TALoS: Enhancing Semantic Scene Completion via Test-time Adaptation on the Line of Sight

Hyun-Kurl Jang*
Visual Intelligence Lab.
KAIST
jhg0001@kaist.ac.kr

Jihun Kim*
Visual Intelligence Lab.
KAIST
jihun1998@kaist.ac.kr

Hyeokjun Kweon*
Visual Intelligence Lab.
KAIST
0327june@kaist.ac.kr

Kuk-Jin Yoon
Visual Intelligence Lab.
KAIST
kjyoon@kaist.ac.kr

Abstract

Semantic Scene Completion (SSC) aims to perform geometric completion and semantic segmentation simultaneously. Despite the promising results achieved by existing studies, the inherently ill-posed nature of the task presents significant challenges in diverse driving scenarios. This paper introduces TALoS, a novel test-time adaptation approach for SSC that excavates the information available in driving environments. Specifically, we focus on that observations made at a certain moment can serve as Ground Truth (GT) for scene completion at another moment. Given the characteristics of the LiDAR sensor, an observation of an object at a certain location confirms both 1) the occupation of that location and 2) the absence of obstacles along the line of sight from the LiDAR to that point. TALoS utilizes these observations to obtain self-supervision about occupancy and emptiness, guiding the model to adapt to the scene in test time. In a similar manner, we aggregate reliable SSC predictions among multiple moments and leverage them as semantic pseudo-GT for adaptation. Further, to leverage future observations that are not accessible at the current time, we present a dual optimization scheme using the model in which the update is delayed until the future observation is available. Evaluations on the SemanticKITTI validation and test sets demonstrate that TALoS significantly improves the performance of the pre-trained SSC model. Our code is available at https://github.com/blue-531/TALoS.

1 Introduction

LiDAR is a predominant 3D sensor in autonomous vehicles, effectively capturing the 3D geometry of the surroundings as a point cloud. However, LiDAR inherently records the surface of objects, leaving the areas behind initial contact points empty. Therefore, it is crucial for safety and driving planning to predict the state of these hidden regions using only the limited information available. Addressing these challenges, Semantic Scene Completion (SSC) has emerged as a pivotal research topic, enabling simultaneous geometric completion and semantic segmentation of the surroundings.

Existing SSC studies have focused on tackling both tasks [1, 2, 3, 4, 5, 6, 7, 8, 9], mainly from an architectural perspective. By amalgamating the models specialized for each task, these approaches have shown promising results over the last few years. Nevertheless, the nature of the completion

*denotes equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

task—filling in the unseen parts from the given observation—heavily relies on the prior structural distribution learned from the training dataset. Therefore, in our view, the classical SSC paradigm is inevitably vulnerable to handling the diverse scene structures encountered in driving scenarios.

As a remedy, this paper pioneers a novel SSC approach based on Test-time Adaptation (TTA), which adjusts a pre-trained model to adapt to each test environment. Due to the absence of ground truths (GTs) during test time, the existing TTA studies for various fields have endeavored to design optimization goals, like meta-learning or auxiliary tasks [10, 11, 12, 13, 14]. Instead, we focus on the driving scenarios assumed by SSC, excavating the information helpful for adapting the model to the scene in test time.

Our main idea is simple yet effective: **an observation made at one moment could serve as supervision for the SSC prediction at another moment**. While traveling through an environment, an autonomous vehicle can continuously observe the overall scene structures, including objects that were previously occluded (or will be occluded later), which are concrete guidances for the adaptation of scene completion. Given the characteristics of the LiDAR sensor, an observation of a point at a specific spatial location at a specific moment confirms not only the occupation at that location itself but also the absence of obstacles along the line of sight from the sensor to that location.

The proposed method, named **Test-time Adaptation via Line of Sight (TALoS)**, is designed to explicitly leverage these characteristics, obtaining self-supervision for geometric completion. Additionally, we extend the TALoS framework for semantic recognition, another key goal of SSC, by collecting the reliable regions only among the semantic segmentation results predicted at each moment. Further, to leverage valuable future information that is not accessible at the time of the current update, we devise a novel dual optimization scheme involving the model gradually updating across the temporal dimension. This enables the model to continuously adapt to the surroundings at test time without any manual guidance, ultimately achieving better SSC performance.

We verify the superiority of TALoS on the SemanticKITTI [15] benchmark. The results strongly confirm that TALoS enhances not only geometric completion but also semantic segmentation performance by large margins. With extensive experiments, including ablation studies, we analyze the working logic of TALoS in detail and present its potential as a viable solution for practical SSC.

2 Related Works

2.1 Semantic scene completion

Starting from SSCNet [1], semantic scene completion task has been extensively studied [16, 2, 3, 4, 5, 6, 7, 8, 9, 17, 18, 19, 20, 21, 22, 23]. In SSC using LiDAR, as both tasks of geometric completion and semantic understanding should be achieved simultaneously, the existing studies have mainly presented architectural approaches. For example, LMSC [4] and UtD [8] utilize UNet-based structures with multi-scale connections. JS3C-Net [5] and SSA-SC [6] propose architectures consisting of semantic segmentation and completion networks to utilize them complementarily. Although these approaches show promise, handling the diversities inherent in outdoor scenes remains a challenging problem. In this light, we would like to introduce test-time adaptation to the field of semantic scene completion.

Notably, SCPNet [3] proposes to use distillation during the training phase, transferring the knowledge from the model using multiple scans to the model using a single scan. Although this approach also aims to use information from various moments, the proposed TALoS is distinct as it leverages such information online, adapting the model to the diverse driving sequence. Also, the recently published OccFiner [24] is noteworthy, as it aims to enhance the already existing SSC model. However, OccFiner is a post-processing method that refines the results of the pre-trained model, performing in an offline manner, unlike our online TTA-based approach.

2.2 Test-time adaptation

Test-time Adaptation (TTA) aims to adapt a pre-trained model to target data in test time, without access to the source domain data used for training. One widely used method involves attaching additional self-supervised learning branches to the model [10, 11, 12]. In point cloud TTAs, using auxiliary tasks such as point cloud reconstruction [13, 14] are actively studied. However, these

approaches require the model to be trained with the additional branches, primarily in the training stage on the source dataset.

To relieve the requirements on the training stage, various online optimization goals have been explored, such as information maximization [25, 26, 27] and pseudo labeling [26, 28, 29, 30] schemes. Similar approaches have also been proposed for the point cloud, as in [31, 32], using pseudo labeling.

Unfortunately, despite the natural fit between these TTA approaches and the goal of SSC, which involves completing diverse driving environments, the use of TTA has been scarcely explored in the SSC field. Against this background, we pioneer the TTA-based SSC method, especially focusing on excavating the information from the point clouds consecutively observed at various moments.

3 Method

3.1 Problem definition

This section begins by defining the formulation of our approach and the notations used throughout the paper. The Semantic Scene Completion (SSC) task aims to learn a mapping function from an input point cloud to the completed voxel representation. We formally denote the input point cloud $\mathbf{X} \in \mathbb{R}^{N\times 3}$ as a set of points, where each point represents its XYZ coordinate. Following the conventional SSC studies, the completion result is denoted as $\mathbf{Y} \in \mathbb{C}^{L\times W\times H}$. Here, L, W, H are the dimensions of the voxel grid, and $\mathbb{C} = \{0, 1, \ldots, C\}$ is a set of class indices indicating whether a voxel is empty (0) or belongs to a specific class $(1, \ldots, C)$.

As our approach is based on TTA, we assume the existence of a pre-trained SSC model $\mathcal{F}$ as follows:

$$\mathbf{p} = \mathcal{F}(\mathbf{X}), \tag{1}$$

where $\mathbf{p} \in [0, 1]^{(C+1)\times L\times W\times H}$ is the probability of each voxel belonging to each class. The final class prediction $\hat{\mathbf{Y}}$ can be obtained by applying an argmax function on $\mathbf{p}$. In this context, the goal of TALoS is to adjust the pre-trained model to adapt to an arbitrary test sample $\mathbf{X}$ by optimizing the parameters, making them more suitable. Here, note that the proposed approach does not have explicit requirements on $\mathcal{F}$, such as the architectures or pre-training policies.

3.2 Test-time Adaptation via Line of Sight (TALoS)

The proposed TALoS targets a realistic application of the pre-trained SSC model, assuming an autonomous vehicle drives through arbitrary environments in test time. In this scenario, we suppose the point clouds captured by LiDAR are continuously provided as time proceeds, and accordingly, the model should perform SSC for each given point cloud instantly. We denote the input sequence of point clouds as $\{\mathbf{X}_i\}$, where $i = \{1, \ldots\}$ indicates the moment when the point cloud is captured.

The main idea behind TALoS is that for guiding the model prediction of a certain moment (let i), the observation made at another moment (let j) can serve as supervision. However, as the ego-vehicle moves as time proceeds, the input point clouds captured at the two different moments, *i.e.*, $\mathbf{X}_i$ and $\mathbf{X}_j$, are on different LiDAR coordinates. To handle this, we use a transformation matrix $\mathbf{T}_{j\to i}$ between two coordinates to transform the jth point cloud $\mathbf{X}_j$ with respect to the ith coordinate system, by

$$\mathbf{X}_{j\to i} = \mathbf{T}_{j\to i}\mathbf{X}_j, \tag{2}$$

where $\mathbf{X}_{j\to i}$ is the transformed point cloud.

Subsequently, we exploit $\mathbf{X}_{j\to i}$ to obtain a binary self-supervision for geometric completion, indicating whether a voxel is empty or occupied. We implement this process as shown in Fig. 1. For voxelization, we first define a voxel grid of a pre-defined size ($L \times W \times H$), initialized with an ignore index (255). Then, using $\mathbf{X}_{j\to i}$, we set the value of voxels containing at least one point to 1 (green color in Fig. 1), while keeping the other voxels as 255. Here, note that we discard the points of the non-static classes (*e.g.*, car), as such object can change their location between ith and jth observation. For this rejection, we use $\mathbf{p}_j$, the model prediction at jth moment. The resulting binary mask indicates which voxels are occupied at the jth moment with respect to the ith coordinate.

Additionally, we use the Line of Sight (LoS), an idea utilized in various fields [33, 34, 35], to further identify which voxels should not be occupied. Considering LiDAR's characteristics, the space

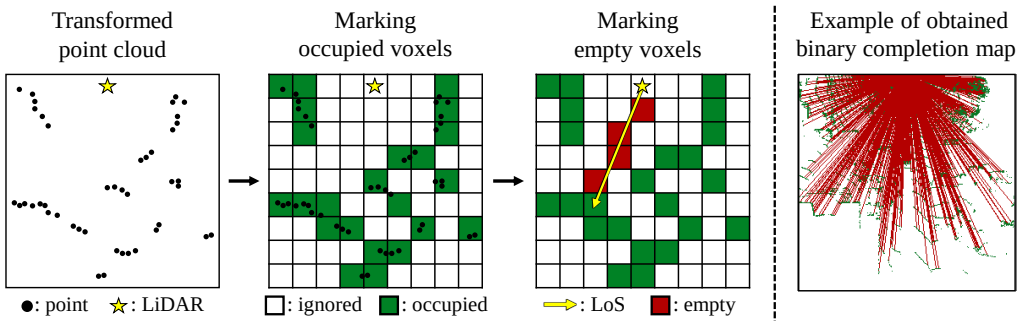

Figure 1: **Left**: Visualization of constructing a binary map $\mathbf{V}_{j\to i}^{comp}$ from the transformed point cloud $\mathbf{X}_{j\to i}$. Although we represent our process using a 2D grid for intuitive visualization, note that the real process is performed on a 3D voxel. **Right**: The real example of the binary map projected on 2D.

between the LiDAR and the voxels filled with 1 should be empty. To check which voxels are crossed by the LoS, we employ Bresenham's algorithm [36]. For this process, we use the LiDAR position at the jth moment converted to the ith coordinate system, not the ith LiDAR position. Finally, we set the value of the identified voxels to 0 (red color in Fig. 1), indicating that the voxels should be empty.

The obtained $\mathbf{V}_{j\to i}^{comp} \in \{0, 1, 255\}^{L\times W\times H}$ then serves as supervision for $\mathbf{p}_i$, the ith prediction. Here, $\mathbf{p}_i$ is obtained by the pre-trained SSC model $\mathcal{F}$ as follows:

$$\mathbf{p}_i = \mathcal{F}(\mathbf{X}_i). \tag{3}$$

Since $\mathbf{p}_i$ is the prediction for all the classes, including the empty class, we convert it into the binary prediction for completion, denoted as $\mathbf{p}_i^{comp} \in [0, 1]^{2\times L\times W\times H}$. Here, the first element of $\mathbf{p}_i^{comp}$ is simply $\mathbf{p}_i^0$, and the second one is $\max_{c=1,\dots,C} \mathbf{p}_i^c$, the maximum value among the scores of the non-empty classes. Here, $\mathbf{p}_i^c$ represents the predicted probability of the voxels belonging to cth class at ith moment.

Finally, the loss function using the binary completion map is as

$$\mathcal{L}_{j\to i}^{comp}(\mathbf{p}_i) = \mathcal{L}_{ce}(\mathbf{p}_i^{comp}, \mathbf{V}_{j\to i}^{comp}) + \mathcal{L}_{lovasz}(\mathbf{p}_i^{comp}, \mathbf{V}_{j\to i}^{comp}), \tag{4}$$

where $\mathcal{L}_{ce}$ and $\mathcal{L}_{lovasz}$ are cross-entropy loss and lovasz-softmax loss [37], respectively.

3.3 Extension for semantic perception

In the previous section, we described a method that leverages an observation made in one moment (j) to obtain the binary occupancy map of another moment (i). As the method mainly focuses on enhancing the model's capability of understanding the test-time scene structure, *i.e.*, scene completion, this section further extends our approach to address the semantic perception of surroundings, another key goal of SSC. Specifically, we carefully identify the reliable regions from the prediction of the pre-trained model at each moment, and then build a consensus among these predictions across various moments.

To achieve this, we first define a metric similar to [25, 26, 27], based on Shannon entropy [38] $\mathcal{H}$, as follows:

$$\mathbf{R} = \mathcal{H}(\mathbf{p}) = -\sum_{c=0}^{C} \mathbf{p}^c \log \mathbf{p}^c, \tag{5}$$

where $\mathbf{R} \in [0, 1]^{L\times W\times H}$ is the measured reliability. As reported in conventional works based on confidence-based self-supervision, we also confirmed a high positive correlation between the reliability $\mathbf{R}$ and the actual accuracy of voxel-wise classification, as in Fig. 2 left.

To identify the confident regions from the model predictions, we simply threshold the measured reliability by a pre-defined value τ (0.75 in ours) as follows:

$$\mathbf{A}(x, y, z) = \begin{cases} \text{argmax}_c\, \mathbf{p}^c(x, y, z) & \text{if } \mathbf{R}(x, y, z) > \tau \\ 255 & \text{otherwise,} \end{cases} \tag{6}$$

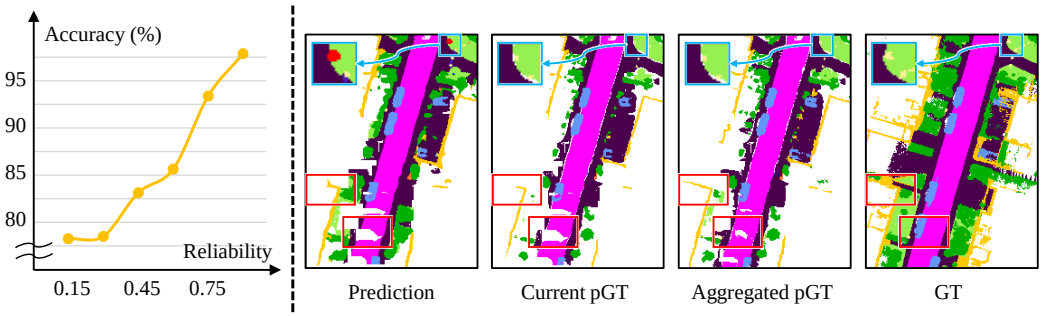

Figure 2: **Left**: Verification of the reliability metric. The voxels having higher reliability show higher semantic completion accuracy. **Right**: Examples of pseudo-GT (pGT) construction. The blue box depicts the successful rejection of misprediction using reliability, while the red boxes show the benefit of using the prediction of another moment, providing more completed pGT.

where (x, y, z) denotes the voxel coordinate. Here, $\mathbf{A} \in (\mathbb{C} \cup \{255\})^{L \times W \times H}$ is the reliable self-supervision, which can function as pseudo-GT for semantic segmentation.

Following the above process, we first acquire $\mathbf{A}_j$ using jth model prediction. Subsequently, we project its coordinate to ith coordinate, similar to Equ. (2). After obtaining the projected pseudo-GT, denoted as $\mathbf{A}_{j \to i}$, we aggregate it with $\mathbf{A}_i$, the pseudo-GT obtained at the current moment i. In detail, we replace only the unconfident voxels in $\mathbf{A}_i$ (indexed as 255) with the corresponding voxels of $\mathbf{A}_{j \to i}$. Meanwhile, if the classes predicted by $\mathbf{A}_i$ and $\mathbf{A}_{j \to i}$ differ on certain voxels and both are confidently predicted, we conservatively drop those voxels as 255. As depicted in the colored boxes in Fig. 2 right, this aggregation helps our framework to build a consensus among the semantic perceptions performed at various moments, significantly enhancing the quality of pseudo-GT from the perspective of SSC.

We denote the result of aggregation as $\mathbf{V}^{sem}_{j \to i} \in (\mathbb{C} \cup \{255\})^{L \times W \times H}$. Since $\mathbf{V}^{sem}_{j \to i}$ contains semantic information about all the classes, we can utilize it as direct pseudo-GT for guiding $\mathbf{p}_i$. Accordingly, the loss function is defined as:

$$\mathcal{L}^{sem}_{j \to i}(\mathbf{p}_i) = \mathcal{L}_{ce}(\mathbf{p}_i, \mathbf{V}^{sem}_{j \to i}) + \mathcal{L}_{lovasz}(\mathbf{p}_i, \mathbf{V}^{sem}_{j \to i}), \tag{7}$$

where the notations are similar to those of Equ. (4).

3.4 Dual optimization scheme for gradual adaptation

The previous sections introduced how TALoS guides the prediction of a pre-trained SSC model at a certain moment (i), leveraging observations made at another moment (j). From a methodological perspective, the remaining step is to consider how to effectively adapt the model by selecting the appropriate moments.

Essentially, we cannot observe the future. Therefore, assuming TTA in real driving scenarios, we can only use past observations when updating the model at the current moment, which implies $i > j$. However, from the perspective of the SSC task, the main region of interest is intuitively the forward driving direction of the autonomous vehicle. This implies that guidance from future observations can be more important and valuable than guidance from past observations.

So, how can we leverage future information without actually observing it at the current moment? Our key idea is to hold the model and its prediction at the current moment without updating and **delay the update until future observations become available**.

We develop this idea as in Fig. 3. In detail, our approach involves two models of $\mathcal{F}^M$ and $\mathcal{F}^G$. The goal of $\mathcal{F}^M$ is an instant adaptation on the sample of the current moment. Therefore, we initialize $\mathcal{F}^M$ every moment with the pre-trained model and discard it at the end of the moment. Here, $\mathcal{F}^M$ can be instantly updated at the moment i, using the past information already observed at jth moment. The loss function for $\mathcal{F}^M$ is defined as:

$$\mathcal{L}^{M}_{j \to i} = \mathcal{L}^{comp}_{j \to i}(\mathbf{p}^M_i) + \mathcal{L}^{sem}_{j \to i}(\mathbf{p}^M_i), \tag{8}$$

where $\mathbf{p}^M_i$ is the output of $\mathcal{F}^M$ at ith moment.

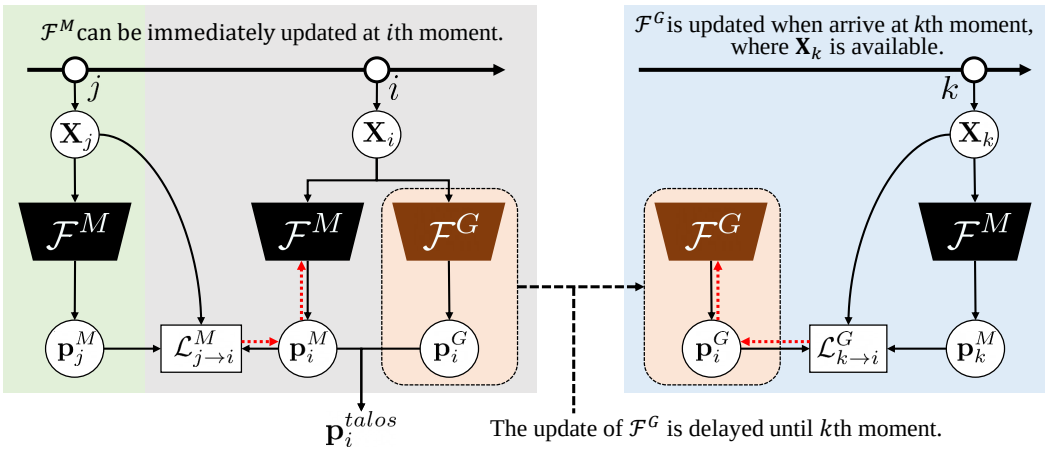

Figure 3: Conceptual visualization of the dual optimization scheme. $\mathcal{F}^M$ is instantly updated at moment i, using the past information provided from jth moment. On the other hand, the update of $\mathcal{F}^G$ using ith prediction is delayed until kth moment, when the future information becomes available. We unify the predictions of the models, $\mathbf{p}_i^M$ and $\mathbf{p}_i^G$, to get the final prediction $\mathbf{p}_i^{talos}$. The red dashed line denotes the back-propagation.

However, as aforementioned, we want to also leverage future information at kth moment, which is not available yet. For this, we define $\mathcal{F}^G$, which aims to gradually learn the overall scene distribution. Therefore, $\mathcal{F}^G$ is initialized only at the first step and continuously used for prediction. As in Fig. 3, the inference of $\mathcal{F}^G$ for $\mathbf{X}_i$ is instantly done and is used for the final output of the ith moment. On the other hand, the update of $\mathcal{F}^G$ should stand by, until $\mathbf{X}_k$ is available. In other words, once the model $\mathcal{F}^G$ arrives at kth moment (which was future at the time of prediction), the update is performed. The loss function for $\mathcal{F}^G$ is defined as:

$$\mathcal{L}_{k\to i}^G = \mathcal{L}_{k\to i}^{comp}(\mathbf{p}_i^G) + \mathcal{L}_{k\to i}^{sem}(\mathbf{p}_i^G), \tag{9}$$

where $\mathbf{p}_i^G$ is the output of $\mathcal{F}^G$ at ith moment.

This update cannot directly affect the prediction of $\mathcal{F}^G$ made at ith moment, as it is already over in the past from the perspective of kth moment when the update occurred. However, this continuous accumulation of future information gradually enhances the model, allowing it to better learn the overall scene structure as time progresses. We provide a detailed illustration in Algorithm 1.

In summary, the proposed TALoS framework involves two models, moment-wisely adapted $\mathcal{F}^M$ and gradually adapted $\mathcal{F}^G$. To obtain the final prediction $\mathbf{p}_i^{talos}$ of ith moment, we individually run $\mathcal{F}^M$ and $\mathcal{F}^G$ using $\mathbf{X}_i$ as an input. Both results $\mathbf{p}_i^M$ and $\mathbf{p}_i^G$ are unified into a single voxel prediction. Here, we use $\mathbf{p}_i^M$ as a base, while trusting $\mathbf{p}_i^G$ only for the voxels predicted as static categories (such as roads or buildings) by $\mathbf{p}_i^G$. The rationale behind this strategy is that continual adaptation makes $\mathcal{F}^G$ gradually adapt to the overall sequence, leading to a better understanding of the distribution of static objects. We empirically found that the continuous adaptation is more facilitated for the static pattern than the movable objects having diverse distribution.

4 Experiments

4.1 Settings

Datasets & Metrics. We primarily experiment on SemanticKITTI [15], the standard benchmark for SSC, comprising 22 LiDAR sequences. Sequences 00 to 10 are used for pre-training SSC models, except for 08, which is employed as a validation set. For testing, we use sequences 11 to 21. Additionally, we verify TALoS on cross-dataset evaluation from SemanticKITTI to SemanticPOSS [39]. From the 6 sequences of SemanticPOSS, we utilize only the validation sequence (02). For more details, please refer to the supplementary material. For evaluation, we employ intersection over union (IoU), a standard metric for semantic segmentation. We report both the completion IoU (cIoU) for binary occupancy prediction and the mean IoU (mIoU) for all classes.

Algorithm 1 Dual optimization scheme (single iteration)

Input: Moment-model $\mathcal{F}^M$, Gradual-model $\mathcal{F}^G$, Pre-trained SSC model parameters θ_0, temporal distance τ, and Buffer B
Output: Adapted gradual-model parameters θ^G
Initialize gradual-model parameters: $\theta^G \leftarrow \theta_0$
for each timestep $t = 1$ to unknown T **do**
  Initialize moment-model parameters: $\theta^M \leftarrow \theta_0$
  $i, j \leftarrow t, t - \tau$
  Receive current LiDAR observation $\mathbf{X}_i$
  Perform prediction with the moment-model: $\mathbf{p}_i^M \leftarrow \mathcal{F}^M(\mathbf{X}_i; \theta^M)$
  Perform prediction with the gradual-model: $\mathbf{p}_i^G \leftarrow \mathcal{F}^G(\mathbf{X}_i; \theta_i^G)$
  Save $\mathbf{p}_i^M$ in the buffer B.
  Save $\mathbf{p}_i^G$ and the corresponding forward propagation graphs of $\mathcal{F}^G$ to the buffer B.
  if $j >= 1$ **then**
    Load $\mathbf{p}_j^M$ from the buffer B.
    Compute $\mathcal{L}_{j \to i}^M$ using $\mathbf{p}_i^M$ and $\mathbf{p}_j^M$
    Update moment-model parameters: $\theta^M \leftarrow \theta^M - \eta \nabla_{\theta^M} \mathcal{L}_{j \to i}^M$
    Load $\mathbf{p}_j^G$ and the corresponding forward propagation graphs of $\mathcal{F}^G$ from the buffer B.
    Compute $\mathcal{L}_{i \to j}^G$ using $\mathbf{p}_i^G$ and $\mathbf{p}_j^G$
    Update moment-model parameters: $\theta^G \leftarrow \theta^G - \eta \nabla_{\theta^G} \mathcal{L}_{i \to j}^G$
    Perform prediction with the updated moment-model: $\mathbf{p}_i^M \leftarrow \mathcal{F}^M(\mathbf{X}_i; \theta^M)$
    Perform prediction with the updated gradual-model: $\mathbf{p}_i^G \leftarrow \mathcal{F}^G(\mathbf{X}_i; \theta_i^G)$
  end if
  $\mathbf{p}_i^{talos} \leftarrow \text{Agg}(\mathbf{p}_i^M, \mathbf{p}_i^G)$
  Return $\mathbf{p}_i^{talos}$ as the final SSC result of the current timestep i
end for
Return adapted gradual-model parameters θ^G

Implementation. We employ the officially provided SCPNet [3], which is pre-trained on the SemanticKITTI [15] train set, as our baseline. To apply TALoS in test time, we only update the last few layers of SCPNet. Additionally, to prevent SCPNet's architecture from automatically making the voxels far from existing points empty, we use a 3D convolution layer to expand the region of sparse tensor computation. This ensures that distant voxels are properly involved in the test-time adaptation. For optimization, we use Adam optimizers [40], where the learning rates are set to 3e-4 and 3e-5 for $\mathcal{F}^M$ and $\mathcal{F}^G$, respectively. For more details, refer to Section A.

4.2 Ablation studies

We conducted ablation studies to evaluate the effectiveness of each component of TALoS. The configurations and results of the ablation studies are demonstrated in Table 1. First, we verify the effectiveness of the loss functions we devised. The result of Exp A confirms that minimizing $\mathcal{L}^{comp}$ indeed increases both cIoU and mIoU performance over the baseline, helping the model adapt to each test sample. In addition, Exp B shows that using $\mathcal{L}^{sem}$ is also effective, especially for semantic perception, resulting in better mIoU performance. Further, Exp C using both losses achieves even higher performance, demonstrating the effectiveness of the proposed TALoS.

Additionally, we check the validity of the dual optimization scheme, ablating either $\mathcal{F}^M$ or $\mathcal{F}^G$. The results of Exp C and Exp D show that the use of moment-wise adaptation and gradual adaptation are both effective. Further, Exp E confirms that our dual scheme effectively unifies both gains into our TALoS framework, significantly outperforming the baseline on both cIoU and mIoU.

4.3 Results on SemanticKITTI

Table 2 provides the performance of existing SSC methods on the SemanticKITTI test set. Compared with SCPNet, which serves as our baseline, the proposed TALoS achieves significantly higher performance on both cIoU and mIoU. Considering that SCPNet also involves knowledge distillation

Table 1: The results of ablation studies for the proposed TALoS framework, conducted on SemanticKITTI val set. COMP and SEM denote the use of loss function defined in Equ. (4) and Equ. (7), respectively. Meanwhile, MOMENT and GRADUAL represent the use of $\mathcal{F}^M$ and $\mathcal{F}^G$ for the optimization scheme in Sec. 3.4, respectively. All metrics are in %. Best results are in **bold**.

	Loss functions		Dual optimization scheme		Metrics	
	COMP	SEM	MOMENT	GRADUAL	mIoU	cIoU
Baseline					37.56	50.24
A	✓		✓		37.97	52.81
B		✓	✓		38.35	52.47
C	✓	✓	✓		38.38	52.95
D	✓	✓		✓	38.81	55.94
E (Ours)	✓	✓	✓	✓	**39.29**	**56.09**

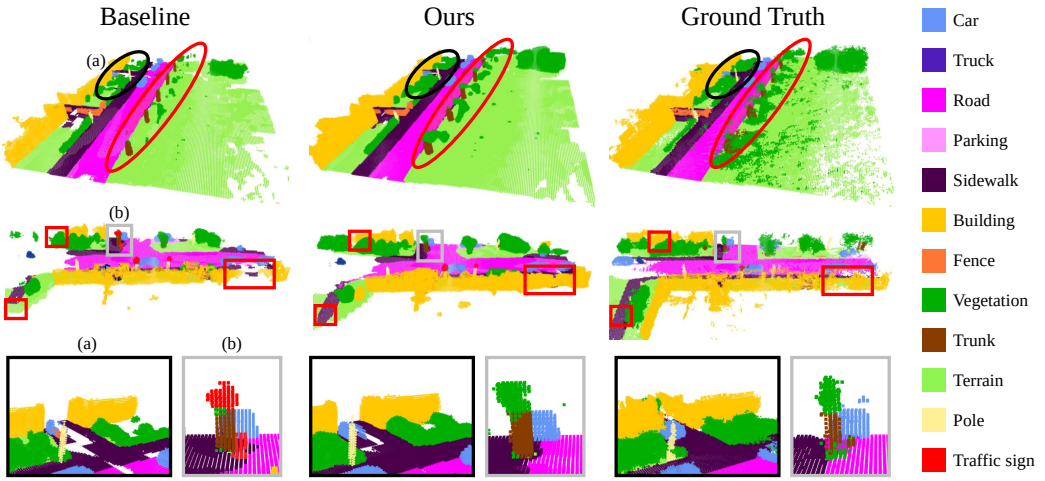

Figure 4: Qualitative comparisons between baseline (SCPNet) and ours TALoS on SemanticKITTI val set. The highlighted regions depict the improvements achieved by TALoS, better completing the scene while also recovering the mispredictions.

using future frames during training, this performance gain confirms that our TTA-based method is more effective for leveraging future information. Figure 4 provides a qualitative comparison between the baseline and ours, demonstrating the advantages of TALoS.

Additionally, it is noteworthy to mention the difference between ours and OccFiner [24]. OccFiner is designed to refine the results of existing SSC methods in an offline manner. It first generates predictions for a LiDAR sequence using an SSC method and then fuses these predictions post-driving to refine the results. In contrast, TALoS aims to perform test-time adaptation instantly in an online manner. We assume the sequential sensing of LiDAR data during driving, and TALoS gradually enhances the model as the test-time adaptation progresses. As both methods have advantages in their respective practical settings, we simply mention it here, rather than comparing them in Table 2.

4.4 Comparisons with the existing TTA methods

To demonstrate the benefit of our approach from the perspective of TTA, we integrated existing TTA studies into our framework and tested them. The results can be found in Table 3. First, we performed optimization via entropy minimization, as in TENT [25], instead of minimizing the proposed loss functions. For this experiment, we exclusively use $\mathcal{F}^M$ for clear comparison. This setting achieved 37.92% and 49.86% in mIoU and cIoU, respectively. Note that cIoU of this setting is slightly lower than that of the baseline. Further, the setting of Exp C in Table 1, which also uses $\mathcal{F}^M$ only, still outperforms TENT in both metrics. These highlight the effectiveness of our losses, which utilize observations made at various moments.

To further clarify the superiority of our dual optimization scheme, we also implemented CoTTA [28], a continuous TTA approach based on the widely used student-teacher scheme. We update both $\mathcal{F}^M$

Table 2: Quantitative comparison between the existing SSC methods with our TALoS on SemanticKITTI [15] test set. We use an online benchmark server for evaluation.

Methods	mIoU	cIoU	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
SSA-SC [6]	23.5	58.8	36.5	13.9	4.6	5.7	7.4	4.4	2.6	0.7	72.2	37.4	43.7	10.9	43.6	30.7	43.5	25.6	41.8	14.5	6.9
JS3C-Net [5]	23.8	56.6	33.3	14.4	8.8	7.2	12.7	8.0	5.1	0.4	64.7	34.9	39.9	14.1	39.4	30.4	43.1	19.6	40.5	18.9	15.9
S3CNet [16]	29.5	45.6	31.2	**41.5**	**45.0**	6.7	16.1	**45.9**	**35.8**	**16.0**	42.0	17.0	22.5	7.9	**52.2**	31.3	39.5	34.0	21.2	31.0	24.3
SCPNet [3]	36.7	56.1	**46.4**	33.2	34.9	13.8	29.1	28.2	24.7	1.8	68.5	**51.3**	49.8	**30.7**	38.8	44.7	46.4	**40.1**	48.7	40.4	**25.1**
TALoS	**37.9**	**60.2**	**46.4**	34.4	36.9	**14.0**	**30.0**	30.5	27.3	2.2	**73.0**	**51.3**	**53.6**	28.4	40.8	**45.1**	**50.6**	38.8	**51.0**	**40.7**	24.4

Table 3: Comparisons between the proposed TALoS and the existing TTA methods [25, 28]. We use SCPNet baseline and conduct evaluation on SemanticKITTI val set.

Methods	Baseline	TENT [25]	Ours (Exp. C)	CoTTA [28]	Ours
mIoU	37.56	37.92	38.38	36.55	39.29
cIoU	50.24	49.86	52.95	50.61	56.09

and $\mathcal{F}^G$ are optimized using entropy minimization, where the update of $\mathcal{F}^G$ is assisted by CoTTA scheme. We verify that this setting achieves 36.55% of mIoU and 50.61% of cIoU, where the mIoU decreases from the baseline. The results strongly confirm the superiority of the proposed optimization goals and schemes, which effectively leverage the information from driving scenarios for SSC.

4.5 Experiments under the severe domain gap

We check the potential of TALoS under the test scenarios of a target domain considerably different from the source domain. Specifically, we tested SCPNet pre-trained on SemanticKITTI, on the driving sequences of SemanticPOSS. Unlike SemanticKITTI, which uses a 64-beam LiDAR, SemanticPOSS is captured with a 40-beam and targets campus rather than on typical roads, resulting in significantly different class distribution. Note that although TTA-based approaches could enhance performance by adapting the pre-trained model to the scene, the capability of the initial model itself is still essential.

Table 4 compares our performance with that of the baseline, which uses the pre-trained SCPNet without any adjustments. Unfortunately, due to the severe drastic gap between SemanticKITTI and SemanticPOSS, the mIoU performance is actually low for both methods, as we expected. Therefore, in this section, we would like to focus on the significant improvement achieved by TALoS over the baseline. In particular, TALoS shows its potential by achieving an improvement of over 10 in cIoU, which is less affected by changes in class distribution. We believe that combining TALoS with the prior studies targeting domain gaps could be an interesting direction for future research.

4.6 Additional experiments

Playback experiment. We further clarify the advantages of our continual approach with an intuitive experiment named "playback". Specifically, we first run TALoS on the SemanticKITTI validation sequence, from start to the end. During this first round, we save the weights of the gradual model $\mathcal{F}^G$ at a certain moment. Subsequently, we initialize the $\mathcal{F}^G$ with the saved weights, and run TALoS once again on the same sequence from the start. Here, during this playback round, we do not update the continual model at all. If the $\mathcal{F}^G$ indeed learned the distribution of the scene while not being biased to a certain moment during the first round, the playback performance would be better than that of the first round. Table 5 verifies this expectation, where the playback performs better compared to not only the baseline but also the first round of TALoS. We also provide a qualitative comparison between their results in Fig. 5, where the prediction of playback is clearly enhanced in terms of SSC.

Impact of number of iterations. Table 6 shows the impact of the number of iterations for updating $\mathcal{F}^M$ on the performance of TALoS. Notably, TALoS achieves significant gains in both mIoU and cIoU, even with a single iteration. Performance is enhanced as the number of iterations increases; however, we observe saturation after five iterations per sample. The results imply that the proposed method efficiently excavates the information we targeted, fully leveraging it with only a few iterations.

Table 4: Results of the cross-dataset evaluation, pre-training on SemanticKITTI [15], and evaluating on SemanticPOSS [39]. We compare the performance of TALoS with the baseline (SCPNet).

Methods	mIoU	cIoU	person	rider	car	trunk	plants	traffic-sign	pole	building	fence	bike	ground
Baseline (SCPNet [3])	7.6	25.2	0.8	0.3	2.5	2.9	16.8	0.6	**9.5**	28.7	**10.0**	3.8	7.2
Ours	**9.6**	**36.2**	**1.0**	**0.6**	**3.5**	**3.5**	**27.7**	**0.9**	8.2	**31.6**	6.1	**9.4**	**13.2**

Table 5: Results of the playback exp.

Methods	Baseline	Ours	Ours-Playback
cIoU	50.24	56.09	**56.91**
mIoU	37.56	39.29	**39.38**

Table 6: Impact of the number of iterations.

# of iterations	0 (baseline)	1	2	3 (ours)	5
cIoU	50.24	55.99	56.05	**56.09**	56.07
mIoU	37.56	39.09	39.22	39.29	**39.31**

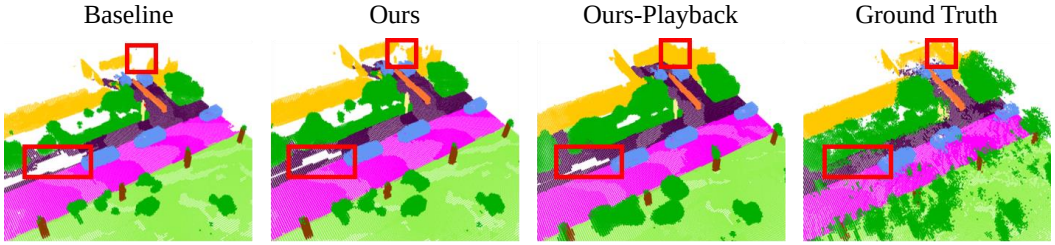

Figure 5: The results of the playback experiment. The red boxes depict the sequential improvements, implying that the gradual model indeed adapts to the scene as TTA proceeds.

5 Limitations

One of our approach's main limitations is that it addresses point clouds only. As there exist a number of SSC approaches using image data, exploring TTA for SSC in this setting could be an interesting research direction. We believe that the main philosophy of this paper, using the observation of one moment to guide the prediction of another moment, can be seamlessly extended to the image-based approaches, enhancing the practicality of SSC.

Further, we are aware of that the current state of SSC performance is still in its infancy, and thereby the gain achieved by TTA can be seem to less meaningful. However, we strongly believe that the proposed TALoS can be a even more promising approach for autonomous driving, as the field of SSC grows in the future.

6 Conclusion

This paper pioneers a novel Semantic Scene Completion (SSC) approach based on Test-time Adaptation (TTA). The proposed method, named TALoS, focuses on that observation made at one moment can serve as Ground Truth (GT) for scene completion at another moment. For this, we present several approaches to acquiring self-supervision that can be helpful in adapting the model to the scene in test time, from both the perspectives of geometric completion and semantic segmentation. To further capitalize on future information that is inaccessible at the current time, we introduced a dual optimization scheme that delays the model update until future observations become available. Evaluations on the Semantic-KITTI validation and test sets confirmed that TALoS significantly enhances the performance of the pre-trained SSC model, compared to the existing TTA approaches.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF2022R1A2B5B03002636).

References

[1] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017.

[2] Christoph B Rist, David Emmerichs, Markus Enzweiler, and Dariu M Gavrila. Semantic scene completion using local deep implicit functions on lidar data. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7205–7218, 2021.

[3] Zhaoyang Xia, Youquan Liu, Xin Li, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, and Yu Qiao. Scpnet: Semantic scene completion on point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17642–17651, 2023.

[4] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 111–119. IEEE, 2020.

[5] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang 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*, volume 35, pages 3101–3109, 2021.

[6] Xuemeng Yang, Hao Zou, Xin Kong, Tianxin Huang, Yong Liu, Wanlong Li, Feng Wen, and Hongbo Zhang. Semantic segmentation-assisted scene completion for lidar point clouds. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3555–3562. IEEE, 2021.

[7] Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 733–749, 2018.

[8] Hao Zou, Xuemeng Yang, Tianxin Huang, Chujuan Zhang, Yong Liu, Wanlong Li, Feng Wen, and Hongbo Zhang. Up-to-down network: Fusing multi-scale context for 3d semantic scene completion. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 16–23. IEEE, 2021.

[9] Martin Garbade, Yueh-Tung Chen, Johann Sawatzky, and Juergen Gall. Two stream 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[10] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

[11] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34:21808–21820, 2021.

[12] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34:3635–3649, 2021.

[13] M Jehanzeb Mirza, Inkyu Shin, Wei Lin, Andreas Schriebl, Kunyang Sun, Jaesung Choe, Mateusz Kozinski, Horst Possegger, In So Kweon, Kuk-Jin Yoon, et al. Mate: Masked autoencoders are online 3d test-time learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16709–16718, 2023.

[14] Ahmed Hatem, Yiming Qian, and Yang Wang. Point-tta: Test-time adaptation for point cloud registration using multitask meta-auxiliary learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16494–16504, 2023.

[15] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019.

[16] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Liu Bingbing. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *Conference on Robot Learning*, pages 2148–2161. PMLR, 2021.

[17] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, et al. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. *arXiv preprint arXiv:2306.09001*, 2023.

[18] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022.

[20] Yining Shi, Jiusi Li, Kun Jiang, Ke Wang, Yunlong Wang, Mengmeng Yang, and Diange Yang. Panossc: Exploring monocular panoptic 3d scene reconstruction for autonomous driving. In *2024 International Conference on 3D Vision (3DV)*, pages 1219–1228. IEEE, 2024.

[21] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9098, 2023.

[22] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023.

[23] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023.

[24] Hao Shi, Song Wang, Jiaming Zhang, Xiaoting Yin, Zhongdao Wang, Zhijian Zhao, Guangming Wang, Jianke Zhu, Kailun Yang, and Kaiwei Wang. Occfiner: Offboard occupancy refinement with hybrid propagation. *arXiv preprint arXiv:2403.08504*, 2024.

[25] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[26] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pages 6028–6039. PMLR, 2020.

[27] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023.

[28] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022.

[29] Francois Fleuret et al. Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9613–9623, 2021.

[30] Devavrat Tomar, Guillaume Vray, Behzad Bozorgtabar, and Jean-Philippe Thiran. Tesla: Test-time self-learning with automatic adversarial augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20341–20350, 2023.

[31] Inkyu Shin, Yi-Hsuan Tsai, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Sparsh Garg, In So Kweon, and Kuk-Jin Yoon. Mm-tta: Multi-modal test-time adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16928–16937, 2022.

[32] Haozhi Cao, Yuecong Xu, Jianfei Yang, Pengyu Yin, Shenghai Yuan, and Lihua Xie. Multi-modal continual test-time adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18809–18819, 2023.

[33] Peiyun Hu, Jason Ziglar, David Held, and Deva Ramanan. What you see is what you get: Exploiting visibility for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11001–11009, 2020.

[34] Li Ding and Chen Feng. Deepmapping: Unsupervised map estimation from multiple point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8650–8659, 2019.

[35] Chao Chen, Xinhao Liu, Yiming Li, Li Ding, and Chen Feng. Deepmapping2: Self-supervised large-scale lidar map optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9306–9316, 2023.

[36] JE Bresenham. Algorithm for computer control of a digital plotter'ibm syst. *J*, 4(1):25–30, 1965.

[37] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018.

[38] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[39] Yancheng Pan, Biao Gao, Jilin Mei, Sibo Geng, Chengkun Li, and Huijing Zhao. Semanticposs: A point cloud dataset with large quantity of dynamic instances. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 687–693. IEEE, 2020.

[40] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[41] Hyeonseong Kim, Yoonsu Kang, Changgyoon Oh, and Kuk-Jin Yoon. Single domain generalization for lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17587–17598, 2023.

[42] Inigo Alonso, Luis Riazuelo, Luis Montesano, and Ana C Murillo. Domain adaptation in lidar semantic segmentation by aligning class distributions. *arXiv preprint arXiv:2010.12239*, 2020.

A Implementation Details

This section provides more details about the implementation of our method. We utilize a voxel size of $256 \times 256 \times 32$, following SCPNet [3]. For cross-domain evaluation, we use the class mapping from semanticKITTI [15] to semanticPOSS [39] shown in Table A.1, following [41, 42]. All experiments are conducted using a single NVIDIA RTX A6000.

Meanwhile, during the test-time adaptation using TALoS, we only updated the segmentation sub-network and fixed the weights of the other modules of SCPNet. We experimentally observe that updating the whole network achieves slightly higher performance but is marginal.

Table A.1: Class mapping from semanticKITTI [15] to semanticPOSS [39]

KITTI	car	bicycle	person	bicyclist	road, sidewalk	building	fence	vegetation	trunk	pole	traffic sign
POSS	car	bike	person	rider	ground	building	fence	plants	trunk	pole	traffic-sign

B Class-wise Results on SemanticKITTI Validation Set

We provide class-wise results on the SemanticKITTI validation set in Table B.2. Notably, the proposed TALoS outperforms the baseline (SCPNet[3]) in most categories, in addition to the significant margins in overall metrics, mIoU and cIoU. These results strongly confirm the superiority of our method.

Table B.2: Quantitative results on SemanticKITTI validation set.

Methods	mIoU	cIoU	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
SCPNet [3]	37.6	50.2	51.2	**25.8**	37.7	57.6	43.8	22.9	18.8	4.2	70.8	**61.5**	53.0	15.3	33.6	32.2	38.9	33.9	52.9	39.8	**19.8**
TALoS	**39.3**	**56.1**	**51.9**	**25.8**	**38.7**	**60.1**	**46.1**	**24.0**	**19.9**	**5.3**	**75.0**	61.3	**55.2**	**17.0**	**36.7**	**33.1**	**44.6**	**35.0**	**57.7**	**40.1**	18.9

C Result on various baseline models

We conducted experiments using different architectures to prove that TALoS is a universally useful approach to various SSC models. As shown in Table C.3, TALoS meaningfully enhances SSC performance (mIoU) across different architectures and datasets. We utilized the baseline weights trained from the training set of the dataset on which validation was to be performed. The results imply that TALoS can be a promising solution for SSC in various settings.

Table C.3: Comparisons between the proposed TALoS and TENT [25] using various SSC models.

SSC method	Dataset	Baseline	TENT [25]	TALoS
SSCNet [1]	KITTI-360	17.0	17.0 (+0.0)	17.4 (+0.4)
SSA-SC [6]	SemanticKITTI	24.5	24.8 (+0.2)	25.3 (+0.8)
SCPNet [3]	SemanticKITTI	37.6	37.9 (+0.3)	39.3 (+1.7)

D Additional Experimental Results

All the experiments in Section D are conducted on the SemanticKITTI validation set.

D.1 Impact of threshold

We verify the impact of τ, the value thresholding the reliability for obtaining pseudo-GT, as in Equ. (6). As shown in Table D.4, the proposed method is quite robust to the change of τ, and both mIoU and cIoU are saturated after a certain point (0.75). Based on this result, we set the thresholding value to 0.75 by default in the main paper.

D.2 Effectiveness of noise

We verify the robustness of the proposed method against the errors that possibly exist in LiDAR calibration, as shown in Table D.5. We model the error by disturbing the transformation matrix in Equ. 2 with the noise sampled from the Gaussian distributions, which have standard deviation values listed in the first column of the table. Specifically, we add the noise to the angles of rotation and translation vectors, to acquire the disturbed projection matrices between the LiDAR coordinate systems. The results show that performance decreases as the level of noise increases, as expected. Nevertheless, we observe that the proposed TALoS achieves substantial performance even with the noise, still meaningfully higher than the baseline. These results imply the robustness and practicality of our method.

D.3 Impact of selecting moments

As the proposed TALoS explicitly leverages the observations made at various moments during adaptation, it is essential to select proper moments. To verify this, we conduct experiments by varying the chosen moments, as shown in Table D.6. For example, frame difference 1 in Table D.6 denotes selecting $j = i - 1$ and $k = i + 1$ as moments. Here, we need to specially handle the boundary cases, *e.g.*, at the start of the sequence (j<0) or the last of the sequence (k is larger than the number of all the samples in the sequence). For these cases, we simply do not use any losses relevant to those moments. Furthermore, for frame difference 0, we use $\mathbf{A}_i$, the pseudo-GT from the current moment, as the only self-supervision.

The results in Table D.6 show that performance decreases if the selected moments are too far from the current step, as the distant observation may not overlap with the current observation or overly force the model to learn completely unpredictable regions, leading to bias. Considering these, we set the frame difference to 1 by default in our setting.

Table D.4: Impact of changes in τ, the thresholding for reliability.

Reliability	mIoU	cIoU
0.65	39.14	55.42
0.7	39.24	55.8
0.75	**39.29**	**56.09**
0.8	39.28	56.05

Table D.5: Effectiveness of noises conducted on semantic-KITTI validation set.

Noise	mIoU	cIoU
0.05	38.58	54.09
0.03	38.76	54.49
0	**39.29**	**56.09**

Table D.6: Impact of selecting different moments in TALoS.

Frame Diff.	mIoU	cIoU
0	38.66	54.81
1	**39.29**	**56.09**
2	39.21	55.94
3	39.14	55.92
4	39.09	55.91

D.4 Comparison with Fusion-based Approaches

To analyze the effect of F^M, we compare our experiments with temporal fusion-based approaches. Specifically, we devise naive temporal fusion methods for the previous and current timesteps in two different ways, named early and late fusion. In early fusion, we merge the raw point clouds of both the previous and current timesteps and use the fused point cloud as input for our baseline (the pre-trained SCPNet). On the other hand, in late fusion, we separately obtain the predictions of the baseline at each timestep and aggregate the results of different timesteps at the voxel level. Here, when the predicted classes differ, we trust the one with lower entropy (which means higher confidence).

Table D.7 compares the performance of these fusion-based approaches with that of TALoS. For a fair comparison, we also replicate the performance of Exp. C in Table 1 of the main paper to the 5th row. Note that this setting exclusively uses F^M. The results show that the cIoU gain of TALoS exceeds the naive temporal fusions, both early and late. Also, for mIoU, the fusion-based methods even harm the mIoU performance, while TALoS achieves significant gain. Finally, it is noteworthy that Exp. C, with F^M only, still outperforms all the other fusion-based methods. This is strong evidence that F^M indeed performs something more and better than the naive temporal fusion.

D.5 Computational overhead

We provide the required time per step of the baseline (SCPNet), conventional TTA methods, and TALoS in Table D.8. We use SCPNet baseline and conduct evaluation on SemanticKITTI val set. Given the significant performance gains of TALoS, the result shows a reasonable trade-off of computational overhead and performance.

Table D.8: Time and performance for proposed TALoS and existing TTA methods [25, 28].

Method	Time per step (s)	overhead (%)	mIoU (%)
Baseline	2.26	-	37.56
TENT [25]	4.09	+81	37.92
CoTTA [28]	6.14	+171	36.55
TALoS	6.65	+194	**39.29**

Table D.7: Comparison with naive temporal fusion-based approaches.

Method	mIoU (%)	cIoU (%)
Baseline	37.56	50.24
Early fusion	35.86	52.85
Late fusion	37.11	52.49
Exp. C (with F^M only)	38.38	52.95
TALoS	**39.29**	**56.09**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state the main claims reflecting the contributions and scope of our paper in the abstract and introduction section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our methods in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the experiment settings or information needed to reproduce the main experimental results throughout the main paper and the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release the code upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all the training and test details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: As our method is deterministic, we believe that we do not have to report the statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We clearly provide the information on the computer resources in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm all the guidelines of the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We believe that there is no special societal impact of this work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite the source of the assets we used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.