Reliable Spatial-Temporal Voxels For Multi-Modal Test-Time Adaptation

Haozhi Cao[1], Yuecong Xu[1,2], Jianfei Yang[1,3](✉), Pengyu Yin[1], Xingyu Ji[1], Shenghai Yuan[1], and Lihua Xie[1]

[1] School of Electrical and Electronic Engineering, Nanyang Technological University
{haozhi002,jianfei.yang,pengyu001,xinyu001,shyuan,elhxie}@ntu.edu.sg
[2] Department of Electrical and Computer Engineering, National University of Singapore
yc.xu@nus.edu.sg
[3] School of Mechanical and Aerospace Engineering, Nanyang Technological University

Abstract. Multi-modal test-time adaptation (MM-TTA) is proposed to adapt models to an unlabeled target domain by leveraging the complementary multi-modal inputs in an online manner. Previous MM-TTA methods for 3D segmentation rely on predictions of cross-modal information in each input frame, while they ignore the fact that predictions of geometric neighborhoods within consecutive frames are highly correlated, leading to unstable predictions across time. To fulfill this gap, we propose Re**L**iable Sp**a**t**t**ial-temporal Vox**e**ls (**Latte**), an MM-TTA method that leverages reliable cross-modal spatial-temporal correspondences for multi-modal 3D segmentation. Motivated by the fact that reliable predictions should be consistent with their spatial-temporal correspondences, Latte aggregates consecutive frames in a slide window manner and constructs Spatial-Temopral (ST) voxels to capture temporally local prediction consistency for each modality. After filtering out ST voxels with high ST entropy, Latte conducts cross-modal learning for each point and pixel by attending to those with reliable and consistent predictions among both spatial and temporal neighborhoods. Experimental results show that Latte achieves state-of-the-art performance on three different MM-TTA benchmarks compared to previous MM-TTA or TTA methods. Visit our project site https://sites.google.com/view/eccv24-latte.

Keywords: Test-time adaptation · Multi-modal learning · 3D semantic segmentation

1 Introduction

3D semantic segmentation is a fundamental task to achieve various fully autonomous applications such as autonomous driving and robot navigation [9, 14]. With the increasing urge for robust sensing, multi-modal sensors (*e.g.*, cameras and LiDARs) are widely adopted in autonomous systems containing rich complementary information from different sensors. However, conventional deep-learning-based solutions lean on expensive point-wise annotations and perform

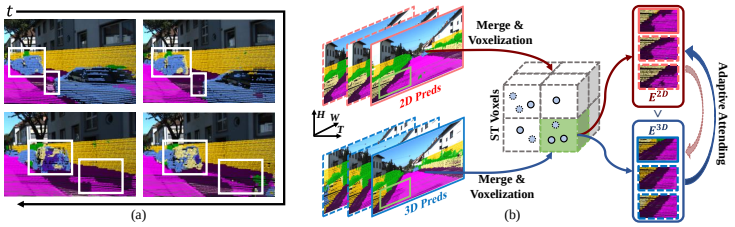


Fig. 1: Illustration of previous MM-TTA methods and our Latte. (a) visualizes the predictions from the previous state-of-the-art MM-TTA method [29] in consecutive frames on SemanticKITTI [1], where regions in white boxes exhibit noisy predictions when considering single-frame input. (b) presents the overall framework of Latte, which firstly obtains ST voxels for each modality given the merged input of consecutive frames within each slide window and then computes the modality-specific ST entropy E^{2D}, E^{3D}. ST entropy indicates the relative reliability of each voxel in each modality and therefore is utilized in the following adaptive cross-modal attending.

poorly on data from different domains due to the domain shift [6, 26, 33]. To resolve this problem, previous works have proposed Multi-Modal Unsupervised Domain Adaptation (MM-UDA) approaches [16, 18, 24, 44] that employ multi-modal information to transfer knowledge learned from the labeled source domain to the unlabeled target domain without expensive human annotations.

Nevertheless, the adaptation process of MM-UDA is completely offline, which requires full access to the source domain dataset and multiple training epochs. Despite its effectiveness, offline training is usually infeasible when encountering distribution shifts during the inference stage that require adapting models in real time. Inspired by Test-Time Adaptation (TTA) [22, 23], a recent work [29] proposes the first Multi-Modal Test-Time Adaptation (MM-TTA) method for 3D segmentation. Similar to the settings of TTA, MM-TTA prohibits access to any raw sample from the source domain dataset and adapts the model during the testing stage. This results in a *quick adaptation* scenario (*i.e.*, one epoch for training only as in [29,32]) that requires stable optimization in an online manner. To this end, MM-TTA methods must seek a reliable source of supervision signals.

To fulfill the efficiency and stability of adaptation, the current state-of-the-art MM-TTA method [29] evaluates modal reliability by the prediction consistency between different experts in a frame-wise manner. However, due to the domain shift between the online frame and the pre-trained source dataset, the single-frame prediction is usually unstable. Existing approaches rely on pseudo labels to refine the single-frame prediction but still fail in some cases. As shown in Fig. 1a, the single-frame refinement strategy [29] suffers from noisy single-frame predictions on the car in the white rectangle across time. Such inconsistent predictions across time degenerate the segmentation performance and aggravate denoising stress in downstream tasks (*e.g.*, semantic-based retrivial [46] and obstacle recognition). On one hand, these temporally unstable single-frame predictions could be wrongly regarded as reliable ones, propagating prediction noise to the other modality and causing error accumulation or even catastrophic forgetting [22,23].

On the other hand, while previous methods [4, 41] alleviate this instability by utilizing the average predictions of multiple augmented frames instead, they are computationally expensive for online adaptation since the inference time grows linearly with the increasing number of augmented frames.

This work aims to efficiently suppress the single-frame instability for MM-TTA by exploiting the correlation between multi-frame correspondences. Specifically, we propose that 3D space can be efficiently divided into multiple voxels [33], each of which contains points and point-corresponding pixels located at the same geometric region. While captured at different timestamps, it can be observed that points or pixels located at the same voxel can be viewed as different observations on the same semantic objects [28]. Reliable predictions should therefore be certain and consistent within their corresponding voxels across time, whether we evaluate them from a temporally global perspective (*i.e.*, all input frames) or a temporally local perspective (*i.e.*, certain consecutive frames). By aggregating the predictions in a voxel-wise manner and re-evaluating their reliability, the single-frame prediction noise as in Fig. 1a can be effectively alleviated.

Motivated by the above observation, we propose a novel MM-TTA method called Re**L**iable Sp**at**ial-**t**emporal Vox**e**ls (**Latte**), which leverages the spatial-temporal correspondences within consecutive frames in a cross-modal manner as in Fig. 1b. Specifically, given frames of point clouds with their estimated poses, we regard points in the same voxel extracted from the merged point cloud frames as spatial-temporal correspondences. Different from previous works that merge all input frames [5, 7] or regularize frame-to-frame consistency [28], Latte aggregates consecutive frames in a sliding window manner to estimate the temporally local prediction consistency through the proposed Spatial-Temporal voxels (ST voxels) and entropy (ST entropy). Based on the reliability estimated by ST entropy, we conduct cross-modal learning in an adaptive attending manner to reduce the contributions of predictions from the noisy modality. To better facilitate the progress of MM-TTA, we evaluate our Latte with more recent backbones (SegFormer [43] and SPVCNN [33]) on three different benchmarks, where Latte consistently outperforms previous TTA or MM-TTA methods.

In summary, our contributions are three-fold. Firstly, we propose Latte, a novel MM-TTA method for multi-modal 3D segmentation, which is the first work that incorporates spatial-temporal correlations for MM-TTA segmentation to our best knowledge. Secondly, Latte efficiently extracts spatial-temporal correspondence through ST voxels and estimates their reliability by ST entropy, which is further incorporated to enhance adaptive cross-modal attending. Thirdly, our experimental results show that Latte outperforms previous state-of-the-art TTA and MM-TTA methods on three different benchmarks.

2 Related Works

Test-Time Adaptation (TTA). TTA is proposed to mitigate the domain shift between the training data and the testing data in an online manner, which is firstly proposed by TENT [39]. Due to its practical and challenging setting,

TTA is attracting more and more attention. Previously proposed TTA methods attempt to address this task from different perspectives, such as entropy minimization [22, 23, 39], self-training with pseudo-labels [12, 40], and augmentation invariance [41, 48]. In addition to regularizing predictions, some methods propose to adapt networks from the feature level instead [21, 32], while more recent methods begin to consider different variants of TTA scenarios, such as continual TTA [4, 22, 41], non-i.i.d TTA [11], or the mix of aforementioned cases [47]. Different from source-free adaptation [19, 30] which also discards the access to the source domain yet requires multiple training epochs, the adaptation process of TTA is completely online, following the "one-pass" protocol as in [32]. While most existing TTA methods are designed for image classification or image segmentation without explicit temporal correlation, we argue temporal information in 3D segmentation can be effectively leveraged for TTA as Latte in this work.

Multi-modal domain adaptation for 3D segmentation. To avoid expensive annotation costs and overcome the poor generalizability of fully supervised solutions, various multi-modal domain adaptation methods have been investigated. Specifically for MM-UDA, xMUDA [16] is the primary works that explore incorporating cross-modal learning with MM-UDA. It regards cross-modal prediction consistency and pseudo-labels as its supervision signals in the unlabeled target domain. Most subsequent works propose solutions to address its limitations from different perspectives, such as more diverse point-pixel correspondence [24, 44], procedures to mitigate domain gaps [18, 20], and alleviating class-imbalanced problem [3]. Since MM-UDA is infeasible when facing online domain shifts, some previous works [4, 29] develop different methods to address the domain gap in an online manner, including cross-modal pseudo-label generation [29] and adaptive modality attending [4]. A recent work [30] proposes a source-free domain adaptation method by estimating cross-modal prediction agreement, yet it requires multiple offline training epochs. Different from previous multi-modal domain adaptation methods that ignore temporal relationships, Latte couples spatial-temporal correspondence to mitigate the single-frame instability and achieve better online adaptation performance.

Temporal processing of point clouds. Since point clouds are naturally in the form of temporally consecutive frames, previous methods have widely explored how to leverage such temporal information for different tasks. Specifically, scene flow [15, 35, 37, 38] contains the movement of every point in the 3D world. Considering such point-wise computation expensive, previous works on different tasks (*e.g.*, segmentation [5,6,28] and object detection [25]) usually interpret the temporal information from a simpler perspective. In terms of fully supervised learning, Choy *et al.* [6] propose the first 4D segmentation method by generalized sparse convolution kernels [13] and Fan *et al.* [8] turn to the fully connected learning scheme as in [26], while Xu *et al.* [45] extract long-term temporal information by an efficient memory bank. On the other hand, considering scenarios with insufficient or zero annotations, some works propose different strategies to couple spatial-temporal information without ground-truth labels. For instance, Chen *et al.* [5] proposes to guide voxelized spatial-temporal representation with

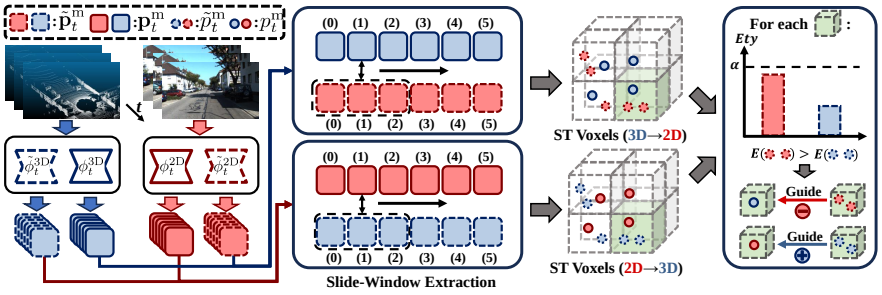


Fig. 2: Overall structure of Latte. Taking a student prediction frame of one modality as the query input, our slide-window extraction searches its spatial-temporal correspondences through voxelization within a time window to establish the temporally local prediction consistency. ST voxels are then generated, where those with high ST entropy (larger than α-quantile) are discarded as unreliable correspondences, while the others are leveraged for adaptive cross-modal learning by attending to the modality with lower ST entropy in a voxel-wise manner.

strong text embedding from CLIP [27] instead of labels in a multi-modal manner, while Saltor *et al*. [28] interpret such representation by nearest neighbor search in a frame-to-frame manner. In this work, we design a slide window aggregation and voxelization rather than aggregating all frames as in [5], which yields a better indication of the prediction reliability for each point.

3 Methodology

Problem definition. During MM-TTA for 3D segmentation, both 2D RGB images and 3D point clouds are captured consecutively in the target domain $\mathcal{T}$, denoted as $\mathbf{x}^{\text{2D}}_{\mathcal{T},t} \in \mathbb{R}^{3\times H\times W}$ and $\mathbf{x}^{\text{3D}}_{\mathcal{T},t} \in \mathbb{R}^{N\times 4}$, respectively, where t represents the frame order and H, W are the height and width of input images. The point-wise input contains the 3D coordinate $\{x, y, z\}$ and the point feature (*e.g.*, intensity, reflectivity, etc.). Modality-specific networks are pre-trained on the labeled source domain before adaptation, denoted as $\phi^{\text{m}}_{\mathcal{T},t}(\cdot)$ with $\text{m} \in \{\text{2D}, \text{3D}\}$. During MM-TTA, both networks are initialized from the parameters pre-trained on the source domain. Following previous works [16,29], 3D points are projected to the 2D image based on the relative projection matrix to obtain cross-modal correspondences within each frame, resulting in the modality-specific predictions $\mathbf{p}^{\text{m}}_{\mathcal{T},t} = \phi^{\text{m}}_{\mathcal{T},t}(\mathbf{x}^{\text{m}}_{\mathcal{T},t}), \mathbf{p}^{\text{m}}_{\mathcal{T},t} \in \mathbb{R}^{N\times K}$ where K denote the number of semantic classes. Data from the labeled source domain are strictly inaccessible and we omit the target domain subscript $\mathcal{T}$ by default.

The multi-modal input of online 3D segmentation is temporally consecutive, which exhibits strong temporal relationships between frames. The geometric neighborhoods in consecutive frames can be viewed as the naturally various observations on the same semantic object that can be leveraged to stabilize noisy single-frame predictions. Motivated by this observation, we develop **Latte**,

which discovers reliable spatial-temporal (ST) correspondences to improve both optimization stability and prediction consistency during MM-TTA. As shown in Fig. 2, given consecutive frames of 2D images and 3D point clouds as input, Latte firstly extracts their frame-wise predictions from both student and teacher networks (Sec. 3.1). These frame-wise predictions are then passed to the slide-window extraction and voxelization (Sec. 3.2). ST voxels and ST entropy are subsequently computed indicating the prediction consistency and certainty (Sec. 3.3). The final ST entropy is leveraged for cross-modal attending to alleviate the noise from unreliable modality-specific predictions (Sec. 3.4).

3.1 Frame-wise Predictions from Students and Teachers

Before exploiting spatial-temporal correspondence, we first generate the frame-wise predictions from each modality. Motivated by the fact that moving average models can provide more stable online predictions [4,34,41], Latte utilizes weight-averaged teacher models and fast student models for each modality as in previous works [29, 41], denoted as $\tilde{\phi}_t^{\mathrm{m}}(\cdot)$ and $\phi_t^{\mathrm{m}}(\cdot)$, respectively. Given each input frame $\mathbf{x}_t^{\mathrm{m}}$, the student predictions $\mathbf{p}_t^{\mathrm{m}}$ and teacher predictions $\tilde{\mathbf{p}}_t^{\mathrm{m}}$ are computed as:

$$\tilde{\mathbf{p}}_t^{\mathrm{m}} = \tilde{\phi}_t^{\mathrm{m}}(\mathbf{x}_t^{\mathrm{m}}), \mathbf{p}_t^{\mathrm{m}} = \phi_t^{\mathrm{m}}(\mathbf{x}_t^{\mathrm{m}}), \tag{1}$$

where the teacher model $\tilde{\phi}_t^{\mathrm{m}}$ is initialized from the source domain pre-trained models. In the following process, the teacher's predictions of one modality are regarded as the cross-modal guidance of student predictions from the other modality. In practice, the gradient propagation of teacher model predictions is discarded to prevent directly updating teacher models.

3.2 Slide Window Frame Aggregation and Voxelization

Our goal is to efficiently leverage spatial-temporal correspondences between predictions and seek reliability estimation through prediction consistency within correspondences. Establishing spatial-temporal correspondences in 3D space has been explored in other deep learning tasks, where they mainly consider temporally global relationships of all input frames [5, 28] or frame-to-frame correspondences [28]. However, they both have some limitations in terms of consistency and certainty evaluation: the former can not highlight the risk of some inconsistent predictions within a short time window (*i.e.*, temporally local inconsistency), while the latter suffers from an insufficient number of correspondences.

To effectively empower spatial-temporal information with Latte, we propose a slide window frame aggregation that focuses on temporally local correspondences that lie within a time window to evaluate temporally local consistency. Specifically, given a student prediction frame $\mathbf{p}_i^{\mathrm{m}}$ with the temporal index i as the query, the correspondence search is conducted within a time window of consecutive frames denoted as $\{j||j-i| \leq w_t\}$, where w_t is the pre-defined time window size. The merged point cloud $\hat{\mathbf{x}}_i^{\mathrm{3D}} \in \mathbb{R}^{\hat{N}\times 3}$ (point features are omitted for simplicity) for correspondence search is then formulated as:

$$\hat{\mathbf{x}}_i^{\mathrm{3D}} = \mathrm{cat}(\{\mathbf{T}_{j\to i} * \mathbf{x}_j^{\mathrm{3D}} || |j-i| \leq w_t\}), \quad \mathbf{T}_{j\to i} = \mathbf{T}_i^{-1}\mathbf{T}_j, \tag{2}$$

where $\mathbf{T}_i$ is the estimated pose at frame i and so as $\mathbf{T}_j$, which can be easily obtained online through off-the-shelf SLAM algorithms [17,36]. $\text{cat}(\cdot)$ is the concatenation operation along the point number dimension while $*$ denotes the pose transformation. The voxelization is then performed on $\hat{\mathbf{x}}_i^{\text{3D}}$, formulated as:

$$\mathbf{v}_i^{\text{3D}} = \mathcal{V}_{\mathbf{s}}(\hat{\mathbf{x}}_i^{\text{3D}}), \quad \mathbf{v}_i^{\text{3D}} \in \mathbb{R}^{N_{\text{v}} \times 3}, \tag{3}$$

$$\mathbf{M}_i^{\text{3D}} = \{k | \forall g \in [1, \hat{N}], \lfloor \hat{\mathbf{x}}_{i,g}^{\text{3D}} / \mathbf{s} \rfloor = \mathbf{v}_{i,k}^{\text{3D}}\}, \quad \mathbf{M}_i^{\text{3D}} \in \mathbb{R}^{\hat{N}}, \tag{4}$$

where $\mathbf{v}_i^{\text{3D}}$ is the extracted voxels and $\mathcal{V}_{\mathbf{s}}$ is the voxelization operation with a voxel size $\mathbf{s}$. After voxelization, points located in the same voxel can be regarded as correspondences since they are geometric neighborhoods, where the hash table $\mathbf{M}_i^{\text{3D}}$ that maps each voxel back to its containing points (*e.g.*, $\mathbf{v}_{i,k}^{\text{3D}}$ to $\hat{\mathbf{x}}_{i,g}^{\text{3D}}$ as in Eq. (4)) is preserved for the subsequent correspondence mapping. In this way, it provides more diverse neighborhood selection strategies with temporally local correspondences for consistency and certainty evaluation compared to globally merging all frames in one single step. In practice, the slide window aggregation and voxelization are conducted iteratively for each frame in the input batch.

3.3 Spatial-Temporal Voxels and Entropy

Given the geometric correspondences captured by voxelization, our goal is to emphasize the predictions with correspondence consistency within each voxel. After slide window aggregation and voxelization, there inevitably exist some unreliable correspondences of two different types: (i) voxels that contain different types of semantic objects (*e.g.*, those located on the contact surface between cars and roads) and (ii) voxels with highly inconsistent predictions due to uncertain predictions across time. These unreliable correspondences contain high uncertainty and inconsistency, which could harm the online optimization of models if they as references. To alleviate the side-effect introduced by such unreliable correspondences, we propose ST voxels that leverage ST entropy for voxel-wise reliability evaluation and cross-modal attending.

Specifically, an ST voxel contains two components, including the query and the reference, where the query is encouraged to be consistent with the reference. Considering that the moving average teacher can usually generate more stable predictions [4, 41], we regard the single-frame student predictions as the query while multi-frame teacher predictions as our reference, forming ST voxels denoted as $\mathbf{v}_i^{\text{ST}}$. Without losing generality, here we take a single ST voxel indexed by k in frame i as an example, denoted as $\mathbf{v}_{i,k}^{\text{ST}} = \{\mathbf{p}_{\text{q}}^{\text{m}}, \mathbf{p}_{\text{r}}^{\text{m}}\}$, where $\mathbf{p}_{\text{q}}^{\text{m}}$ and $\mathbf{p}_{\text{r}}^{\text{m}}$ are the point-wise student predictions at the query frame i and teacher predictions in the frame searching range j, respectively, formulated as:

$$\mathbf{p}_{\text{q}}^{\text{m}} = \{\mathbf{p}_{i,g}^{\text{m}} | \lfloor \mathbf{x}_{i,g}^{\text{3D}} / \mathbf{s} \rfloor = \mathbf{v}_{i,k}^{\text{3D}}\}, \quad \mathbf{p}_{\text{q}}^{\text{m}} \in \mathbb{R}^{N_{\text{q}} \times K} \tag{5}$$

$$\mathbf{p}_{\text{r}}^{\text{m}} = \{\tilde{\mathbf{p}}_{j,g}^{\text{m}} | \lfloor \mathbf{x}_{j,g}^{\text{3D}} / \mathbf{s} \rfloor = \mathbf{v}_{i,k}^{\text{3D}}, |j - i| \leq w_t\}, \quad \mathbf{p}_{\text{r}}^{\text{m}} \in \mathbb{R}^{N_{\text{r}} \times K}. \tag{6}$$

The reliability of each voxel is then evaluated as the Shannon's entropy [42] of the average teacher predictions, where the unreliable ST voxels are then filtered

out given a pre-defined quantile α as in Fig. 2, denoted as:

$$E_{i,k}^{\mathrm{m}} = -\sum_{c}^{K} \bar{\mathbf{p}}_{\mathrm{r},c}^{\mathrm{m}} \log \bar{\mathbf{p}}_{\mathrm{r},c}^{\mathrm{m}}, \quad \bar{\mathbf{p}}_{\mathrm{r}}^{\mathrm{m}} = \sum_{n=1}^{N_{\mathrm{r}}} \psi(\mathbf{p}_{\mathrm{r},n}^{\mathrm{m}})/N_{\mathrm{r}}, \tag{7}$$

$$h_{i,k} = \begin{cases} 0, & \text{if } \ E_{i,k}^{m} > Q^{m}(\alpha) \\ 1, & \text{if } \ E_{i,k}^{m} \leq Q^{m}(\alpha) \end{cases}, \tag{8}$$

where $E_{i,k}^{\mathrm{m}}$ is regarded as the ST entropy of voxel k for modality m, indicating the modal-specific prediction reliability within this voxel. $Q^{m}(\alpha)$ is the α-quantile of all ST entropy in modality m in the same batch and $\psi(\cdot)$ denotes the Softmax function along the class dimension.

3.4 ST Voxel Aided Cross-Modal Learning

Each modality has its pros and cons under different conditions and cross-modal learning for online scenarios like MM-TTA should therefore possess an adaptive mechanism that can attend to "pros" and meanwhile suppress the "cons". To achieve that, Latte performs cross-modal learning by attending to the modality with more consistent and certain predictions from a spatial-temporal perspective. Specifically, taking ST voxel $\mathbf{v}_{i,k}^{\mathrm{ST}}$ and its corresponding ST entropy $E_{i,k}^{\mathrm{m}}$ as an example, the voxel-wise cross-modal attending weights and the weighted cross-modal consistency loss $\mathcal{L}_{i,k}^{\mathrm{xM}}$ are formulated as:

$$w_{\mathrm{v}}^{\mathrm{2D}} = \frac{\exp(E_{i,k}^{\mathrm{2D}})}{\exp(E_{i,k}^{\mathrm{2D}}) + \exp(E_{i,k}^{\mathrm{3D}})}, \quad w_{\mathrm{v}}^{\mathrm{3D}} = 1 - w_{\mathrm{v}}^{\mathrm{2D}}, \tag{9}$$

$$\mathcal{L}_{i,k}^{\mathrm{xM}} = w_{\mathrm{v}}^{\mathrm{2D}} D_{\mathrm{KL}}(\bar{\mathbf{p}}_{\mathrm{q}}^{\mathrm{3D}} \| \bar{\mathbf{p}}_{\mathrm{r}}^{\mathrm{2D}}) + w_{\mathrm{v}}^{\mathrm{3D}} D_{\mathrm{KL}}(\bar{\mathbf{p}}_{\mathrm{q}}^{\mathrm{2D}} \| \bar{\mathbf{p}}_{\mathrm{r}}^{\mathrm{3D}}), \tag{10}$$

where $D_{\mathrm{KL}}(\cdot)$ represents the KL-divergence between two probability. $\bar{\mathbf{p}}_{\mathrm{q}}^{\mathrm{3D}}$ is the average query predictions in the ST voxel, similarly computed as in Eq. (7).

The cross-modal attending is further extended from the voxel level to the point level for better online predictions and cross-modal pseudo-label generation. Specifically, the ST entropy is propagated from the ST voxel to its reference point-wise prediction, indicating its confidence and consistency within its spatial-temporal neighborhoods. If one's ST entropy has been filtered out as in Eq. (8), its ST entropy would fall back to the point-level entropy. For arbitrary point-wise predictions $\mathbf{p}^{\mathrm{m}} \in \mathbf{p}_{\mathrm{r}}^{\mathrm{m}}$, the cross-modal predictions y^{xM} are formulated as:

$$\hat{E}_{\mathrm{r}}^{\mathrm{m}} = h_{i,k} E_{i,k}^{\mathrm{m}} - (1 - h_{i,k}) \sum_{c}^{K} \mathbf{p}_{c}^{\mathrm{m}} \log \mathbf{p}_{c}^{\mathrm{m}}, \tag{11}$$

$$w_{\mathrm{p}}^{\mathrm{2D}} = \frac{\exp(\hat{E}^{\mathrm{2D}})}{\exp(\hat{E}^{\mathrm{2D}}) + \exp(\hat{E}^{\mathrm{3D}})}, \quad w_{\mathrm{p}}^{\mathrm{3D}} = 1 - w_{\mathrm{p}}^{\mathrm{2D}}, \tag{12}$$

$$\mathbf{p}^{\mathrm{xM}} = w_{\mathrm{p}}^{\mathrm{2D}} \mathbf{p}^{\mathrm{2D}} + w_{\mathrm{p}}^{\mathrm{3D}} \mathbf{p}^{\mathrm{3D}}, \quad y^{\mathrm{xM}} = \arg\max_{c} \mathbf{p}^{\mathrm{xM}}. \tag{13}$$

Algorithm 1: Adaptation and Online Prediction Process of Latte

Target Domain: Multi-modal input $\mathcal{X}_\mathcal{T} = \{(\mathbf{x}_t^{\text{2D}}, \mathbf{x}_t^{\text{3D}})\}$
Init Model: Teacher $\tilde{\phi}_t^{\text{m}}$, student ϕ_t^{m}, m $\in \{\text{2D}, \text{3D}\}$, Slide window size w_t
for $\boldsymbol{x} = \{(\boldsymbol{x}_t^{2D}, \boldsymbol{x}_t^{3D})\}_{t=t_0+1}^{B} \in \mathcal{X}_\mathcal{T}$ **do**
1. Get student and teacher predictions $\tilde{\mathbf{p}}_t^{\text{m}}, \mathbf{p}_t^{\text{m}}$ by Eq. (1)
for $i \in [t_0+1, t_0+B]$ **do**
2. Get $\hat{\mathbf{x}}_i^{\text{3D}}, \mathbf{v}_i^{\text{3D}}, \mathbf{M}_i^{\text{3D}}$ within the time window of size w_t by Eq. (2)-(4)
3. Compute ST voxels $\mathbf{v}_i^{\text{ST}}$ and ST entropy E_i^{m} by Eq. (5)-(8)
4. Compute $\mathcal{L}_i^{\text{xM}}$ by Eq. (9)-(10)
end
5. Compute $\hat{E}_{\text{r}}^{\text{m}}$ and get $\mathbf{y}_t^{\text{xM}}$ as predictions by Eq. (11)-(13)
6. Compute overall loss by Eq. (14) and update models by Eq. (15)
end

3.5 Online Predictions and Optimization

As shown in Fig. 2, given a batch of B consecutive frames $\{\mathbf{x}_t^{\text{m}}\}_{t=t_0+1}^{t_0+B}$ as input, we first extract the student and teacher predictions through Eq. (1). Subsequently, frames are aggregated and voxelized as in Eq. (2), resulting in a batch of merged point cloud $\{\hat{\mathbf{x}}_t^{\text{3D}}\}_{t=t_0+1}^{t_0+B}$ and voxels $\{\mathbf{v}_t^{\text{3D}}\}_{t=t_0+1}^{t_0+B}$. The ST voxels and ST entropy computation process is then applied to each frame of the merged point cloud and voxels. Due to the overlap of the slide window, multiple ST entropy values could be propagated to the same point prediction, where we empirically take the average value of all ST entropy received to proceed Eq. (11). The overall loss function and the updating scheme can be formulated as:

$$\mathcal{L} = \sum_t \mathcal{F}(\mathbf{p}_t^{\text{m}}, \mathbf{y}_t^{\text{xM}}) + \frac{\lambda_{\text{xM}}}{B} \sum_t \sum_k \mathcal{L}_{t,k}^{\text{xM}}, \tag{14}$$

$$\tilde{\theta}_t^{\text{m}} = \lambda_s \tilde{\theta}_{t-1}^{\text{m}} + (1-\lambda_s)\theta_t^{\text{m}}, \tag{15}$$

where $\mathcal{F}(\cdot)$ is the cross-entropy function and $\mathbf{y}_t^{\text{xM}}$ is the point-wise cross-modal pseudo-label frame from Eq. (13), which is also regarded as our cross-modal prediction for evaluation. $\tilde{\theta}_t^{\text{m}}$ and θ_t^{m} are the model parameters of teacher and student models, respectively, where λ_s is the momentum update coefficient. λ_{xM} is pre-defined coefficient for the cross-modal consistency loss. Our adaptation and inference process is summarized in Algo. 1

4 Experimental Results

In this section, we present our thorough experimental results on three different benchmarks. Details of benchmarks, backbones, and settings are first present in Sec. 4.1. The main results are then illustrated in Sec. 4.2, followed by detailed ablation studies and qualitative results in Sec. 4.3.

4.1 Benchmarks and settings

Benchmark Details. To thoroughly investigate the effectiveness of Latte, we conduct our experiments on three different MM-TTA benchmarks: (i) **USA-to-Singapore (U-to-S)**, (ii) **A2D2-to-SemanticKITTI (A-to-S)**, and (iii) **Synthia-to-SemanticKITTI (S-to-S)**. Specifically, U-to-S is developed from NuScenes-LiDARSeg [2] dataset, where the domain gap is mainly credited to the infrastructure difference between countries. Same as [16], the source domain and the target domain data are selected by filtering country keywords based on the data recording description in NuScenes. Different from previous UDA methods [16,24] utilizing self-defined class mappings, we follow the official mapping in NuScenes-LiDARSeg, forming a 16-class segmentation benchmark. The remaining two benchmarks follow a similar setting as in [29], where the domain gap of A-to-S [1, 10] lies in different LiDAR mounting positions and image resolutions while the one of S-to-S [1] lies in different patterns between synthetic and real data for both modalities. For A-to-S, we adopt the same class-mapping as in [16,29] which leads to a 10-class benchmark, while we design our class-mapping for S-to-S since its original class map has not been revealed previously [29], forming a 9-class benchmark. *More details are presented in our appendix.*

Baseline methods. Previous TTA and MM-TTA methods are included in comparison with Latte. For TTA methods, we compare Latte with point-wise pseudo-labels with filtering (PsLabel), TENT [39], ETA [22], SAR [23] with and without restoring (denoted as SAR-rs and SAR, respectively). MMTTA [29] as well as its online version of xMUDA [16] (including pure xMUDA and pseudo-label version xMUDA+PL) are additionally included as the previous SOTA methods for comparison. Considering that EATA [22] with Fisher regularizer requires pre-access to the target domain samples, we compare Latte with ETA, an official variant of EATA without this regularizer for a fair comparison. We further present a performance lower bound by directly testing with source pre-trained models (*i.e.*, Source only) and an upper bound by online adapting networks with ground-truth labels from the target domain for one epoch (*i.e.*, Oracle TTA).

Implementation details. With the rapid progress in 2D and 3D segmentation, we employ all baseline methods and Latte with more recent SegFormer [43] (SegFormer-B1) and SPVCNN [33] (SPVCNN-cr1) as their 2D and 3D backbones, respectively, to investigate the improvement of existing TTA methods when combined with advanced networks. The pre-training procedures on the source domain follow a similar setting as in [16], except for 2D SegFormer which is trained with a base learning rate of 6E-5 with an AdamW optimizer and Poly scheduler as in [43]. For Latte, we maintain the same learning rate and optimizer as in the pre-training stage, while the scheduler is disabled. For our voxelization, we leverage the off-the-shelf SLAM algorithm KISS-ICP [36] to generate poses and utilize a window size w_t of 3 with a voxel size of 0.2m, while the quantile α and coefficient λ_{xM} are set to 0.9 and 0.3, respectively. λ_s is empirically set to 0.99 as in [4,31]. For all methods, we strictly follow the *one-pass* protocol to first evaluate the predictions of the input batch and then update the networks, where we only update the trainable parameters in normalization layers. For all base-

Table 1: Performance (mIoU) comparison of Latte. Latte outperforms all previous SOTA methods on the cross-modal metric (xM) across three benchmarks. Here "MM" denotes whether the method contains multi-modal interaction or not. Cross-modal predictions are computed as the Softmax average of modal outputs except for Latte. "Avg" present the average xM performance across three benchmarks.

Method	Publication	MM	U-to-S			A-to-S			S-to-S			Avg
			2D	3D	xM	2D	3D	xM	2D	3D	xM	
Source only	-	✗	31.4	41.1	43.9	47.4	17.9	44.3	23.4	36.4	38.2	42.1
Oracle TTA	-	✗	38.7	45.5	50.3	49.1	61.3	62.5	36.0	54.5	54.6	55.8
TENT [39]	ICML-21	✗	36.8	36.1	41.1	43.3	44.4	49.1	22.4	35.5	37.5	42.6
ETA [22]	ICML-22	✗	36.7	34.7	43.7	43.2	42.5	49.7	22.4	30.6	32.9	42.1
SAR [23]	ICLR-23	✗	36.6	37.2	43.9	43.0	42.7	50.3	20.7	31.3	32.9	42.4
SAR-rs [23]	ICLR-23	✗	36.6	35.5	43.9	43.1	43.8	50.1	24.2	25.1	28.8	40.9
xMUDA [16]	PAMI-22	✓	19.4	22.9	24.2	13.1	33.0	32.2	12.2	14.8	14.1	23.5
xMUDA+PL [16]	PAMI-22	✓	36.2	38.2	43.0	42.8	48.3	50.9	30.9	34.1	36.3	43.4
MMTTA [29]	CVPR-22	✓	37.2	**41.5**	45.4	44.5	51.7	53.7	27.5	35.1	35.5	44.9
PsLabel	-	✓	**37.7**	35.8	41.8	**46.7**	44.3	50.0	29.2	33.3	35.3	42.4
Latte (Ours)	-	✓	37.4	41.0	**46.0**	46.1	**52.6**	**54.3**	**33.2**	**39.3**	**41.6**	**47.3**

lines, we conduct a parameter search and report their best results. We utilize the mean intersection over union (mIoU) as our evaluation metric. All experiments are conducted with PyTorch on a single RTX 3090. *More details are presented in our appendix.*

4.2 Overall Results

Tab. 1 presents the overall results of our Latte on three benchmarks compared with previous SOTA methods. In terms of average cross-modal predictions (Avg), Latte surpasses all previous TTA or MM-TTA methods, with a significant relative improvement of more than 5.3%. For the less challenging U-to-S and A-to-S (*i.e.*, gap between Oracle TTA and best TTA less than 10%), Latte achieves consistent relative improvements of 1.3% on U-to-S and 1.1% on A-to-S compared to MMTTA, respectively. Most existing methods except for xMUDA [16] can achieve superior performance compared to source-only results when one modality is reliable as in U-to-S (3D prevails) and A-to-S (2D prevails).

In terms of the most challenging S-to-S, Latte exhibits significant performance improvement compared to previous SOTA MM-TTA methods such as MMTTA [29] and xMUDA+PL [16], with a relative gap of 14.6%. In fact, all previous methods except for our Latte perform inferiorly compared to source pre-trained models (Source only), due to the significant domain gaps caused by the non-trivial pattern difference in both modalities between synthetic and real data. In this case, the instability of single-frame predictions becomes exacerbated, leading to performance degeneration in MMTTA which mainly relies on single-frame refinement. This justifies the effectiveness of Latte under either challenging or less challenging TTA scenarios. Another interesting observation is that almost all cross-modal methods outperform other single-modal TTA methods, *e.g.*, even the simplest PsLabel can outperform recent TTA methods like SAR [23]. This reveals the non-trivial benefits of cross-modal learning for TTA with segmentation. *More comparisons about efficiency are included in appendix.*

Table 2: Ablation studies about the effectiveness of cross-modal learning. When disabling $\mathbf{y}^{\mathrm{xM}}$, we utilize the average Softmax of all modalities as the pseudo-labels as a replacement. In experiment No. 8, we replace the ST entropy (ST ety) with point-wise entropy in Eq. (13) and set $w_{\mathrm{v}}^{\mathrm{m}}$ in Eq. (10) to 0.5. Here $\mathcal{L}_{32}^{\mathrm{xM}}$ and $\mathcal{L}_{23}^{\mathrm{xM}}$ refer to the former half and the latter half in Eq. 10, respectively.

No.	Description	$\mathbf{y}^{\mathrm{xM}}$	$\mathcal{L}_{23}^{\mathrm{xM}}$	$\mathcal{L}_{32}^{\mathrm{xM}}$	U-to-S			A-to-S			S-to-S		
					2D	3D	xM	2D	3D	xM	2D	3D	xM
0	Source only	-	-	-	31.4	41.1	43.9	47.4	17.9	44.3	23.4	36.4	38.2
1	only ST ety label	✓			36.8	34.5	43.7	43.3	42.2	49.9	22.4	30.6	33.3
2	only ST 2D→3D		✓		32.5	34.5	41.0	41.7	42.2	48.2	27.7	30.6	32.9
3	only ST 3D→2D			✓	36.8	39.6	42.6	43.3	50.2	48.4	22.4	26.7	25.6
4	ST ety label + ST 2D→3D	✓	✓		**37.5**	35.8	41.0	43.6	47.5	49.8	31.0	36.0	38.4
5	ST ety label + ST 3D→2D	✓		✓	37.4	40.9	45.9	**46.2**	52.1	53.4	32.5	38.7	39.7
6	ST 2D↔3D		✓	✓	36.8	40.3	43.8	44.0	47.0	51.2	30.4	35.6	37.7
7	w/o percentile in Eq. (8)	✓	✓	✓	37.4	40.3	45.4	45.7	51.4	53.2	32.9	39.0	**41.7**
8	Point ety vs. ST ety	✓	✓	✓	37.0	40.6	45.1	45.1	51.4	52.7	31.6	36.7	38.2
9	Latte	✓	✓	✓	37.4	**41.0**	**46.0**	46.1	**52.6**	**54.3**	**33.2**	**39.3**	41.6

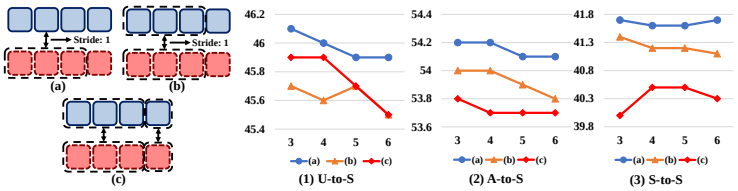


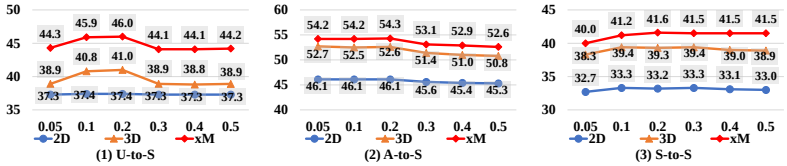
Fig. 3: Ablation studies of different frame aggregation mechanisms. Besides (a) Latte's slide window aggregation, we test two different aggregation methods, including (b) replacing single frame student predictions with multiple frames in Eq. (5) and (c) adding non-overlapping aggregation to (b). Results show that our slide window aggregation can better evaluate the local consistency and results in consistent improvement.

4.3 Ablation Studies and Visualziation

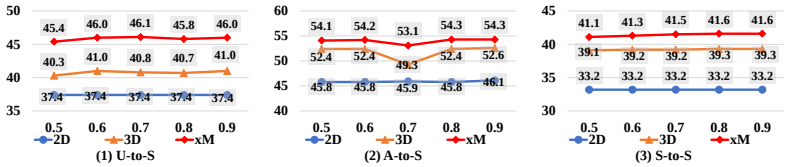
In this section, we investigate the effectiveness of each component in Latte by thorough ablation studies and qualitative results.

Effectiveness of cross-modal learning. The main optimization objectives of Latte contain three parts, including ST-entropy-based cross-modal consistency (both 2D-to-3D and 3D-to-2D) and cross-entropy loss between predictions and cross-modal pseudo-labels. We justify the effectiveness by using different combinations of these components as in Tab. 2. As shown from experiments No. 1 to No. 6, disabling any part of Latte causes performance decreases of more than 1.6% relatively on the cross-modal prediction (xM), which proves the effectiveness of each component we included. A special case is No. 5 on U-to-S, where disabling ST $\mathcal{L}_{23}^{\mathrm{xM}}$ only leads to a minor 0.1% drop in accuracy. This is mainly because the 2D performance is already close to saturation (37.4 vs. 38.7 of Oracle TTA) without the guidance of 3D predictions. In this case, introducing 3D information to 2D is less effective compared to the other benchmarks.

Besides optimization objects, we further testify two important components of Latte, including ST-entropy-based weighting in Eq. (10) and Eq. (13) as well as percentile filtering in Eq. (8). Specifically, disabling percentile filtering as in No.



(a) Ablation studies of various ST voxel sizes.



(b) Ablation studies of filtering percentile α in Eq. (8).

Fig. 4: Sensitivity analysis about the ST voxel size and confident percentile α.

7 leads to a relative accuracy drop of more than 0.5% across all benchmarks while utilizing point-wise entropy instead of our ST entropy as in No. 8 degenerates the performance by more than 2.0% relatively, which justifies that ST entropy can better estimate the prediction reliability compared to the point-wise entropy.

Different frame aggregation mechanisms. One of the core designs of Latte is the slide window frame aggregation, which is designed to capture temporally local correspondences and estimate their prediction consistency. To justify the aforementioned claim, we compare our slide window aggregation with two aggregation variants as shown in Fig. 3. Specifically, variant (b) regards all student prediction frames rather a single frame within the time window as the query in Eq. (5), while variant (c) further utilizes a non-overlapping sliding strategy. According to Fig. 3, our slide window aggregation surpasses the other two aggregation variants under all sizes of time windows across three benchmarks, which justifies the effectiveness of our method. The inferior performance of variants (b) and (c) could be due to two factors. Firstly, aggregating student predictions could be an inferior option since student predictions tend to be less stable, and aggregating multiple frames can therefore lead to noisy queries for cross-modal learning. Secondly, non-overlapping aggregation can result in a miss-checking area on the boundary of each time window, which leads to unrepresentative entropy estimation for some ST voxels. Another observation is that utilizing a smaller slide window can usually yield better performance since temporally local consistency can be captured. This justifies the effectiveness of estimating prediction reliability through temporally local consistency.

Parameter sensitivity analysis. To testify whether Latte is sensitive towards hyper-parameter settings, we conduct sensitivity analysis on Latte across three benchmarks. Here we illustrate the sensitivity analysis of two hyper-parameters of Latte, including ST voxel size and filtering percentile α as shown in Fig. 4. In terms of ST voxel sizes, one can tell from Fig. 4a that all benchmarks share the same optimal voxel size as 0.2, where performance drops as the voxel size

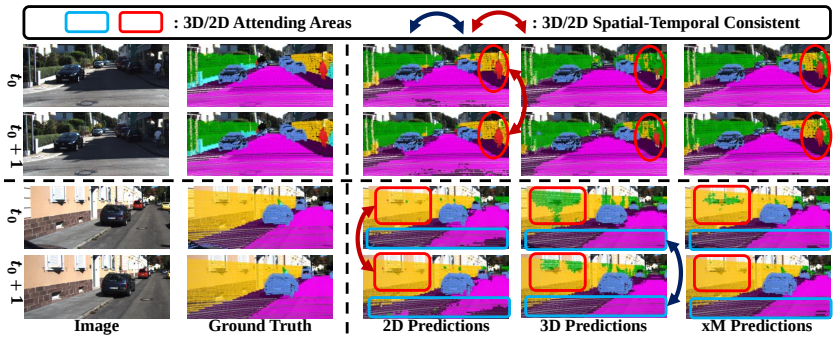


Fig. 5: Qualitative results of Latte on S-to-S. Here we visualize modality-specific predictions and cross-modal predictions from Latte during two sets of consecutive frames.

becomes too large or too small. This is mainly because large voxels could cause boundary ambiguity while small voxels contain insufficient representative points for evaluation. As for filtering percentile α, non-trivial improvement can be observed on U-to-S and A-to-S when $\alpha \geq 0.8$, while a lower α begins to discard more confident ST voxels leading to a slight drop in accuracy. On S-to-S, Latte without filtering (*i.e.*, No. 7 in Tab. 2) performs slightly better on cross-modal predictions with a minor gap of 0.1%, while it can also be observed that setting α as 0.9 leads to an improvement of 0.3% on both modality-specific predictions, which justifies the overall effectiveness of our filtering procedures.

Qualitative results. Our main motivation is utilizing prediction consistency of spatial-temporal correspondences to estimate modality reliability and achieve better cross-modal attending. As shown in Fig. 5, Latte can effectively attend the modality with more consistent predictions in spatial-temporal correspondences and therefore improve the cross-modal prediction consistency across time. For instance, Latte successfully attends to the stable 2D predictions pedestrian in the upper set of consecutive frames, suppressing the spatial-temporal inconsistency from 3D predictions. Similar observations can also be found in the lower sample, where the more consistent 2D predictions of buildings in the red rectangle and 3D predictions of roads in the blue rectangle prevail. *More qualitative comparisons with previous SOTA methods are presented in our appendix.*

5 Conclusion

In this paper, we propose a novel MM-TTA method called Latte, which leverages prediction consistency of spatial-temporal correspondences in MM-TTA for 3D segmentation. Latte utilizes a slide window frame aggregation to extract ST voxels and estimates temporally local consistency by the ST entropy of each modality. By attending to the modality with more consistency, Latte can achieve stable improvement on challenging MM-TTA scenarios. Despite its effectiveness, Latte struggles to rectify consistently wrong predictions across time, which is challenging and worth further investigation in the future.

Acknowledgment

This research is supported by the National Research Foundation, Singapore, under the NRF Medium Sized Centre for Advanced Robotics Technology Innovation (CARTIN). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

1. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9297–9307 (2019)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631 (2020)
3. Cao, H., Xu, Y., Yang, J., Yin, P., Yuan, S., Xie, L.: Mopa: Multi-modal prior aided domain adaptation for 3d semantic segmentation. arXiv preprint arXiv:2309.11839 (2023)
4. Cao, H., Xu, Y., Yang, J., Yin, P., Yuan, S., Xie, L.: Multi-modal continual test-time adaptation for 3d semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 18809–18819 (October 2023)
5. Chen, R., Liu, Y., Kong, L., Zhu, X., Ma, Y., Li, Y., Hou, Y., Qiao, Y., Wang, W.: Clip2scene: Towards label-efficient 3d scene understanding by clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7020–7030 (2023)
6. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3075–3084 (2019)
7. Fan, H., Yang, Y., Kankanhalli, M.: Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14204–14213 (2021)
8. Fan, H., Yu, X., Ding, Y., Yang, Y., Kankanhalli, M.: Pstnet: Point spatio-temporal convolution on point cloud sequences. arXiv preprint arXiv:2205.13713 (2022)
9. Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., Dietmayer, K.: Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems **22**(3), 1341–1360 (2020)
10. Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A.S., Hauswald, L., Pham, V.H., Mühlegg, M., Dorn, S., et al.: A2d2: Audi autonomous driving dataset. arXiv preprint arXiv:2004.06320 (2020)
11. Gong, T., Jeong, J., Kim, T., Kim, Y., Shin, J., Lee, S.J.: Robust continual test-time adaptation: Instance-aware bn and prediction-balanced memory. arXiv preprint arXiv:2208.05117 (2022)
12. Goyal, S., Sun, M., Raghunathan, A., Kolter, J.Z.: Test time adaptation via conjugate pseudo-labels. In: Advances in Neural Information Processing Systems (2022)

13. Graham, B.: Sparse 3d convolutional neural networks. arXiv preprint arXiv:1505.02890 (2015)
14. Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M.: Deep learning for 3d point clouds: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(12), 4338–4364 (2020)
15. Huang, S., Gojcic, Z., Huang, J., Wieser, A., Schindler, K.: Dynamic 3d scene analysis by point cloud accumulation. In: European Conference on Computer Vision. pp. 674–690. Springer (2022)
16. Jaritz, M., Vu, T.H., De Charette, R., Wirbel, É., Pérez, P.: Cross-modal learning for domain adaptation in 3d semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 1533–1544 (2022)
17. Ji, X., Yuan, S., Yin, P., Xie, L.: Lio-gvm: an accurate, tightly-coupled lidar-inertial odometry with gaussian voxel map. IEEE Robotics and Automation Letters (2024)
18. Li, M., Zhang, Y., Xie, Y., Gao, Z., Li, C., Zhang, Z., Qu, Y.: Cross-domain and cross-modal knowledge distillation in domain adaptation for 3d semantic segmentation. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3829–3837 (2022)
19. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International Conference on Machine Learning. pp. 6028–6039. PMLR (2020)
20. Liu, W., Luo, Z., Cai, Y., Yu, Y., Ke, Y., Junior, J.M., Gonçalves, W.N., Li, J.: Adversarial unsupervised domain adaptation for 3d semantic segmentation with multi-modal learning. ISPRS Journal of Photogrammetry and Remote Sensing **176**, 211–221 (2021)
21. Liu, Y., Kothari, P., Van Delft, B., Bellot-Gurlet, B., Mordan, T., Alahi, A.: Ttt++: When does self-supervised test-time training fail or thrive? Advances in Neural Information Processing Systems **34**, 21808–21820 (2021)
22. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: International Conference on Machine Learning. pp. 16888–16905. PMLR (2022)
23. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. arXiv preprint arXiv:2302.12400 (2023)
24. Peng, D., Lei, Y., Li, W., Zhang, P., Guo, Y.: Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7108–7117 (2021)
25. Piergiovanni, A., Casser, V., Ryoo, M.S., Angelova, A.: 4d-net for learned multi-modal alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15435–15445 (2021)
26. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660 (2017)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
28. Saltori, C., Krivosheev, E., Lathuiliére, S., Sebe, N., Galasso, F., Fiameni, G., Ricci, E., Poiesi, F.: Gipso: Geometrically informed propagation for online adaptation in 3d lidar segmentation. In: European Conference on Computer Vision. pp. 567–585. Springer (2022)

29. Shin, I., Tsai, Y.H., Zhuang, B., Schulter, S., Liu, B., Garg, S., Kweon, I.S., Yoon, K.J.: Mm-tta: multi-modal test-time adaptation for 3d semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16928–16937 (2022)
30. Simons, C., Raychaudhuri, D.S., Ahmed, S.M., You, S., Karydis, K., Roy-Chowdhury, A.K.: Summit: Source-free adaptation of uni-modal models to multi-modal targets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1239–1249 (October 2023)
31. Song, J., Lee, J., Kweon, I.S., Choi, S.: Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11920–11929 (2023)
32. Su, Y., Xu, X., Li, T., Jia, K.: Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering regularized self-training. arXiv preprint arXiv:2303.10856 (2023)
33. Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S.: Searching efficient 3d architectures with sparse point-voxel convolution. In: European Conference on Computer Vision. pp. 685–702. Springer (2020)
34. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in Neural Information Processing Systems **30** (2017)
35. Vedula, S., Baker, S., Rander, P., Collins, R., Kanade, T.: Three-dimensional scene flow. In: Proceedings of the Seventh IEEE International Conference on Computer Vision. vol. 2, pp. 722–729. IEEE (1999)
36. Vizzo, I., Guadagnino, T., Mersch, B., Wiesmann, L., Behley, J., Stachniss, C.: Kiss-icp: In defense of point-to-point icp–simple, accurate, and robust registration if done the right way. IEEE Robotics and Automation Letters **8**(2), 1029–1036 (2023)
37. Vogel, C., Schindler, K., Roth, S.: 3d scene flow estimation with a rigid motion prior. In: 2011 International Conference on Computer Vision. pp. 1291–1298. IEEE (2011)
38. Vogel, C., Schindler, K., Roth, S.: Piecewise rigid scene flow. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1377–1384 (2013)
39. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=uXl3bZLkr3c
40. Wang, J.K., Wibisono, A.: Towards understanding gd with hard and conjugate pseudo-labels for test-time adaptation. arXiv preprint arXiv:2210.10019 (2022)
41. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7201–7211 (2022)
42. Wyner, A.: Recent results in the shannon theory. IEEE Transactions on information Theory **20**(1), 2–10 (1974)
43. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
44. Xing, B., Ying, X., Wang, R., Yang, J., Chen, T.: Cross-modal contrastive learning for domain adaptation in 3d semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2974–2982 (2023)
45. Xu, J., Miao, Z., Zhang, D., Pan, H., Liu, K., Hao, P., Zhu, J., Sun, Z., Li, H., Zhan, X.: Int: Towards infinite-frames 3d detection with an efficient framework. In: European Conference on Computer Vision. pp. 193–209. Springer (2022)

46. Yin, P., Cao, H., Nguyen, T.M., Yuan, S., Zhang, S., Liu, K., Xie, L.: Outram: One-shot global localization via triangulated scene graph and global outlier pruning. arXiv preprint arXiv:2309.08914 (2023)
47. Yuan, L., Xie, B., Li, S.: Robust test-time adaptation in dynamic scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15922–15932 (2023)
48. Zhang, M., Levine, S., Finn, C.: Memo: Test time robustness via adaptation and augmentation. Advances in Neural Information Processing Systems **35**, 38629–38642 (2022)