], which provide vital resources for model training and evaluation.

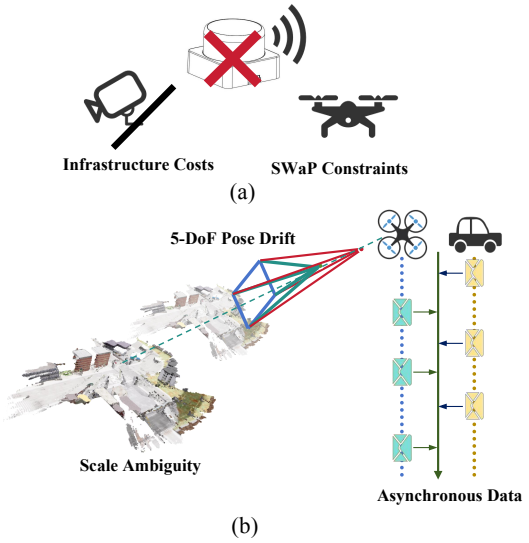

Figure 1. The challenge of existing methods. (a) The impracticality of LiDAR due to high infrastructure costs and strict SWaP constraints on agents like drones. (b) The inherent limitations of vision-only reconstruction, including scale ambiguity, 5-DoF pose drift, and handling asynchronous data streams.

However, the large-scale creation of these datasets is severely constrained by the challenges of 3D annotation: Some datasets remain limited to the relatively simple 2D annotation [1, 15] ,while others with 3D labels rely on expensive LiDAR sensors and labor-intensive manual annotation pipelines [6, 8, 36]. As shown in Fig.1(a) , the high equipment costs and reliance on manual labor render large-scale annotation prohibitively expensive and time-consuming. Furthermore, cooperative perception systems increasingly incorporate diverse agents, such as Unmanned Aerial Vehicles (UAVs) [8, 24], which offer wider vantage points but face strict Size, Weight, and Power (SWaP) con-

*Project Lead
†Corresponding authors

straints. These constraints render multi-LiDAR arrays for surround-view coverage impractical, while single-sensor payloads struggle with sparse point density and a narrow Field of View (FoV) at altitude. Therefore, **automated, low-cost, and accurate 3D annotation techniques** are needed to support multi-agent dataset construction.

Purely vision-based 3D automatic annotation presents a promising, low-cost alternative. However, leveraging 3D reconstruction models for this task introduces several challenges, as shown in Fig.1(b). First, scale ambiguity makes it difficult to maintain a consistent scale and orientation in multi-view, multi-agent collaborative settings. Second, the vision-based 3D reconstruction models usually suffer from 5-DoF limitations, making it hard to locate the exact positions of cameras. Third, asynchronous acquisition among agents means data cannot be perfectly aligned in space and time, degrading overall annotation accuracy. Solving automatic 3D annotation therefore requires new approaches that directly address these scale, pose and asynchrony issues.

To address these challenges, this paper proposes **CoSA-3D (Cooperative, Scale-aware, Asynchronous 3D Auto-Annotation)**, an automatic 3D annotation framework that is LiDAR-free and operates primarily on multi-agent visual data. The framework exploits relative pose information between different agents to resolve scale ambiguity and correct pose drift. It addresses asynchronous data capture through static-scene alignment and a decoupled handling of dynamic objects. By eliminating the need for LiDAR, CoSA-3D drastically reduces annotation cost and time, offering a scalable and accurate solution for large-scale multi-agent dataset annotation. The main contributions of this work are threefold:

- CoSA-3D introduces, for the first time, a LiDAR-free automatic annotation pipeline for collaborative 3D perception using 3D reconstruction models, significantly reducing the cost of dataset collection and labeling.
- We propose a novel geometric fiducial alignment method. It uses inter-agent relative poses to recover a true metric scale for the scene and refine poses, tackling the scale ambiguity and pose noise inherent in visual reconstruction.
- The proposed framework achieves state-of-the-art (SOTA) annotation accuracy on the Griffin dataset, demonstrating robust performance, particularly in challenging long-range scenarios. We further validate our framework's practical effectiveness and generalization on a custom real-world cooperative dataset, where it also demonstrates strong performance.

2. Related Work

We organize the related work into the following three parts, as they respectively provide the data foundations and task-specific methodologies underpinning CoSA-3D.

2.1. Cooperative Perception Dataset

Cooperative perception uses multi-agent data (from vehicles, roadside fixed cameras, and drones) for joint analysis, thereby extending the capabilities of autonomous driving(AD) and traffic management. Currently, there are already many simulation datasets [5, 14, 24, 27, 30] and simulation testing platforms [16]. They are low-cost, but there is a clear gap between them and the real world. To better reflect how autonomous driving models perform in real environments, real-world cooperative datasets with real scenes are indispensable. However, real datasets heavily rely on LiDAR and manual annotation [6, 8, 31, 36] making them costly and inefficient. Therefore, we aim to achieve automatic labeling through LiDAR-free methods to reduce costs and improve the practical feasibility of annotation.

2.2. Automatic 3D Annotation for AD

There are already multiple automatic labeling solutions for autonomous driving. Among them, ZOPP [17] integrates the strong zero-shot recognition capability of vision foundation models with 3D information derived from point clouds, being the first attempt toward multi-modal panoptic perception and automatic annotation for autonomous-driving scenarios. However, these methods share the following problems: First of all, they [17–20, 35] use LiDAR point clouds and are not purely vision-based. Therefore, they are limited by the insufficient depth-prediction capability of current vision models. In addition, it is difficult to transfer multi-sensor auto-labeling methods [20] to multi-agent collaborative scenarios: they require sensor data to be captured synchronously, but in collaborative scenarios, different agents cannot achieve hardware-level synchronization.

Thus, CoSA-3D aims to use a purely vision-based collaborative approach for 3D annotation, while also working to address the problems of asynchronous data collection and noisy localization between agents.

2.3. Vision foundation model

Since CoSA-3D is a pure vision model, it needs zero-shot vision foundation models to detect targets in 2D images and lift to 3D.

At present, there are already many models [37] that can provide 2D semantic information. SAM [13] is the foundation of other similar models [12, 21, 22] for image segmentation. Grounded SAM [22], with the ability of controllable image editing and to tackle a wide range of vision tasks perfectly, stands out among SAM-based models. Since CoSA-3D is constructed on top of 2D annotations, it uses Grounded SAM [22] to generate 2D detection.

Frameworks and methods for depth estimation from monocular images are popular research topics, with many models [4, 32, 33] created in this field. To make the most of cooperative data, we choose to use multi-view scene recon-

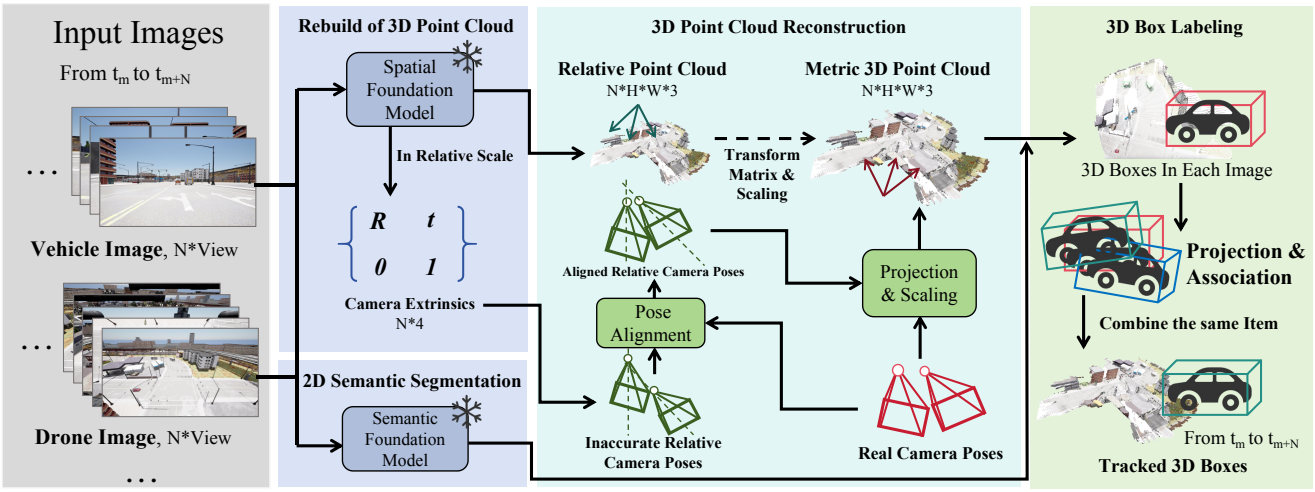

Figure 2. The main pipeline of CoSA-3D. Our model takes images from both vehicle and drone sides as input. The vision foundation models generate a relative point cloud and 2D masks, and they are reconstructed to metric scale then. In 3D Box labeling, the target point cloud turns into tracked 3D boxes as output.

struction to solve the 3D annotation problem. At present, there are already multiple multi-view reconstruction methods [10]. DUSt3R [26] provided a foundation for this field, on top of it, MUSt3R [2] proposes a multi-view network for stereo 3D reconstruction. When handling multi-view data in collaborative scenarios, the model can take advantage of the latest LiDAR-free multi-view scene reconstruction method, VGGT [25], which runs fast and performs well in 3D scene reconstruction and depth prediction. To improve VGGT's [25] poor performance on large-scale scenes, its successor Pi3 [28] provides its solution: by predicting affine-invariant camera poses and scale-invariant local point maps, it breaks the dependence of visual geometric reconstruction on a fixed reference view. These better 3D scene reconstruction works make it possible to vision-only 3D labeling.

Building on Grounded SAM [22] and Pi3 [28], CoSA-3D can directly utilize the obtained sparse point cloud. Notably, our method is not limited to these two models and has strong generalization capabilities for other similar frameworks. This approach saves the step of using LiDAR sensors to generate a dense point cloud, thereby saving time and improving efficiency.

3. Method

3.1. Pipeline Overview

The primary objective of our training-free, offline framework is to achieve fully vision-based, automatic 3D annotation for cooperative perception datasets, providing an efficient and low-cost alternative to systems reliant on LiDAR or manual labeling. As shown in Fig.2, we formally define this task as an offline label generation process, generating accurate 3D bounding boxes, complete with temporally continuous trajectories, for all objects within a pre-recorded scene. The input consists of potentially asynchronous surround-view images captured from multiple agents. The final output is a unified set of static 3D annotations represented in a single world coordinate system, which can be directly utilized to train downstream cooperative perception models. To accomplish this, the CoSA-3D framework is structured into three main components:

1. **Vision Foundation Model**: extracting semantic information and object masks from multi-view images using general-purpose vision foundation models such as PI3 and Grounded-SAM;
2. **3D Point Cloud Reconstruction**: revising the camera extrinsics using geometric fiducial alignment, recovering 3D point clouds at real-world scale via multi-view geometric relations, and registering them into a unified coordinate frame;
3. **3D Box Labeling**: generating oriented 3D bounding boxes (OBBs) for each object in the aligned point cloud, followed by multi-view fusion and temporal association for unified labeling and cross-frame tracking.

The resulting 3D annotations are expressed in real-scale parameters relative to each agent's coordinate system and can be flexibly converted to meet the format requirements of different datasets:

$$[(x, y, z, l, w, h, yaw(\theta))]$$

where (x, y, z) represents the target center position in a certain coordinate system, (l, w, h) represents the target's metric scale and $yaw(\theta)$ is its direction.

3.2. Vision Foundation Model

Since the 3D reconstruction foundation model addresses asynchronous data capture through static-scene alignment

and decoupled handling of dynamic objects, we can take either synchronized or asynchronous images and order them in a continuous sequence. Subsequently, multiple consecutive frames are selected within a continuous time window and organized into several image subsets according to the viewing order: $ImageSet = \{Image(H, W)_{t \to t+N}\}$. Adjacent subsets share overlapping frames to enable cross-frame object association and tracking.

3.2.1. Semantic Segmentation

By employing a well-established 2D semantic segmentation model, object positions can be efficiently obtained for images from different viewpoints. For the multi-view images collected from multiple intelligent agents, our framework applies a unified semantic segmentation model to each image independently to generate the corresponding 2D object masks. These masks offer precise pixel-level delineations of object regions. Compared with directly using 2D bounding boxes, leveraging pixel-wise masks effectively mitigates recognition errors caused by occlusions, thereby improving the reliability of cross-view object identification. Through this process, the model extracts each object's semantic label and its associated pixel mask from the 2D images, i.e. for one pixel mask $P_{mask,j}$, it is defined as the set of pixel coordinates (u, v) belonging to the target:

$$P_{mask,j} = \{(u, v) \mid \text{pixel } (u, v) \text{ from 2D mask}\}$$

They are subsequently utilized in the stages of spatial alignment and 3D reconstruction.

3.2.2. Relative-Scale Point Cloud Reconstruction

Once we obtain the 2D information, the model needs to lift the 2D mask into 3D space to generate 3D bounding boxes for the targets in the scene. Since obtaining pixel-level depth is key to producing 3D boxes, we perform 3D scene reconstruction for each image subset. CoSA-3D adopts the Structure-from-Motion (SfM)–based reconstruction method to establish the initial mapping between the images and the point cloud.

For each image subset $ImageSet(M, H, W), M = N * N_{views}$, N_{view} meaning the number of all agents views, the SfM model will generate a world point cloud containing all image in the subset that is in one-to-one correspondence with the image pixels: $P_{pred} = (M, H, W, 3)$. This point cloud maps points from all images into a unified reconstructed 3D world; at the same time, the model can also retrieve the point cloud corresponding to each individual image. The base vision model also provides the predicted camera extrinsics for all images: $E_{pred,cam} = (M, 4)$.

3.2.3. Limitations of SfM Models

However, the scale and camera extrinsics of point clouds produced by SfM cannot directly meet high-precision annotation requirements. Because a SfM model does not know the real-world scale of the reconstructed scene, it cannot estimate absolute depth values; therefore, we refer to the point clouds reconstructed by SfM as "relative."

Moreover, most SfM literature [7, 23] notes a 5-DoF limit in camera pose estimation: rotations are typically estimated more accurately than translations, with translation/scale more prone to drift—especially under short baselines, sparse keyframes, or long-range scenes, which brings huge difficulty in scale recovery. Since the camera orientation is relatively accurate, the model can accept the optical axis in the predicted extrinsics to be approximately correct, while the distance along that axis remains uncertain, as shown in Fig.1. For UAV-acquired data, this problem is even more pronounced due to the larger distance scales.

3.3. Recovery of Prediction Extrinsics

In this subsection, we first address geometric fiducial alignment to correct the model's predicted camera extrinsics. To reduce the impact of uncertainty in the predicted camera poses, we fully leverage the original dataset by using the actual camera positions to correct the predicted extrinsics $E_{pred,cam} = \{R, t\}$. As shown in the previous section, we may regard the optical axes of all predicted camera extrinsics as accurate—that is, the rotation R in all camera extrinsics is reliable—so the quantity that needs to be corrected is the translation t. We present the optical axes as:

$$A_{pred}(k) = t_{pred} + k \cdot \hat{D}_{pred} \tag{1}$$

Here, k represents the position of the camera on the axis, $\hat{D}_{pred}$ is the unit vector of the axis. Because the optical axis is already fixed-putting it vividly-we only need to "move" the camera along that axis.

Using the raw data for recovery, we need to identify a common "fixed point" on one agent between the real and prediction positions. In our method, shown in Fig.3, this fixed point is defined as the point with the smallest total

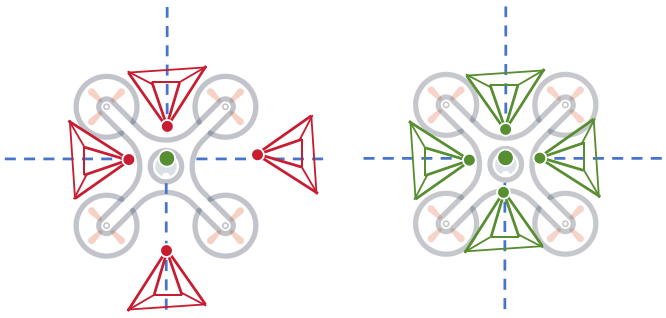

Figure 3. The original camera extrinsics from the SfM model cannot find the correct location along its optical axis, shown as the red one. Green represents the camera position of the real world. But we can use one fixed point (the green point in the image) which is considered accurate relative to the agent.

distance to all optical axes, which is easier to locate the position relationship between them. Given the accuracy of the optical axes, we assume the relative position relations between the set of axes and the fixed point is approximately the same in both predicted and real world. Therefore, using the direction vector from a real camera position to the fixed point on the agent:

$$\hat{V}_{real,ego\to cam} = P_{real,cam} - P_{real,fixed} \tag{2}$$

representing the relations between the cameras and the fixed point, where $V_{real,ego\to cam}$ means the normalized vector from fixed point $P_{real,fixed}$ to camera position $P_{real,cam}$.

Using (1) (2), we trace the same direction from the predicted fixed point and find the point closest to the corresponding camera's predicted optical axis:

$$\begin{cases} L_{pred,ego\to cam}(s) = P_{pred,ego} + s \cdot \hat{V}_{real,ego\to cam} \\ (s^*, k^*) = \arg\min_{s,k} \|L_{pred,ego\to cam}(s) - A_{pred}(k)\|^2 \end{cases} \tag{3}$$

Solving the function (3), we can find the point that is taken as the corrected camera position $t' = t_{pred} + k^* \cdot \hat{D}_{pred}$, yielding the corrected camera extrinsics $E'_{pred,cam}$.

3.4. Real Point Cloud Registration

With the predicted camera extrinsics corrected, we proceed to register the predicted point clouds, turning the relative scale mentioned in Sec.3.2.3 to real-world scale. Our model separates the registration into two steps: because the model's predicted world is accurate in terms of relative positions, we firstly need to align it with the real world in length scale as factor k_{scale}; since the reference coordinate system of the predicted world and the real world are not the same, an extra rigid transformation $T_{pred\to real}$ is required to unify the two point clouds. Therefore, we assume that mapping the predicted point cloud P_{pred} to the real point cloud P_{real} is the result of a global scaling in length together with a rigid transformation:

$$P_{real} = T_{pred\to real} \cdot k_{scale} \cdot P_{pred} \tag{4}$$

Since a rigid transformation does not change distances between objects, the global scale can be determined from the real and the predicted agent position. Using the relations between the real cameras and agents position, we can obtain the predicted agent position O_{pred}. Taking consideration of cooperative distance between two agents, their distance presents as an excellent reference for calculating scale:

$$k_{scale} = \frac{1}{N_p} \sum_{i_1 \neq i_2}^{N_p} \{ \frac{||O_{real,i_1} - O_{real,i_2}||}{||O_{pred,i_1} - O_{pred,i_2}||} \} \tag{5}$$

We calculate the Euclidean distance for each corresponding pair (total pairs number N_p), then take the average to obtain scale. After aligning the length scale, we need to estimate the transformation matrix. After aligning the global scale, we employ the Kabsch algorithm [11] to estimate the optimal rigid transformation matrix $T_{\text{pred}\to\text{real}} = \{R, t\}$. This is formulated as finding the optimal rotation matrix R and translation vector t that minimize the Root Mean Square Error between the scaled predicted agent positions $O^{(i)}_{\text{pred,scaled}}$ and their corresponding real-world positions $O^{(i)}_{\text{real}}$:

$$\min_{R,t} \sum_{i=1}^{N_p} \|(R \cdot O^{(i)}_{\text{pred,scaled}} + t) - O^{(i)}_{\text{real}}\|^2 \tag{6}$$

subject to the constraints $R^T R = I$ and $\det(R) = 1$.

After the two steps, the relative point cloud P_{pred} from the prediction vision model is transformed into the real-world metric scale.

3.5. 3D bbox Labeling

After obtaining the 3D point cloud for each image subset, in this subsection, we mainly map the previously extracted 2D mask information onto the point cloud to obtain the 3D points group for each item detected. This point group then generate 3D oriented bounding boxes for the targets in the scene. We then aggregate and fuse multi-view results for the same target based on the spatial relationships among these 3D boxes, yielding a unified 3D annotation for all targets within the subset.

For data captured at the same timestamp from different agents' viewpoints, let C_i be the structured 3D point cloud from agent i, which is registered to its 2D image. To isolate a specific target j, we first use its 2D semantic mask $P_{mask,j}$. We then filter the full point cloud C_i using this mask to extract the valid point set $S_{i,j}$ for the target:

$$S_{i,j} = \{C_i(u, v) \mid (u, v) \in P_{mask,j}\}$$

This set $S_{i,j}$ contains all 3D points from viewpoint i that correspond to the target j. On this basis, we generate an oriented bounding box (OBB), $B_{i,j}$, for each target-specific point set $S_{i,j}$. Compared with axis-aligned bounding boxes (AABBs), OBBs account not only for an object's position and scale in 3D space but also accurately reflect its orientation and pose. Concretely, we apply Principal Component Analysis (PCA) to fit the geometric features of the target point cloud. By computing the eigenvectors and eigenvalues of the covariance matrix, we obtain the cloud's principal directions (i.e., the object's length, width, and height axes), from which we determine the OBB's center, orientation, and edge lengths. This procedure adaptively captures the object's spatial orientation and improves annotation accuracy for targets with complex poses.

Subsequently, we map each view's OBB back to a unified world coordinate system using the known camera extrinsics $E'_{pred,cam}$, and then aggregate multi-view results

Table 1. Model Accuracy Performance on Griffin Dataset

Model Name	Griffin-25m-100m[a]							
	Overall (0-100m)[b]				Long Scope (50-100m)[b]			
	AP ↑	ATE ↓	ASE ↓	AOE ↓	AP ↑	ATE ↓	ASE ↓	AOE ↓
V2X-ViT (ECCV 2022)	<u>0.201</u>	0.606	<u>0.164</u>	0.306	<u>0.011</u>	1.3	**0.3**	1.0
where2comm (NIPS 2022)	0.131	<u>0.597</u>	**0.160**	0.338	0.000	1.3	0.4	<u>0.9</u>
UniV2X (AAAI 2025)	0.160	0.700	0.200	<u>0.300</u>	0.007	1.3	0.4	1.1
CoSA-3D (ours)	**0.299**	**0.262**	0.306	**0.183**	**0.048**	**1.1**	0.8	**0.2**

[a]Model performance comparison on Griffin-25m datasets. Evaluate the targets range from 0 to 100 m.
[b]Bold values denote the best performance. Underlined values indicate the second-best performance.

from different agents. During fusion, we match and merge boxes based on the 3D Intersection over Union (IoU) computed on the world-space point clouds associated with each OBB: if the IoU of two OBBs exceeds a preset threshold, they are regarded as the same target instance and fused, yielding a globally consistent set of 3D bounding boxes for the scene at that timestamp.

Finally, for multi-frame data over a continuous time span from $t \rightarrow t+N$, we repeat the above annotation procedure on each frame and associate the same targets across frames by enforcing temporal continuity. Concretely, we compare IoU of OBBs in adjacent frames and the change in their centers to achieve consistent matching of the same target over time, producing 3D target tracks on the temporal axis.

4. Experiments

4.1. Datasets and Metrics

Griffin. This is a pioneering public large-scale dataset specifically designed for aerial-ground cooperative 3D perception [24]. Since CoSA-3D is a zero-shot, training-free offline annotation pipeline, it does not require any training data. We report our test results on two validation splits of *Griffin-25m* subset.

Evaluation Metrics. Both benchmarks adopt established metrics from the NuScenes benchmark [3] for 3D object detection and tracking, including Average Precision (AP) to assess annotation agreement and Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE) to assess labeling quality.

4.2. Quantitative Results

The labeling accuracy performance is shown in Table 1. We tested three other models [9, 29, 34] of cooperative vision-only 3D detection models on the Griffin dataset and compared the results with ours. Following the main task in most datasets and cooperative models [5, 24], we focus the metrics of the vehicles (including *Car, Bus and Truck*) in the Griffin dataset. The quantitative results of our proposed model, CoSA-3D, against state-of-the-art (SOTA) methods on the Griffin-25m-100m dataset. The evaluation is performed across two spatial ranges: overall (0-100m) and long-scope (50-100m). In the comprehensive 0-100m evaluation, CoSA-3D demonstrates superior performance by achieving the best results in three of the four metrics. This indicates a substantial enhancement in 3D object detection accuracy with our pipeline. Furthermore, CoSA-3D excels in localization, registering the lowest ATE (0.262) and AOE (0.183). This corresponds to a 56.1% reduction in translation error compared to the second-best ATE and a 39.0% reduction in orientation error relative to the next-best AOE.

The strength of our model is further highlighted in the challenging long-scope (50-100m) evaluation. CoSA-3D maintains its lead, securing the top performance in AP (0.048), ATE (1.1), and AOE (0.2). This robust performance at extended distances is critical in 3D vision labeling, effectively alleviating the problem of recognizing distant objects. Specifically, our model's AP is over 4.3 times higher than that of V2X-ViT (0.011), and its orientation error (0.2) is 77.8% lower than the second-best (0.9 from where2comm), underscoring its resilience to the sparse data challenges inherent in long-range perception.

To comprehensively evaluate the effectiveness of CoSA-3D as an annotation framework, we conducted an in-depth assessment specifically focused on the quality of the generated labels (see Table 2), in addition to the standard detec-

Table 2. Label Category Accuracy on Griffin Dataset

Object Category	IoU Threshold[a]	Precision (%)	Recall (%)
Car	0.3	81.4	93.6
	0.5	**88.1**	91.2
	0.8	82.8	**94.5**

[a]In the last projection and association part, filter and combine the boxes.

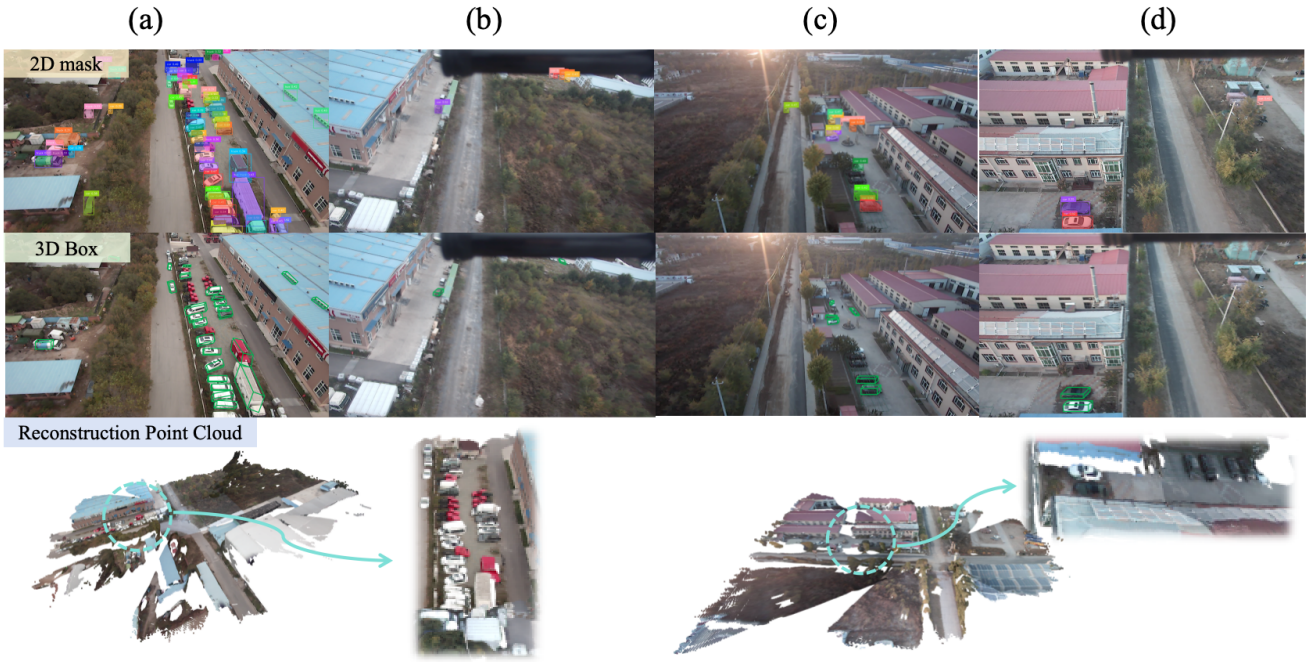

Figure 4. Qualitative results for the model performance on a real world cooperative dataset. They contain the 2D masks, final 3D box labeling and the 3D reconstruction point cloud for each image. (a) (c) show the drone's front view. (b) (d) show the drone's right view.

tion metrics. For dataset construction tasks, an extremely low miss rate (high recall) and high-confidence bounding boxes (high precision) are of paramount importance.

Discussion on Average Scale Error. While our framework achieves significant improvements in Average Precision and Average Translation Error, we observe a slight degradation in Average Scale Error compared to baseline methods. This limitation is largely inherent to the vision-only reconstruction pipeline. Although our geometric alignment module successfully recovers the global metric scale of the scene using inter-agent relative distances, estimating the precise local scale (i.e., exact length, width, and height) of individual objects remains challenging without explicit depth sensors or multi-modal priors. However, considering that CoSA-3D completely eliminates the reliance on expensive hardware and manual labor, this minor compromise in object-level scale is an acceptable trade-off for large-scale, low-cost dataset annotation.

4.3. Ablation Study

Effect of Geometric Fiducial Alignment. To rigorously quantify the impact of our proposed geometric fiducial alignment module, we compare our full model ("With correction") against a variant where this correction mechanism is deactivated ("Without correction"). The results shown in Table 3 unequivocally demonstrate that the extrinsic correction module is a critical component of our framework. When the correction module is removed, the model's performance collapses across nearly all key metrics.

Table 3. Influence of Geometric Fiducial Alignment

Configuration	AP (Overall) ↑	ATE ↓	ASE ↓	AOE ↓
Without correction	0.0712	1.4	0.8	1.0
With correction	0.299	0.262	0.606	0.183

Most notably, the overall Average Precision (AP) plummets from 0.299 to just 0.0712, a relative decrease of 76.2%. This signifies a catastrophic failure in the model's ability to accurately detect 3D objects. This degradation is mirrored in the localization accuracy, where the Average Translation Error (ATE) explodes by over 5.3 times, increasing from 0.262 to 1.4. The Average Scale Error (ASE) also deteriorates, worsening from 0.606 to 0.8.

This stark performance drop highlights that our proposed extrinsic correction is not an incremental optimization but a fundamental mechanism. It is essential for stabilizing the geometric perception and enabling the high-precision, robust 3D detection that our final model achieves.

Effect of Cooperative labeling. To investigate the individual contributions of each sensing agent, we evaluated the performance of three distinct configurations: labeling only from vehicle, labeling only from drone, and our full method.The results shown in Table 4 clearly illustrate the severe limitations of single-agent perception and the synergistic advantage of fusion. The "Vehicle-only" configuration yields a minimal overall AP of 0.063. This performance is substantially inferior to the "Drone-only" config-

Table 4. Ablation Study of Different Agents

Configuration	AP (Overall)
Fusion labeling	**0.299**
Vehicle-only labeling	0.063 (-0.236)
Drone-only labeling	0.153 (-0.146)

uration, which achieves an AP of 0.153. This suggests that the aerial perspective from the drone provides a more comprehensive and less occluded view for 3D object detection compared to the ground-level vehicle perspective alone.

This study confirms that both data streams are complementary and critical. The drone provides a broad, top-down context, while the vehicle provides high-resolution, ground-level details. Our proposed fusion mechanism is essential for effectively synthesizing these two disparate and individually incomplete viewpoints into a coherent and rich spatial representation, which is indispensable for achieving robust and accurate 3D labeling.

4.4. Qualitative Results

In addition to the quantitative evaluation on the Griffin benchmark, we conduct a generalization study to demonstrate our framework's practical applicability. The qualitative results are shown in Fig.4. We collected a real-world cooperative dataset, using one 4-rotor drone, lasting for 2 hours, to test our model's generalization performance. We present the 3D point cloud reconstruction and both 2D and 3D labeling from the drone's view. To provide a comprehensive view of the collaborative labeling without occlusion, our presentation focuses on the reconstructions from the drone's macro perspective. The model performed well on our real world dataset, showing robustness in different types of scenes.

5. Conclusion

In this work, we presented CoSA-3D, a LiDAR-free automatic annotation pipeline for collaborative 3D perception, significantly reducing the cost of labeling collaborative datasets. Our model combines 2D detection from 2D segmentation model and Structure-from-Motion (SfM) foundation model outputs to predict the depth of the targets. By applying inter-agent relative poses to recover a metric scale for the prediction point cloud, our model can label and track 3D bounding boxes across frames. Experiments show CoSA-3D achieves state-of-the-art detection and bounding box labeling on the Griffin dataset.

Acknowledgments

This work was supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM123) and the National Natural Science Foundation of China, Key Project (No. 52131201).

References

[1] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient Interactive Annotation of Segmentation Datasets with Polygon-RNN++. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 859–868, Salt Lake City, UT, 2018. IEEE. 1

[2] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. MUSt3R: Multi-view Network for Stereo 3D Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1050–1060, 2025. 3

[3] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, Seattle, WA, USA, 2020. IEEE. 6

[4] Yue-Jiang Dong, Yuan-Chen Guo, Ying-Tian Liu, Fang-Lue Zhang, and Song-Hai Zhang. PPEA-Depth: Progressive Parameter-Efficient Adaptation for Self-Supervised Monocular Depth Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2):1609–1617, 2024. 2

[5] Xiangbo Gao, Yuheng Wu, Fengze Yang, Xuewen Luo, Keshu Wu, Xinghao Chen, Yuping Wang, Chenxi Liu, Yang Zhou, and Zhengzhong Tu. AirV2X: Unified Air-Ground Vehicle-to-Everything Collaboration, 2025. arXiv:2506.19283 [cs]. 2, 6

[6] Ruiyang Hao, Siqi Fan, Yingru Dai, Zhenlin Zhang, Chenxi Li, Yuntian Wang, Haibao Yu, Wenxian Yang, Jirui Yuan, and Zaiqing Nie. RCooper: A Real-world Large-scale Dataset for Roadside Cooperative Perception. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22347–22357, Seattle, WA, USA, 2024. IEEE. 1, 2

[7] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge Univ. Press, Cambridge, 2. edition, 17.printing edition, 2018. 4

[8] Yunhao Hou, Bochao Zou, Min Zhang, Ran Chen, Shangdong Yang, Yanmei Zhang, Junbao Zhuo, Siheng Chen, Jiansheng Chen, and Huimin Ma. AGC-Drive: A Large-Scale Dataset for Real-World Aerial-Ground Collaboration in Driving Scenarios, 2025. arXiv:2506.16371 [cs]. 1, 2

[9] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-Efficient Collaborative Perception via Spatial Confidence Maps. *Advances in Neural Information Processing Systems*, 35:4874–4886, 2022. 6

[10] Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. DEFOM-Stereo: Depth Foundation Model Based Stereo Matching. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21857–21867, 2025. 3

[11] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976. Publisher: International Union of Crystallography. 5

[12] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment Anything in High Quality. In *Advances in Neural Information Processing Systems*, pages 29914–29934. Curran Associates, Inc., 2023. 2

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[14] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2X-Sim: Multi-Agent Collaborative Perception Dataset and Benchmark for Autonomous Driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022. 1, 2

[15] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation. pages 3159–3167, 2016. 1

[16] Genjia Liu, Yue Hu, Chenxin Xu, Weibo Mao, Junhao Ge, Zhengxiang Huang, Yifan Lu, Yinda Xu, Junkai Xia, Yafei Wang, and Siheng Chen. Toward Collaborative Autonomous Driving: Simulation Platform and End-to-End System. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6566–6584, 2025. 2

[17] Tao Ma, Hongbin Zhou, Qiusheng Huang, Xuemeng Yang, Jianfei Guo, Bo Zhang, Min Dou, Yu Qiao, Botian Shi, and Hongsheng Li. ZOPP: A Framework of Zero-shot Offboard Panoptic Perception for Autonomous Driving. In *Advances in Neural Information Processing Systems*, pages 140266–140291. Curran Associates, Inc., 2024. 2

[18] Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R. Qi, Xinchen Yan, Scott Ettinger, and Dragomir Anguelov. Unsupervised 3D Perception with 2D Vision-Language Distillation for Autonomous Driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8568–8578, Paris, France, 2023. IEEE.

[19] Charles R. Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa Vo, Boyang Deng, and Dragomir Anguelov. Offboard 3D Object Detection from Point Cloud Sequences. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6130–6140, Nashville, TN, USA, 2021. IEEE.

[20] Xiaoyan Qian, Chang Liu, Xiaojuan Qi, Siew-Chong Tan, Edmund Lam, and Ngai Wong. Context-Aware Transformer for 3D Point Cloud Automatic Annotation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2082–2090, 2023. 2

[21] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos, 2024. arXiv:2408.00714 [cs]. 2

[22] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks, 2024. arXiv:2401.14159 [cs]. 2, 3

[23] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle Adjustment — A Modern Synthesis. In *Vision Algorithms: Theory and Practice*, pages 298–372. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000. Series Title: Lecture Notes in Computer Science. 4

[24] Jiahao Wang, Xiangyu Cao, Jiaru Zhong, Yuner Zhang, Haibao Yu, Lei He, and Shaobing Xu. Griffin: Aerial-Ground Cooperative Detection and Tracking Dataset and Benchmark, 2025. arXiv:2503.06983 [cs]. 1, 2, 6

[25] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5294–5306, 2025. 3

[26] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3

[27] Yuchao Wang, Zhirui Wang, Peirui Cheng, Pengju Tian, Ziyang Yuan, Jing Tian, Wensheng Wang, and Liangjin Zhao. AVCPNet: An AAV-Vehicle Collaborative Perception Network for 3-D Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–16, 2025. 2

[28] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. pi^3: Permutation-Equivariant Visual Geometry Learning, 2025. arXiv:2507.13347 [cs]. 3

[29] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer. In *Computer Vision – ECCV 2022*, pages 107–124. Springer Nature Switzerland, Cham, 2022. Series Title: Lecture Notes in Computer Science. 6

[30] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. OPV2V: An Open Benchmark Dataset and Fusion Pipeline for Perception with Vehicle-to-Vehicle Communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589, 2022. 1, 2

[31] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, Hongkai Yu, Bolei Zhou, and Jiaqi Ma. V2V4Real: A Real-World Large-Scale Dataset for Vehicle-to-Vehicle Cooperative Perception. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13712–13722, Vancouver, BC, Canada, 2023. IEEE. 2

[32] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10371–10381, Seattle, WA, USA, 2024. IEEE. 2

[33] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards Zero-shot Metric 3D Prediction from A Single Image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 2

[34] Haibao Yu, Wenxian Yang, Jiaru Zhong, Zhenwei Yang, Siqi Fan, Ping Luo, and Zaiqing Nie. End to End Autonomous Driving Through V2X Cooperation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (9):9598–9606, 2025. 6

[35] Sergey Zakharov, Wadim Kehl, Arjun Bhargava, and Adrien Gaidon. Autolabeling 3D Objects With Differentiable Rendering of SDF Shape Priors. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12221–12230, Seattle, WA, USA, 2020. IEEE. 2

[36] Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C. Knoll. TUMTraf V2X Cooperative Perception Dataset. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22668–22677, Seattle, WA, USA, 2024. IEEE. 1, 2

[37] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment Everything Everywhere All at Once. In *Advances in Neural Information Processing Systems*, pages 19769–19782. Curran Associates, Inc., 2023. 2