Multi-Modal Continual Test-Time Adaptation for 3D Semantic Segmentation

Haozhi Cao[1]* Yuecong Xu[2]* Jianfei Yang[1]† Pengyu Yin[1] Shenghai Yuan[1] Lihua Xie[1]
[1]Nanyang Technological University [2]Institute for Infocomm Research, A*STAR, Singapore
{haozhi002, xuyu0014, yang0478, pengyu001, syuan003}@e.ntu.edu.sg elhxie@ntu.edu.sg

Abstract

Continual Test-Time Adaptation (CTTA) generalizes conventional Test-Time Adaptation (TTA) by assuming that the target domain is dynamic over time rather than stationary. In this paper, we explore Multi-Modal Continual Test-Time Adaptation (MM-CTTA) as a new extension of CTTA for 3D semantic segmentation. The key to MM-CTTA is to adaptively attend to the reliable modality while avoiding catastrophic forgetting during continual domain shifts, which is out of the capability of previous TTA or CTTA methods. To fulfill this gap, we propose an MM-CTTA method called Continual Cross-Modal Adaptive Clustering (CoMAC) that addresses this task from two perspectives. On one hand, we propose an adaptive dual-stage mechanism to generate reliable cross-modal predictions by attending to the reliable modality based on the class-wise feature-centroid distance in the latent space. On the other hand, to perform test-time adaptation without catastrophic forgetting, we design class-wise momentum queues that capture confident target features for adaptation while stochastically restoring pseudo-source features to revisit source knowledge. We further introduce two new benchmarks to facilitate the exploration of MM-CTTA in the future. Our experimental results show that our method achieves state-of-the-art performance on both benchmarks. Visit our project website at https://sites.google.com/view/mmcotta.

1. Introduction

Test-Time Adaptation (TTA) proposes a realistic domain adaptation scenario, which adapts pre-trained models to the target domain during the testing process. Unlike previous Unsupervised Domain Adaptation (UDA) [17, 45, 52, 47], TTA performs adaptation online without accessing the data from the source domain. Typical TTA [40, 35, 36] methods assume a stationary target domain, while real-world scenarios are dynamic over time. To fulfill this gap, a previous work [42] proposes a general extension of TTA named Continual Test-Time Adaptation (CTTA) which assumes that the target domain is continually changing rather than static. Compared to TTA, CTTA is more challenging since the performance is easily affected by error accumulation [6] and catastrophic forgetting [27] during the continual adaptation.

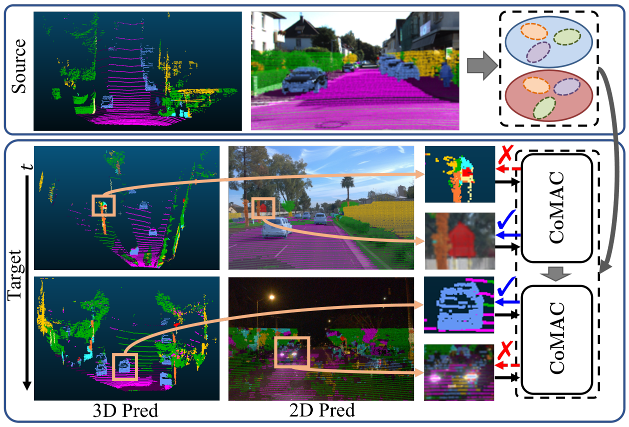

Figure 1. Illustration of how continual domain shifts affect multi-modal segmentation and our method. Unlike the source domain where predictions from both modalities are reliable, the reliability of each modality varies in MM-CTTA due to different domain shifts. CoMAC tackles MM-CTTA by attending to the reliable modality (✓) rather than the noisy one (✗). Meanwhile, source knowledge is stochastically revisited to avoid catastrophic forgetting. Figure best viewed in color and zoomed in.

For 3D semantic segmentation, multi-modal sensors are frequently leveraged in different tasks, such as scene understanding [4, 28] and semantic map construction [29, 43]. For some specific applications like semantic-based localization [8] and autonomous driving [46, 44], multi-modal information is the key to robust performance under adverse conditions. However, their collaboration has been proven to be sensitive toward domain drifts [1, 35]. In real-world scenarios, such collaboration deterioration could be more severe considering that the target domain is continually changing as in CTTA (e.g., the operating environments of 24/7 AGVs are continually changing due to altering weather or illumination conditions). Hence, it is essential for multi-

*Equal contributions.
†Corresponding author.

modal networks to adapt to the dynamic target domain in an online manner. In this work, we aim to study Multi-Modal Continual Test-Time Adaptation (MM-CTTA) for 3D semantic segmentation, where networks are continually adapted to a changing target domain taking 3D point clouds and 2D images as input without accessing the source data.

Intuitively, one can address MM-CTTA by utilizing CTTA [30, 42] or TTA [3, 38] methods on 2D and 3D networks separately. However, this simple extension can not achieve satisfactory performance since it cannot correctly attend to the reliable modality for adaptation when others suffer from severe domain shifts. Take Fig. 1 as an example: predictions from the 2D image are more accurate at the beginning of the target domain, while 2D results become unreliable and 3D predictions prevail as the illumination level significantly changes over time. Although previous CTTA or TTA methods have attempted to mitigate this intra-modal prediction noise by augmentations [42] or entropy minimization [40, 30], the domain drift in Fig. 1 is too severe to be effectively rectified by image input itself, leading to inevitable error accumulation. Previous works have proposed different cross-modal fusion methods to mitigate the effect of the noisy modality during adaptation, such as cross-modal consistency [18] for UDA or pseudo-label fusion based on student-teacher consistency [35] for TTA. However, their methods rely on the premise that the target domain is static, and therefore suffer from catastrophic forgetting in MM-CTTA, leading to degenerated results.

For MM-CTTA, an ideal solution is to suppress the contribution from the noisy modality and attend more to the reliable one in an online manner. Typically, the reliability of prediction can be estimated based on its corresponding feature location in the latent space. A prediction with features close to the centroid is more likely to be reliable, suffering less from the domain shift and vice versa. On the other hand, to ensure the validity of centroid-based reliability estimation, class-wise centroids should actively adapt to the continually changing target domain while stochastically revisiting the source knowledge to avoid catastrophic forgetting. To this end, we propose an MM-CTTA method called **C**ontinual Cr**o**ss-**M**odal **A**daptive **C**lustering (CoMAC) for 3D semantic segmentation. CoMAC consists of three main modules: (i) Intra-Modal Prediction Aggregation (iMPA), (ii) Inter-Modal Pseudo-Label Fusion (xMPF), and (iii) Class-Wise Momentum Queues (CMQs). On one hand, the proposed iMPA and xMPF are utilized to suppress prediction noise based on the centroid-based reliability estimation from intra-modal and inter-modal perspectives, respectively. On the other hand, CMQs are designed to actively adapt class-wise centroids for iMPA and xMPF while avoiding catastrophic forgetting. Additionally, a class-wise contrastive loss is introduced to regularize the extracted features from drifting too far from centroids.

In summary, our main contributions are three-fold. (i) We explore a new task MM-CTTA where multi-modal input is utilized to perform continual test-time adaptation for 3D semantic segmentation and propose an effective method named CoMAC to leverage multi-modal information for MM-CTTA. (ii) We propose iMPA and xMPF to generate accurate cross-modal pseudo-labels by attending to a reliable modality. CMQs are introduced to actively adapt to the target domain without catastrophic forgetting. (iii) We introduce two 3D semantic segmentation benchmarks to facilitate the future exploration of MM-CTTA. Extensive experiments show that our method achieves state-of-the-art performance, outperforming previous methods significantly by 6.9% on the challenging benchmark.

2. Related Works

Test-Time Adaptation Describing a more realistic adaptation scenario, Test-Time Adaptation (TTA) is receiving more and more attention. Different from previous Unsupervised Domain Adaptation (UDA), TTA forbids access to raw source data and adapts the source pre-trained model during test time. As one of the primary works, TENT [40] highlights the fully test-time setting and proposes to update the batch normalization layers by entropy minimization. The following works mainly address TTA by aligning batch normalization statistics [30, 23, 51], self-training with pseudo labeling [13, 41], feature alignment [25], or augmentation invariance [50, 21]. The aforementioned TTA methods strictly follow the one-pass protocol as mentioned in [36], where networks immediately infer each sample in an online manner and forbid multiple training epochs. On the other hand, some previous works [22, 10, 7, 48] follow the multi-pass protocol by adapting the model for multiple epochs in an offline manner. As one of the primary works, Source Hypothesis Transfer (SHOT) [22] proposes to update only the encoder parameters and align source and target representation by entropy minimization and pseudo-labeling. While most existing TTA methods are proposed for image classification, some recent works [10, 20, 26] aim to explore TTA for image semantic segmentation. Specifically, [10] proposes to minimize the prediction entropy while maximizing its robustness toward feature noise. [20] divides the TTA problem into source domain generalization and target domain adaptation. [26] proposes a dual attention distillation method to transfer contextual knowledge and patch-level pseudo labels for self-supervised learning.

Continual Test-Time Adaptation The definition of Continual Test-Time Adaptation (CTTA) is first proposed in CoTTA[42], which aims to adapt the model to continually changing target domains in an online manner without any source data. Specifically, CoTTA proposes to use the moving teacher model and augmentation-average predictions for noise suppression and the model stochastic restoration to avoid catastrophic forgetting. Following the scheme of CoTTA, some recent works [11, 12, 30] have addressed CTTA from different perspectives. Specifically, [12] leverages the temporal correlations of streamed input by reservoir sampling and instance-aware batch normalization. [11] proposes domain-specific prompts and domain-agnostic prompts to preserve domain-specific and domain-shared knowledge, respectively. EATA [30] performs adaptation on non-redundant samples for an efficient update.

Multi-Modal Adaptation for Segmentation Thanks to the emerging multi-modal datasets [2, 5, 37, 33, 49], recent works start to explore how to leverage multi-modal information between 2D images and 3D point clouds to perform domain adaptation under different settings. xMUDA [18] proposes the first Multi-Modal Unsupervised Domain Adaptation (MM-UDA) method for 3D semantic segmentation. Specifically, it leverages the cross-modal prediction consistency by minimizing the prediction discrepancy from additional classifiers. Following the scheme of cross-modal learning, DsCML [31] designs a deformable mapping between pixel-point correlation for MM-UDA while [24] introduces adversarial training to mitigate domain discrepancy. In addition to UDA settings, [35] proposes the first multi-modal test-time adaptation for semantic segmentation which generates intra-modal and inter-modal pseudo labels by attending to the one with more consistent predictions across student and teacher models.

In this work, we propose MM-CTTA as a new extension of CTTA with a specific method CoMAC. While our proposed dual-stage modules (i.e., iMPA and xMPA) and CMQs share some similar merit with previous TTA methods [35, 36], our method is explicitly designed for MM-CTTA. Unlike previous work [35] measuring prediction reliability by teacher-student consistency, our reliability measurement relies on the feature-centroid distance to encourage feature clustering around the adapting centroids. Different from [36] utilizing queues to measure target distribution, the objective of our CMQs is to actively update class-wise centroids to ensure the validity of iMPA and xMPA while avoiding catastrophic forgetting.

3. Proposed Method

Problem definition and notations. At timestamp t, the 2D image $x^{\text{2D}}_{\mathcal{T},t} \in \mathbb{R}^{H\times W\times 3}$ and the 3D point cloud $x^{\text{3D}}_{\mathcal{T},t} \in \mathbb{R}^{N\times 4}$ are observed in the target domain $\mathcal{T}$, where N denotes the number of 3D points located in the camera FOV. Modal-specific pre-trained networks $\phi^m_{\theta,t}(\cdot) = f^m_t(g^m_t(\cdot)), m \in \{\text{2D}, \text{3D}\}$ consisting of the feature extractor f^m_t and the classifier g^m_t are used to predict the semantic labels for each point. Inspired by the fact that moving average models can provide more stable predictions [39], we utilize a fast student network $\phi^m_\theta(\cdot) = f^m_\theta(g^m_\theta(\cdot))$ and a slow teacher network $\phi^m_{\theta'}(\cdot) = f^m_{\theta'}(g^m_{\theta'}(\cdot))$ for each modality similar to previous works [35, 42]. Given $x^m_{\mathcal{T},t}$ as input, the features extracted by f^m_t is denoted as $z^m_{\mathcal{T},t} \in \mathcal{R}^{N\times F^m}$, where F^m denotes the channel number. Here we adopt the same projection protocol of previous works [18, 35] for cross-modal alignment, which projects features from the 2D branch back to 3D points, resulting in the 2D feature $z^{\text{2D}}_{\mathcal{T},t}$ of shape $N \times F^{\text{2D}}$. By default, the subscript of the target domain $\mathcal{T}$ and the timestamp t are omitted for clarity. Given the multi-modal input $x^{\text{2D}}_{\mathcal{T},t}$, $x^{\text{3D}}_{\mathcal{T},t}$, the goal of MM-CTTA is to output reliable cross-modal predictions by continuously adapting to the changing target domain. In this work, we interpret the core of MM-CTTA as two-fold: (i) attending to the reliable modality for noisy suppression, and (ii) revisiting source knowledge to prevent catastrophic forgetting.

To this end, we propose Continual Cross-Modal Adaptive Clustering (CoMAC) to tackle MM-CTTA from the aforementioned perspectives. Specifically, as shown in Fig. 2, the class-wise centroids are initialized from the source domain and pseudo-source features are randomly sampled around the centroids as source knowledge (Sec. 3.1). Given the raw and augmentation-average predictions from the teacher models as input, Intra-Modal Prediction Aggregation (iMPA) generates stable intra-modal predictions as their weighted sum based on their feature distance to the class-wise centroids (Sec. 3.2). Inter-Modal Pseudo-Label Fusion (xMPF) then combines the intra-modal predictions from each modality based on their reliability and output cross-modal pseudo-labels as supervision signals for student networks (Sec. 3.3). Given raw samples as input, confident target features from student networks are utilized to update Class-Wise Momentum Queues (CMQs), while pseudo-source features are stochastically restored to avoid catastrophic forgetting (Sec. 3.4).

3.1. Source Models and Class-Wise Centroids

To effectively avoid catastrophic forgetting and inspired by previous TTA methods [9, 36] which preserve trivial source domain information, we utilize pseudo-source features sampled around source offline feature centroids as source representatives. The source offline centroids and pseudo-source features are treated as the prior knowledge from the source domain, which plays an essential role in preventing catastrophic forgetting detailed in Sec. 3.4. Specifically, given a specific semantic category k, the corresponding source offline centroid is modeled as Gaussian distribution, and pseudo-source features are denoted as:

$$C^{m,k}_{src} = \mathcal{N}(\mu^{m,k}_{src}, \sigma^{m,k}_{src}), \tag{1}$$

$$\mathcal{Z}^{m,k}_{src} = \{z^{m,k}_{src,i} \sim C^{m,k}_{src} \mid i \in [1, N_q]\}, \tag{2}$$

where N_q is the number of pseudo-source features. $\mu^{m,k}_{src}$ and $\sigma^{m,k}_{src}$ denote the mean and standard deviation of normalized source features from the category k, respectively.

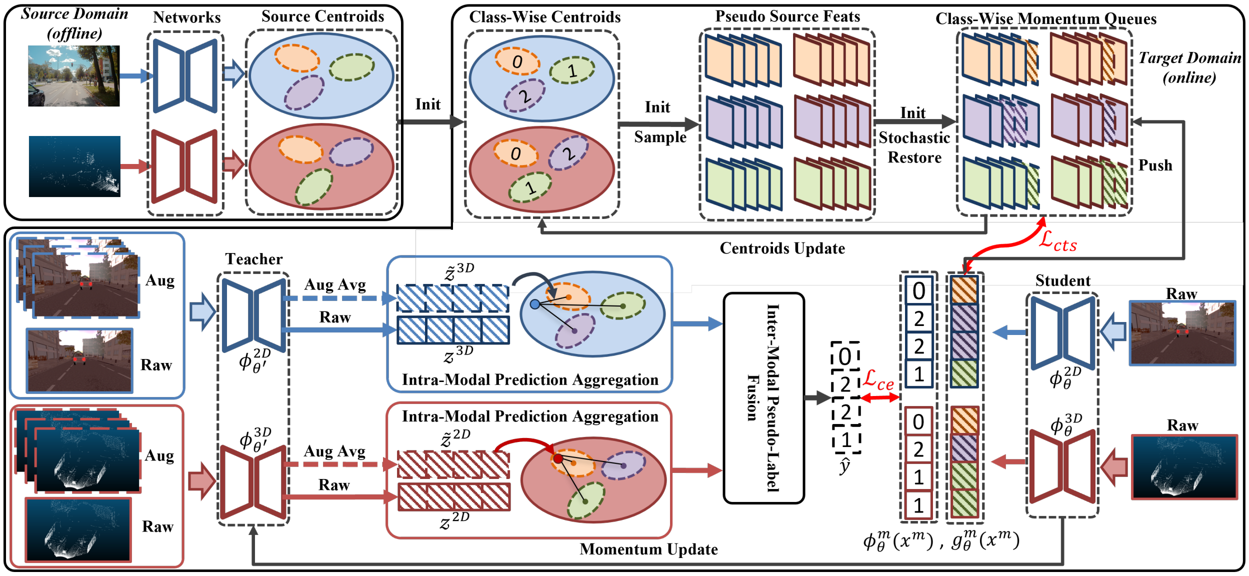

Figure 2. The main structure of our proposed method. From the source domain, we preserve the source centroids and the pre-trained model for each modality as the initialization of the class-wise centroids and all the models in the target domain, respectively. To generate reliable intra-modal predictions, Intra-Modal Prediction Aggregation (iMPA) attends to the reliable prediction whose feature shares a closer distance with the class-wise centroids in an intra-modal manner. Inter-Modal Pseudo-Label Fusion (xMPF) then fuses the intra-modal predictions by estimating the reliability of predictions from each modality for noise suppression. Class-Wise Momentum Queues (CMQs) are designed to achieve a good balance between target domain adaptation and source knowledge preservation.

Here we use the features generated by the pre-trained extractor $g^m_0(\cdot)$ for the centroid construction.

3.2. Intra-Modal Prediction Aggregation

In MM-CTTA, the prediction from each modality could be unreliable due to continual domain shifts, causing severe noise to the multi-modal fusion. To mitigate the intra-modal noise, iMPA aims to generate reliable intra-modal predictions for each modality. Although previous methods [42, 38] regard the augmentation-average prediction as a more stable alternative, we argue that its superiority is not certain since test-time augmentations may introduce inductive bias to the prediction [34] (e.g., resizing image as in [42] could cause ambiguity to small-scale classes). Unlike the previous work [42], we propose an adaptive mechanism to fuse the raw and the augmentation-average prediction as the weighted sum based on their feature distance between class-wise centroids. Given the input sample x^m and its augmented variants $\{\widetilde{x}^m_i \mid i \in [1, N^m_{aug}]\}$, the features and predictions are extracted by the teacher model $\phi^m_{\theta'}$ as:

$$p(x^m) = f^m_{\theta'}(g^m_{\theta'}(x^m)) = f^m_{\theta'}(z^m), \tag{3}$$

$$\widetilde{p}(x^m) = \frac{1}{N^m_{aug}} \sum_{i=1}^{N^m_{aug}} f^m_{\theta'}(g^m_{\theta'}(\widetilde{x}^m_i)), \tag{4}$$

where $p(x^m)$ and $\widetilde{p}(x^m)$ represent the raw and the augmentation-average prediction, respectively. Both predictions are of shape $N \times N_c$, where N_C is the number of classes and N^m_{aug} denotes the number of augmentations. The normalized augmentation-average feature is computed as $\widetilde{z}^m = \sum_{i=1}^{N^m_{aug}} g^m(\widetilde{x}^m_i)/(\|g^m(\widetilde{x}^m_i)\| \cdot N^m_{aug})$. Subsequently, the weights for predictions $p(x^m), \widetilde{p}(x^m)$ are computed based on their corresponding feature distance to the class-wise centroid in a point-wise manner, so that each point of the intra-modal prediction $\hat{p}(x^m)$ can attend more to the confident prediction located closer to the cluster in the latent space. Taking a point (j) as an example, given its predictions $p_{(j)}(x^m)$, $\widetilde{p}_{(j)}(x^m)$ and features $z^m_{(j)}$, $\widetilde{z}^m_{(j)}$:

$$w^m_{(j)} = \frac{\exp(\langle \mu^{m,k}, z^m_{(j)} \rangle)}{\sum_{i \in N_c} \exp(\langle \mu^{m,i}, z^m_{(j)} \rangle)}, \tag{5}$$

$$\widetilde{w}^m_{(j)} = \frac{\exp(\langle \mu^{m,\widetilde{k}}, \widetilde{z}^m_{(j)} \rangle)}{\sum_{i \in N_c} \exp(\langle \mu^{m,i}, \widetilde{z}^m_{(j)} \rangle)}, \tag{6}$$

$$\hat{p}_{(j)}(x^m) = \frac{w^m_{(j)} p_{(j)(x^m)} + \widetilde{w}^m_{(j)} \widetilde{p}_{(j)}(x^m)}{w^m_{(j)} + \widetilde{w}^m_{(j)}}, \tag{7}$$

where $w^m_{(j)}$, $\widetilde{w}^m_{(j)}$ are termed as the raw weight and augmentation-average weight, respectively. k, $\widetilde{k}$ are the largest element locations of $p_{(j)}(x^m)$, $\widetilde{p}_{(j)}(x^m)$. $\mu^{m,k}$ denotes the mean of class k centroid, which is actively updated as shown in Sec. 3.4. $\langle a, b \rangle$ represents the inner product of vectors a and b. Our empirical results show that this adaptive fusion between the raw and augmentation-average predictions results in more stable intra-modal predictions.

3.3. Inter-Modal Pseudo-Label Fusion

Given the output of iMPA, the goal of xMPF is to generate pseudo-labels by estimating their reliability in a cross-modal manner. Since domain shifts can affect the reliability of each modality variously (e.g., images in day and night), the cross-modal pseudo-label should adaptively attend to the reliable modality for noise suppression. Motivated by this idea, the proposed xMPF generates the cross-modal pseudo-labels as the weighted sum of intra-modal predictions from iMPA. Specifically, taking point (j) as an example, the inter-modal weight $\hat{w}^m_{(j)}$ is computed as the summation of both weights if their corresponding predictions indicate the same class, or the maximum one otherwise. This can be viewed as an implicit way to encourage point-wise prediction consistency for each modality. Formally, the inter-modal weight $\hat{w}^m_{(j)}$ is computed as follows:

$$\hat{w}^m_{(j)} = \begin{cases} w^m_{(j)} + \widetilde{w}^m_{(j)}, & \text{if } k = \widetilde{k} \\ \max(w^m_{(j)}, \widetilde{w}^m_{(j)}), & \text{if } k \neq \widetilde{k} \end{cases}. \tag{8}$$

The cross-modal prediction is then computed as the weighted sum in a cross-modal manner:

$$\hat{p}^{\text{xM}}_{(j)}(x) = \frac{\hat{w}^{\text{2D}}_{(j)} \hat{p}^{\text{2D}}_{(j)}(x) + \hat{w}^{\text{3D}}_{(j)} \hat{p}^{\text{3D}}_{(j)}(x)}{\hat{w}^{\text{2D}}_{(j)} + \hat{w}^{\text{3D}}_{(j)}}, \tag{9}$$

$$\hat{y}_{(j)} = \arg\max_{k \in N_c} (\hat{p}^{\text{xM}}_{(j)}(x)^{(k)}), \tag{10}$$

where $\hat{p}^{\text{xM}}_{(j)}(x)$, $\hat{y}_{(j)}$ denote the cross-modal prediction and pseudo-label, respectively. The cross-modal pseudo-label can therefore attend to the more reliable modality with confident intra-modal predictions.

3.4. Class-Wise Moving Queues

Both iMPA and xMPF greatly depend on the quality of class-wise centroids. Initialized from the source offline centroids as in Eq. 1, the class-wise centroids should continually adapt to the target domain to ensure the validity of iMPA and xMPF. Meanwhile, the source knowledge should be occasionally played back during the centroid adaptation so that it can be revisited during the reliability estimation in iMPA and xMPF without catastrophic forgetting. To achieve a balance between adaptation and source knowledge preservation, we propose CMQs which utilize momentum queues [15] to actively estimate feature clustering of the target domain by capturing confident target features while stochastically restoring pseudo-source features for each modality. Specifically, the CMQ of the class k is initialized by the pseudo-source features $\mathcal{Z}^{m,k}_{src}$ as in Eq. 2. For each iteration, taking the raw sample as input, the normalized point features from the student network ϕ^m_θ with confident predictions are preserved by threshold-based filtering. Given the CMQ of class k from the previous step denoted as $\hat{\mathcal{Q}}^{m,k}_{t-1}$, the CMQ of current step is updated as:

$$\mathcal{Z}^{m,k}_{cf} = \mathbb{1}^k_{cf}[g^m_\theta(x^m)/\|g^m_\theta(x^m)\|], \tag{11}$$

$$\widetilde{\mathcal{Q}}^{m,k}_t = \Phi(\hat{\mathcal{Q}}^{m,k}_{t-1}, \mathcal{Z}^{m,k}_{cf}), \tag{12}$$

$$\hat{\mathcal{Q}}^{m,k}_t = \begin{cases} \widetilde{\mathcal{Q}}^{m,k}_t, & \text{if } \gamma_t > p_{rs} \\ \Phi(\hat{\mathcal{Q}}^{m,k}_{t-1}, \mathbb{1}^k_{rs}\mathcal{Z}^{m,k}_{src}), & \text{if } \gamma_t \leq p_{rs} \end{cases}, \tag{13}$$

where $\mathbb{1}^k_{cf}$ is the confidence index for class k where $\mathbb{1}^k_{cf}\phi^m_\theta(x^m)^{(k)} \geq \tau_{cf}$ and $\mathbb{1}^k_{rs}$ denotes the source restoring index which randomly samples N_{enq} pseudo-source features from $\mathcal{Z}^{m,k}_{src}$. $\Phi(a, b)$ denotes the operation of enqueuing b in a as in [15]. γ_t is a restoring flag uniformly sampled from $[0, 1]$ at timestamp t and p_{rs} denotes the hyper-parameter of restoring probability. Different from the image classification problem which can capture all confident samples [36], the number of confident points from each class can be huge (i.e., more than 1,000), which is too aggressive and expensive to enqueue all points. Therefore, we limit the index number of $\mathbb{1}^k_{cf}$ by an upper bound N_{enq} through selecting features with the top N_{enq} confident predictions.

For each optimization step, the mean of the class-wise centroid for the next step $\mu^{m,k}_{t+1}$ is updated as the average of CMQ of the current timestamp $\hat{\mathcal{Q}}^{m,k}_t$ so that the class-wise centroids can adapt to the feature clustering of target domain without catastrophic forgetting. Additionally, we propose a class-wise contrastive loss modified from [19] as a regularizer so that the target confident features can revisit source knowledge through clustering around the class-wise centroids. Specifically, given any confident target feature from Eq. 11 as the anchor, the positive samples are the features from the CMQ that shares the same modality and semantic class, denoted as $P(k) = \{\hat{q}_p \mid \hat{q}_p \in \hat{\mathcal{Q}}^{m,k}_t\}$. The negative samples are the features from the CMQs of the same modality but different classes denoted as $A(k) = \{\hat{q}_n \mid \hat{q}_n \in \hat{\mathcal{Q}}^{m,a}_t, \forall a \neq k \bigcap a \in N_c\}$. The class-wise contrastive loss $\mathcal{L}^m_{cts}$ is computed as:

$$\mathcal{L}^m_{cts} = \sum_{i \in |\mathcal{Z}^{m,k}_{cf}|} \frac{-1}{|P(k)|} \sum_{q_p \in P(k)} \frac{\exp(\hat{z}^{m,k}_{t,(i)} \cdot q_p)}{\sum_{q_n \in N(k)} \exp(\hat{z}^{m,k}_{t,(i)} \cdot q_n)}. \tag{14}$$

3.5. Main Structure and Optimization

The main structure of our proposed method is shown in Fig. 2. Given the multi-modal input x^m, the teacher model is used to infer x^m and its augmented variants $\{\widetilde{x}^m_i | i \in [1, N^m_{aug}]\}$ and then generate reliable cross-modal pseudo-labels denoted as $\hat{y}$ through iMPA and xMPF. The student model is directly updated by minimizing the weighted sum

Algorithm 1: The proposed CoMAC

```
Init Model: Teacher $\phi^m_{\theta'}$, student $\phi^m_{\theta}$, $m \in \{2\text{D}, 3\text{D}\}$
Init CMQ: $\hat{\mathcal{Q}}^{m,k}_0 = \mathcal{Z}^{m,k}_{src}$, $m \in \{2\text{D}, 3\text{D}\}, k \in N_c$
for $t \in |\mathcal{X}_\mathcal{T}|$, $x = (x^{2D}, x^{3D}) \in \mathcal{X}_\mathcal{T}$ do
    for $m \in \{2D, 3D\}$ do
        1. Augmented input to get $\{\tilde{x}^m_i \mid i \in N^m_{aug}\}$
        2. Get predictions and features with $\phi^m_{\theta'}$ by Eq. 3-4
        3. Get intra-modal $\hat{p}^m(x^m)$ by iMPA through Eq. 5-7
    end
    4. Get inter-modal $\hat{y}$ by xMPF through Eq. 8-10
    5. Update CMQs and compute $\mathcal{L}^m_{cts}$ by Eq. 11-14
    6. Update $\phi^m_\theta$ by Eq. 16 and $\phi^m_{\theta'}$ by Eq. 17
end
```

of the standard cross-entropy loss between its prediction and the pseudo-label $\hat{y}$, and the contrastive loss in Eq. 14. The teacher model is updated as the exponential moving average of the student's weights as in previous works [35, 42]:

$$\mathcal{L}^m_{ce} = \mathbb{1}_{\hat{y}} \log \phi^m(x^m), \tag{15}$$

$$\mathcal{L}_{total} = \sum_{m}^{\{2\text{D},3\text{D}\}} (\mathcal{L}^m_{ce} + \lambda_{cts}\mathcal{L}^m_{cts}), \tag{16}$$

$$\theta'_{t+1} = (1 - \lambda_s)\theta + \lambda_s \theta', \tag{17}$$

where λ_{cts} and λ_s are the hyper-parameters denoting the coefficient of the contrastive loss and the momentum factor. Our adaptation process is summarized as Algorithm 1.

4. Experiments

In this section, we present our experimental results based on two new benchmarks. The details of our proposed benchmarks, baselines, and implementation are first introduced in Sec. 4.1. Sec. 4.2 presents our overall results and Sec. 4.3 justifies our method by extensive ablation studies.

4.1. Benchmarks and Settings

Proposed benchmarks. To evaluate the performance under MM-CTTA settings, we proposed two benchmarks in this work, including (i) **SemanticKITTI-to-Synthia (S-to-S)** and (ii) **SemanticKITTI-to-Waymo (S-to-W)**. Both benchmarks leverage SemanticKITTI [2] as the source-domain dataset, which utilizes a 0.7MP camera and a 64-line LiDAR. For the target domain, we utilize two different datasets including Synthia [33] and Waymo [37] thanks to the various environmental conditions they cover. The former shares a larger domain gap with more diverse seasons, weather, and illumination conditions while the latter is less challenging but closer to the real-world application. For S-to-S, it is constructed by a sequence of different videos from Synthia [33] without any repetition. Since Synthia includes depth images instead of 3D point clouds, we generate a simulated point cloud for each sample by randomly sampling depth pixels as 3D points following the structural characteristic of LiDAR. For each sequence, all samples are included for adaptation and we use "No.-Conditions" to indicate their corresponding video sequence in Synthia (i.e, "01-Spring" indicates video sequence 01 under Spring condition). Different from the previous CTTA benchmark [42] which only includes 4 different unique sequences, the proposed S-to-S consists of 12 different sequences recorded under various environmental conditions without repetition. By default, we organize the sequences following the order of seasons→illumination→weathers as a challenge ascending order. For S-to-W, we sort out the illumination conditions of each sequence in Waymo dataset [37] based on the sequence descriptions and divide all samples into three types, including Day (**D**), Dawn-Dusk (**DD**), and Night (**N**). Since the number of sequences of **D** is much larger than others, we further split the sequence of **D** based on their recording location, including Phoenix (**P**), San Francisco (**S**), and Others (**O**). Each sequence is separated in two halves to simulate the revisiting situation as in [42] but without repeated samples (i.e., "D-O-1" indicates the first half of the D-O sequence). Similar to S-to-S, S-to-W also follows a challenge ascending order. For both datasets, we adopt the images of the front-view camera similar to the source-domain dataset. We utilize a similar class mapping strategy as [18, 35] with slight modification since some classes are missing in the target domain. *More details are illustrated in the appendix.*

Baselines. Previous TTA, MM-TTA, and CTTA methods are evaluated on both S-to-S and S-to-M as baselines. For TTA methods, we compare with pseudo-labels with threshold filtering (Pslabel), TENT [40] and LAME [3], while xMUDA with pseudo-labels (xMUDA-pl) [18] and MMTTA [43] are regarded as the representatives of MM-TTA methods. CoTTA [42] and EATA [30] are the recent works that focus on CTTA problems. Specifically, xMUDA-pl is originally designed for UDA, whereas we modify it into a TTA version by discarding the source data during adaptation. Considering our CoMAC is the only MM-CTTA method, we implement an MM-CTTA version of MMTTA [35] integrated with model-based stochastic restoration in [42] for a fair comparison. We report the softmax-average mIoU of 2D and 3D predictions as results.

Implementation details. Following previous works [18, 35], we utilize UNet [32] with ResNet-34 [16] encoder as the 2D backbone and SCN [14] based on UNet as the 3D backbone for all baselines. The pre-training procedures on SemanticKITTI dataset follow [18]. For our method, we empirically choose resizing and z-rotation as the 2D and 3D augmentations following previous works [42, 18]. Specifically, we utilize multiple resizing factors of [0.5,

Time		t →													t →										
		SemanticKITTI-to-Synthia													SemanticKITTI-to-Waymo										
Methods	DA Type	*01-Spring*	*02-Summer*	*04-Fall*	*05-Winter*	*01-Dawn*	*02-Night*	*04-Sunset*	*05-W-night*	*01-Fog*	*02-S-Rain*	*04-R-Night*	*05-Rain*	Avg	*D-O-1*	*D-P-1*	*D-S-1*	*DD-1*	*N-1*	*D-O-2*	*D-P-2*	*D-S-2*	*DD-2*	*N-2*	Avg
Source	-	28.3	30.7	37.7	27.3	23.1	27.3	38.2	26.5	27.0	21.6	27.8	16.4	27.7	33.2	36.9	29.5	28.3	14.0	32.8	38.0	30.4	28.4	14.6	28.6
Pslabel	TTA	29.6	27.1	36.4	28.4	21.4	26.6	32.3	26.1	27.4	21.8	25.0	18.0	26.7	39.1	41.9	37.0	36.8	25.7	38.5	43.7	37.6	37.6	24.6	36.9
TENT [40]	TTA	**35.4**	24.9	24.9	20.7	18.7	16.6	16.0	13.6	13.6	11.5	8.4	9.1	17.8	33.6	22.3	18.8	16.9	10.8	15.7	12.9	9.8	11.3	10.6	16.3
LAME [3]	TTA	14.0	12.4	17.3	13.0	12.6	11.6	17.0	13.2	19.2	7.7	7.8	6.0	12.7	11.9	10.4	13.4	9.2	8.7	13.4	11.2	12.2	9.7	8.6	12.7
xMUDA-pl [18]	MM-TTA	28.8	26.9	35.9	28.2	21.3	26.5	32.2	26.0	27.0	21.6	24.9	17.9	26.4	39.3	42.1	37.4	37.0	25.9	38.7	**43.9**	37.9	37.7	25.0	36.5
MMTTA [35]	MM-TTA	31.1	24.4	30.7	25.8	28.3	24.2	26.8	23.2	29.6	20.7	22.1	20.6	25.6	39.9	40.0	30.9	31.5	29.6	30.6	32.2	23.9	26.4	23.8	30.9
CoTTA [42]	CTTA	29.7	27.2	34.7	27.0	26.4	25.6	33.0	27.3	28.1	18.0	22.7	18.4	26.5	32.9	28.0	22.9	22.2	20.3	26.1	24.9	23.6	24.7	19.7	24.5
EATA [30]	CTTA	34.0	30.0	38.6	30.2	30.2	28.4	36.5	30.1	32.2	21.3	25.3	20.1	29.7	40.1	40.8	36.3	34.3	28.9	39.3	41.7	36.5	37.9	28.8	36.5
MMTTA-rs [35]	MM-CTTA	31.9	28.1	37.7	29.3	28.7	27.6	35.2	29.7	30.2	20.4	25.6	20.3	28.7	40.4	41.5	36.3	34.7	30.2	39.9	41.5	36.4	38.3	29.9	36.9
CoMAC (Ours)	MM-CTTA	34.0	**34.7**	**44.9**	**38.1**	**35.0**	**37.8**	**45.1**	**39.0**	**39.1**	**29.4**	**35.1**	**27.1**	**36.6**	**41.4**	**43.7**	**38.8**	**37.8**	**32.2**	**42.2**	41.8	**40.7**	**38.9**	**30.4**	**38.8**
Time		t ←													t ←										
TENT [40]	TTA	10.2	8.8	8.6	9.0	10.2	9.5	9.8	11.4	18.3	13.2	16.9	18.9	12.1	12.0	11.5	9.2	9.8	9.7	12.6	13.0	12.5	19.8	21.2	13.1
MMTTA-rs [35]	MM-CTTA	31.4	27.4	37.2	28.7	26.2	27.2	33.1	29.0	28.5	19.9	22.3	19.7	27.6	40.4	41.4	36.4	31.6	29.1	37.6	40.2	35.4	35.2	27.8	35.5
CoMAC (Ours)	MM-CTTA	**34.6**	**32.8**	**40.9**	**33.2**	**31.8**	**32.4**	**39.2**	**33.2**	**33.8**	**25.1**	**30.3**	**23.3**	**32.4**	**40.8**	**41.9**	**37.7**	**33.2**	**30.3**	**39.6**	**43.6**	**38.8**	**38.6**	**30.0**	**37.5**

Table 1. Performance (mIoU) of SemanticKITTI-to-Synthia. Here we report the softmax-average mIoU of 2D and 3D prediction.

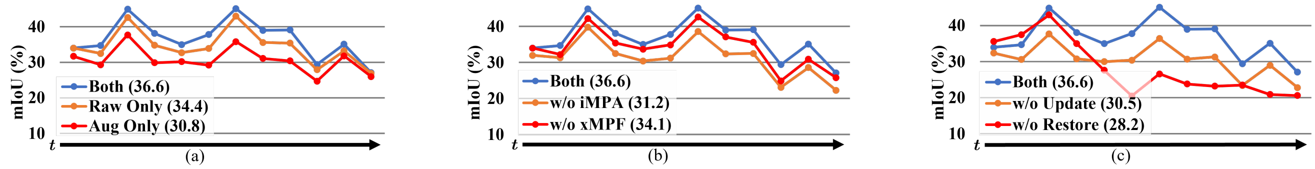

Figure 3. Ablation studies to justify our design of CoMAC. Specifically, (a) co mpares CoMAC with variants only using raw or augmented samples as input of iMPA and (b) justifies the necessity of weighting mechanisms in iMPA and xMPF. (c) compares CoMAC with variants without either target feature enqueuing or pseudo-source features restoring in CMQs. Scores within the parentheses are the average mIoU.

0.625, 0.75, 0.875] and z-rotation of [60°, 120°, 180°, 240°, 300°]. N_q and N_{enq} are set to 4096 and 200, respectively. We utilize a restore rate p_{rs} of 0.5 and a momentum factor λ_s of 0.999 with a coefficient λ_{cts} of 1. All methods strictly follow the one-pass protocol [36] (i.e., the training epoch is one where inference is conducted immediately for each sample). *More details are included in the appendix.*

4.2. Overall Results

As shown in Table 1, our CoMAC achieves SOTA performance on the average mIoU of all sequences for both benchmarks. Specifically, the proposed CoMAC significantly outperforms previous methods by more than 6.9% on S-to-S and 1.9% on S-to-W, which justifies the effectiveness of our method. It can also be observed that most TTA methods (TENT [40] and LAME [3]) are outperformed by CTTA methods, which reveals the importance of avoiding catastrophic forgetting. Additionally, previous MM-TTA (i.e., xMUDA and MMTTA) methods can not consistently outperform previous single-modal methods (e.g., EATA [30]), which suggests the sensitivity of multi-modal collaboration toward continual domain shifts. In fact, by intuitively utilizing model-based stochastic restoration, MMTTA-rs can achieve a noticeable improvement of 3.1% and 6.0% on S-to-S and S-to-W, respectively. Yet MMTTA-rs is still outperformed by our method. To justify the effect of the sequence arrangement, we conduct additional experiments by reversing the sequences of S-to-S and S-to-W for TENT, MMTTA-rs, and our CoMAC. While the performance of all methods decreases, the improvement of CoMAC compared to others can be observed across all sequences.

For S-to-S, we notice that our method achieves the second-best performance at the first sequence 01-Spring and then surpasses all previous methods in the following sequences since we utilize CMQs to gradually adapt to the changing target domain without forgetting. Compared with TTA methods, our method consistently prevails with an increasing gap as they struggle with error accumulation. A similar observation can be found at D-O-2 of S-to-W. Interestingly, xMUDA-pl as well as Pslabel perform competitively on S-to-W and outperform CoMAC in D-P-2 mainly because is less challenging and simply filtering pseudo-labels can effectively mitigate the prediction noise. Nevertheless, this performance gap is trivial as our CoMAC surpasses both xMUDA-pl and Pslabel on all other sequences.

4.3. Ablation Studies

In this section, we provide our ablation studies on S-to-S to justify our design of CoMAC.

Utilization of raw and augmented input. In Sec. 3.2, a key hypothesis is that while the augmentation-average prediction can be a potential reliable alternative, it may introduce inductive bias to the result. To justify this assumption, we conduct experiments by simply using the pure raw

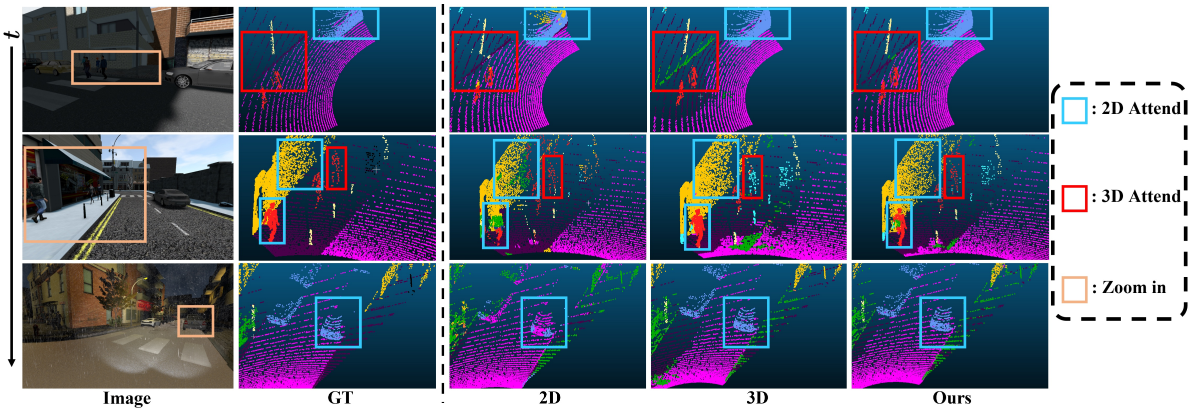

Figure 4. Visualization of online segmentation results. Here we present the cross-modal pseudo-labels from xMPF (Ours) as well as the individual prediction from 2D and 3D teachers in comparison with the ground truth (GT).

or augmentation-average prediction as the intra-modal prediction as in Eq. 7. As shown in Fig. 3(a), both variants achieve inferior performance compared to CoMAC and these performance gaps are consistent across most sequences. Specifically, using pure augmentation-average predictions lead to serious degradation of 5.8%, which justifies the necessity of mitigating the inductive noise brought by test-time augmentations and our motivation for iMPA.

Weighting mechanisms of iMPA and xMPF. The proposed iMPA and xMPF utilize centroid-based weighting mechanisms to attend to the reliable modality. To justify their effectiveness, we compared CoMAC with variants by replacing either Eq. 7 or Eq. 8 with a simple average fusion (indicated as w/o iMPA and w/o xMPF). As presented in Fig. 3(b), CoMAC outperforms both variants by more than 2.5%. This performance gap justifies that both iMPA and xMPF play an important part in noise suppression.

Updating mechanisms of CMQs. The goal of CMQs is to achieve a good balance between adaptation and knowledge preservation. To justify the role of CMQs, we compared CoMAC with variants without enqueuing target features or restoring pseudo-source features, respectively. For the variant without enqueuing target features, the corresponding class-wise centroids are identical to the source centroids across the whole testing process. As in Fig. 3(c), disabling enqueuing target features (w/o Update) causes a significant performance decrease of 6.1%, which justifies the necessity of adapting class-wise centroids for the validity of iMPA and xMPF. On the other hand, while disabling restoring pseudo-source features (w/o Restore) leads to a quicker adaptation at the first two sequences, it eventually causes catastrophic forgetting after the third sequence, resulting in a noticeable gap of 8.4%. Overall, the performance gain brought by CMQs justifies their effectiveness.

Number of Augmentations. To investigate the effect of augmentations, we conduct a grid search by utilizing different numbers of 2D and 3D augmentations. Specifically, we adopt resize and rotation as 2D and 3D augmentation, respectively. The factor range for 2D scaling is $[0.5, 1)$ while 3D one is set to $(0^\circ, 360^\circ)$, where the list of augmentations is computed as evenly spaced factors given the number of augmentations. As shown in Table 2, increasing the number of either 2D or 3D augmentations leads to performance improvement, while the 2D improvement is less noticeable compared to 3D probably due to the inductive bias brought by 2D augmentations. While all settings perform competitively, we adopt the combination of three 2D augmentations and five 3D augmentations to achieve the best result.

CoMAC		No. of 2D Aug					
		1	2	3	4	5	Avg
No. of 3D Aug	1	33.0	33.1	33.2	33.2	33.3	33.2
	2	34.1	34.2	34.3	34.4	34.4	34.3
	3	35.2	35.2	35.4	35.3	35.5	35.3
	4	36.4	36.4	36.5	36.5	36.5	36.5
	5	36.4	36.5	**36.7**	36.6	**36.7**	36.6
	Avg	35.0	35.1	35.2	35.2	35.3	35.2

Table 2. Augmentation analysis of CoMAC. CoMAC is evaluated with different numbers of augmentations.

	p_{rs}	$(\tau_{cf}, \lambda_{cts})$	Avg	τ_{cf}	(p_{rs}, λ_{cts})	Avg	λ_{cts}	(p_{rs}, τ_{cf})	Avg
CoMAC	0.5	(0.8, 1.0)	36.6	0.8	(0.5, 1.0)	36.6	1.0	(0.5, 0.8)	36.6
	0.3	-	34.5	0.0	-	34.8	0.0	-	35.8
	0.7	-	36.4	0.5	-	36.5	0.01	-	36.4
	1.0	-	36.3	0.9	-	36.4	0.1	-	36.4

Table 3. Sensitivity analysis of CoMAC. Here “xM” is the softmax-average of 2D and 3D predictions. The parameter value with “-” indicates the same value as default settings.

Sensitivity analysis. We perform sensitivity analysis over the hyper-parameters of CoMAC as shown in Table 3. There are some extreme cases that can lead to the performance degeneration of CoMAC. For restoring rate p_{rs}, the performance relatively drops about 5.7% when $p_{rs} = 0.3$, indicating the restoring rate is too low to prevent forgetting. Nevertheless, when $p_{rs} \geq 0.5$, CoMAC is robust to the arbitrary choice of p_{rs}. Similar observations can be found when

Method	Aug Samples	Inf. Time (s)/sample		mIoU
		mean	std	
TENT [40]	✗	0.379	0.010	17.8
CoTTA [42]	✓	7.220	0.399	26.5
Ours	✓	1.888	0.082	36.6
Ours (less augs)	✓	1.004	0.038	33.0

Table 4. Comparison of processing time per sample.

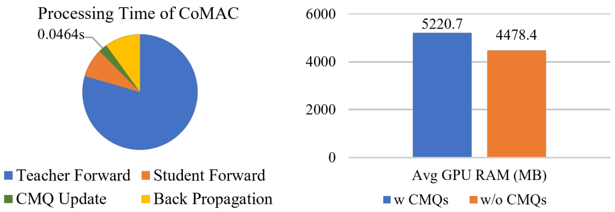

Figure 5. Processing time of CoMAC components per sample (left) and GPU RAM with or without CMQs (right). Experiments are conducted on S-to-S with the default backbones and settings.

$\tau_{cf} = 0.0$ or $\lambda_{cts} = 0.0$, where the performance decreases relatively by 4.9% and 2.2%. This performance degeneration indicates the importance of confident feature filtering and class-wise contrastive loss. Overall, when $\tau_{cf} \geq 0.0$ and $\lambda_{cts} \geq 0.01$, the performance of CoMAC is also robust to the hyper-parameter settings, falling into a relative margin of 0.5%. Note that all settings of CoMAC surpass the previous SOTA method EATA [30] by more than 4.8%.

Visualization of segmentation. To further justify the effectiveness of our proposed CoMAC, we provide some visualization results as shown in Fig. 4. Specifically, we visualize the 2D and 3D predictions from teacher models and our cross-modal pseudo-labels from xMPF. Compared to single-modal predictions, our cross-modal pseudo-labels can achieve better segmentation results by attending to the reliable modality, such as the first row in Fig. 4, where the sidewalk prediction attends more to the 2D image while the car prediction attends to the reliable 3D point cloud.

Computational and storage overhead. To investigate the computational and storage overhead of CoMAC, we analyze the processing time and GPU occupancy when online testing with our CoMAC as illustrated in Table 4 and Fig. 5. Similar to previous augmentation-based methods, our proposed CoMAC is limited by the computational overhead brought by multiple augmentations. However, compared to the augmentation-based SOTA method CoTTA [42], our method outperforms it in terms of performance and efficiency (a relative performance improvement of 38.1% with 73.9% less processing time). Compared to the single-sample-based method TENT, despite a computational overhead of 1.51s, our method significantly improves relatively by more than 100%. The computational overhead can be further lessened by using fewer augmentations, reducing the processing time relatively by more than 46.8% with a minor performance drop of 3.6%. Besides augmentations, another potential source of overhead lies in the utilization of CMQs as they capture class-wise features for each iteration. As shown in Fig .5, however, the computational overhead brought by CMQs is rather trivial, which takes less than 0.05s to update and process each frame. Moreover, since CMQs only preserve the high-level features instead of raw samples, the storage cost is also acceptable with a minor usage increase (less than 800 MB) of GPU RAM compared to the CoMAC invariant without CMQs.

5. Conclusion

In this paper, we present a new task, named multi-modal continual test-time adaptation (MM-CTTA) for 3D semantic segmentation. We further propose a novel method called CoMAC that tackles MM-CTTA from two perspectives. On one hand, reliable cross-modal pseudo-labels are generated by iMPA and xMPF by adaptively attending to the more reliable modality in a dual-stage manner. On the other hand, CMQs are proposed to leverage pseudo-source features and reliable target features to perform adaptation without catastrophic forgetting. We introduce two new benchmarks for MM-CTTA and our methods outperform previous works by a noticeable margin in both benchmarks. In the future, we hope both our method and benchmarks can facilitate the exploration of MM-CTTA and promote the development of reliable multi-modal systems in altering environments.

Acknowledgement

This research is supported by the National Research Foundation, Singapore under its Medium Sized Center for Advanced Robotics Technology Innovation. This research is jointly supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-027).

References

[1] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, pages 1–32, 2021. 1

[2] 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. 3, 6

[3] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8344–8353, 2022. 2, 6, 7

[4] Simon Bultmann, Jan Quenzel, and Sven Behnke. Real-time multi-modal semantic fusion on unmanned aerial ve-

hicles with label propagation for cross-domain adaptation. *Robotics and Autonomous Systems*, 159:104286, 2023. 1

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 3

[6] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 627–636, 2019. 1

[7] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022. 2

[8] Xieyuanli Chen, Andres Milioto, Emanuele Palazzolo, Philippe Giguere, Jens Behley, and Cyrill Stachniss. Suma++: Efficient lidar-based semantic slam. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4530–4537. IEEE, 2019. 1

[9] C. Eastwood, I. Mason, C. Williams, and B. Schölkopf. Source-free adaptation to measurement shift via bottom-up feature restoration. In *10th International Conference on Learning Representations (ICLR)*, Apr. 2022. 3

[10] 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. 2

[11] Yulu Gan, Xianzheng Ma, Yihang Lou, Yan Bai, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: visual domain prompt for continual test time adaptation. *arXiv preprint arXiv:2212.04145*, 2022. 3

[12] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. In *Advances in Neural Information Processing Systems*, 2022. 3

[13] Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and J Zico Kolter. Test time adaptation via conjugate pseudo-labels. In *Advances in Neural Information Processing Systems*, 2022. 2

[14] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9224–9232, 2018. 6

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[17] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1989–1998. Pmlr, 2018. 1

[18] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Perez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020. 2, 3, 6, 7

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 5

[20] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7046–7056, 2021. 2

[21] Jogendra Nath Kundu, Akshay R Kulkarni, Suvaansh Bhambri, Deepesh Mehta, Shreyas Anand Kulkarni, Varun Jampani, and Venkatesh Babu Radhakrishnan. Balancing discriminability and transferability for source-free domain adaptation. In *International Conference on Machine Learning*, pages 11710–11728. PMLR, 2022. 2

[22] 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. 2

[23] Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. *arXiv preprint arXiv:2302.05155*, 2023. 2

[24] Wei Liu, Zhiming Luo, Yuanzheng Cai, Ying Yu, Yang Ke, José Marcato Junior, Wesley Nunes Gonçalves, and Jonathan Li. Adversarial unsupervised domain adaptation for 3d semantic segmentation with multi-modal learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176:211–221, 2021. 3

[25] 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. 2

[26] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021. 2

[27] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Elsevier, 1989. 1

[28] John Mccormac, Ronald Clark, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Fusion++: volumetric object-level slam. In *2018 International Conference on 3D Vision (3DV)*, pages 32–41, 2018. 1

[29] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic map-

ping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4628–4635. IEEE, 2017. 1

[30] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, pages 16888–16905. PMLR, 2022. 2, 3, 6, 7, 9

[31] Duo Peng, Yinjie Lei, Wen Li, Pingping Zhang, and Yulan Guo. 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*, pages 7108–7117, 2021. 3

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 6

[33] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3234–3243, 2016. 3, 6

[34] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1214–1223, 2021. 4

[35] 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. 1, 2, 3, 6, 7

[36] Yongyi Su, Xun Xu, and Kui Jia. Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. 1, 2, 3, 5, 7

[37] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 3, 6

[38] 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. 2, 4

[39] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[40] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 1, 2, 6, 7, 9

[41] Jun-Kun Wang and Andre Wibisono. Towards understanding gd with hard and conjugate pseudo-labels for test-time adaptation. *arXiv preprint arXiv:2210.10019*, 2022. 2

[42] 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. 1, 2, 3, 4, 6, 7, 9

[43] Zejie Wang, Zhen Zhao, Zhao Jin, Zhengping Che, Jian Tang, Chaomin Shen, and Yaxin Peng. Multi-stage fusion for multi-class 3d lidar detection. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3113–3121, 2021. 1, 6

[44] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1887–1893. IEEE, 2018. 1

[45] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4376–4382. IEEE, 2019. 1

[46] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16024–16033, 2021. 1

[47] Yuecong Xu, Haozhi Cao, Kezhi Mao, Zhenghua Chen, Lihua Xie, and Jianfei Yang. Aligning correlation information for domain adaptation in action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 1

[48] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Min Wu, and Zhenghua Chen. Source-free video domain adaptation by learning temporal consistency for action recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 147–164. Springer, 2022. 2

[49] Jianfei Yang, Yuecong Xu, Haozhi Cao, Han Zou, and Lihua Xie. Deep learning and transfer learning for device-free human activity recognition: A survey. *Journal of Automation and Intelligence*, 1(1):100007, 2022. 3

[50] Marvin Mengxin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. 2

[51] Bowen Zhao, Chen Chen, and Shu-Tao Xia. Delta: degradation-free fully test-time adaptation. *arXiv preprint arXiv:2301.13018*, 2023. 2

[52] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019. 1