RobIA: Robust Instance-aware Continual Test-time Adaptation for Deep Stereo

Jueun Ko[1*] **Hyewon Park**[1*] **Hyesong Choi**[2] **Dongbo Min**[1†]
[1]Ewha Womans University [2]Soongsil University
{jueun.ko, hwpark}@ewha.ac.kr hyesong@ssu.ac.kr dbmin@ewha.ac.kr
https://github.com/0ju-un/RobIA

Abstract

Stereo Depth Estimation in real-world environments poses significant challenges due to dynamic domain shifts, sparse or unreliable supervision, and the high cost of acquiring dense ground-truth labels. While recent Test-Time Adaptation (TTA) methods offer promising solutions, most rely on static target domain assumptions and input-invariant adaptation strategies, limiting their effectiveness under continual shifts. In this paper, we propose **RobIA**, a novel *Robust, Instance-Aware* framework for Continual Test-Time Adaptation (CTTA) in stereo depth estimation. RobIA integrates two key components: (1) Attend-and-Excite Mixture-of-Experts (AttEx-MoE), a parameter-efficient module that dynamically routes input to frozen experts via lightweight self-attention mechanism tailored to epipolar geometry, and (2) Robust AdaptBN Teacher, a PEFT-based teacher model that provides dense pseudo-supervision by complementing sparse handcrafted labels. This strategy enables input-specific flexibility, broad supervision coverage, improving generalization under domain shift. Extensive experiments demonstrate that RobIA achieves superior adaptation performance across dynamic target domains while maintaining computational efficiency.

1 Introduction

Stereo Depth Estimation (SDE) is a fundamental task for 3D scene understanding, with applications in autonomous driving and robotics. While deep learning-based stereo approaches have achieved notable accuracy improvements [1, 2], their success relies largely on supervised training with dense ground-truth disparity maps, which are costly and labor-intensive to obtain. As a result, they are typically pre-trained on large-scale synthetic datasets [2] and later adapted to real world training datasets. However, they suffer from domain shifts by challenging conditions unseen in the training datasets, leading to performance degradation during inference. To deal with these challenges, Test-Time Adaptation (TTA) has recently emerged [3–6], aiming to adapt a model on-the-fly to unseen target domains during inference, generally conducted in an unsupervised manner.

Current TTA approaches for SDE [7, 8] operate under the assumption of a single, stationary target domain, overlooking more realistic scenarios where the domain distribution evolves over time, including changing weather, lighting, or scene structure. Continual Test-time Adaptation (CTTA) [9–12] has recently emerged as a framework for continuously adapting models to consistently evolving target domains, and has been applied to other vision tasks such as classification and semantic segmentation. In this regard, we explore a more practical and challenging CTTA setting for SDE. In general, CTTA introduces two major challenges: *catastrophic forgetting*, where the model gradually loses the knowledge acquired from source domains while adapting to new target domains, and *error*

[*]Equal contribution.
[†]Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

accumulation, where noisy pseudo-labels progressively degrade model performance. Addressing these issues requires a careful balancing between stability, to preserve prior knowledge, and plasticity, to accommodate new domain-specific variations.

To address the stability-plasticity trade-off, recent CTTA methods have explored Parameter-Efficient Fine-Tuning (PEFT) strategies [10, 12], which preserve the representation capacity of pre-trained backbones by freezing them, while enabling adaptation via a small set of trainable components such as prompt and meta-networks. These approaches have shown promising results in classification tasks, but their effectiveness diminishes in more complex settings such as semantic segmentation [13], which requires dense, spatially structured predictions. A key limitation lies in the input-invariant nature of standard PEFT modules where adapters or prompts are fixed for all inputs, making it difficult to capture instance-specific variations [14, 15]. This underscores the need for PEFT methods that dynamically adapt to each instance, particularly under continual domain shifts.

In addition to architectural adaptability, the quality of supervision plays a crucial role in effective adaptation. Existing methods for SDE commonly rely on photometric consistency loss [7, 8] or pseudo-labels generated by handcrafted stereo matching algorithms [16], which are considered relatively robust to domain shifts [17]. Prior works typically filter these pseudo-labels with confidence-based thresholding, since they are often unreliable in challenging regions such as occlusions, reflective surfaces, or low-texture areas. While this selective supervision improves pseudo-label quality, it introduces a critical drawback: the model is trained only on a subset of the input target domains, leading to over-reliance to confident pseudo labels and weak generalization to uncertain or structurally complex regions. These observations highlight the need for more comprehensive supervision signals that can cover the full input distribution and mitigate the risk of pseudo-label over-reliance during continual adaptation.

To enable instance-aware adaptation and improve pseudo-supervision under dynamic conditions, we propose Robust, Instance-Aware CTTA approach, termed **RobIA**, which is tailored for stereo depth estimation in continually shifting domains. RobIA addresses the limitations of conventional PEFT and proxy-labeling approaches through two key components. First, we introduce the Attend-and-Excite Mixture-of-Experts (AttEx-MoE), a compact yet effective MoE architecture that enables input-specific adaptation without updating the backbone. Inspired by selective channel excitation [18], AttEx-MoE dynamically activates the convolutional channel experts via a self-attention, conditioned on instance-aware global features. To reduce computational overhead, we constrain the attention operation to be row-wise along epipolar lines, leveraging stereo geometry while preserving long-range contextual reasoning. This design allows AttEx-MoE to maintain the efficiency of PEFT while introducing fine-grained, content-aware adaptability crucial for dense stereo predictions.

Second, we propose the Robust AdaptBN Teacher, a complementary PEFT-based model that enhances the coverage of pseudo-labels during adaptation. Prior work [3, 19] has presented an effective test-time adaptation mechanism by updating only the affine parameters of batch normalization layers, known as AdaptBN [20, 21]. Building on this, we leverage an AdaptBN-trained teacher model to complement the sparsity of handcrafted stereo pseudo-labels in low-confidence regions. Specifically, we adopt a hybrid supervision scheme: reliable pseudo-labels from handcrafted stereo matching algorithms are used in high-confidence areas, while predictions from the Robust AdaptBN Teacher supervise low-confidence regions. This dual-source guidance allows the model to retain the precision of proxy labels where they are reliable, while extending supervision coverage to previously ignored areas, thus promoting better generalization across the entire input space.

Our key contributions are summarized as follows: **(1)** We propose RobIA, a novel CTTA framework specifically designed for stereo depth estimation under dynamic domain shifts. **(2)** We introduce a parameter-efficient, instance-aware adaptation module (AttEx-MoE) that dynamically routes input through frozen convolutional experts using a lightweight row-wise self-attention mechanism. **(3)** We design a complementary PEFT-based (AdaptBN) teacher that provides pseudo-supervision in low-confidence regions, enhancing coverage and robustness of pseudo labels. **(4)** We propose a dual-source supervision scheme that combines reliable handcrafted stereo pseudo-labels with predictions from the AdaptBN Teacher, enhancing coverage and robustness of pseudo labels.

2 Related Works

Test-Time Adaptation for Stereo Depth Estimation Test-Time Adaptation (TTA) in stereo depth estimation (SDE) aims to adapt a model to new domains in an online or real-time manner without

access to source data or ground-truth labels. Early approaches include modularized model update [8], meta-learning-based adaptation [7], and pixel-wise focused adaptation [22]. Despite these advancements, prior stereo TTA approaches have mostly focused on *single-domain adaptation* or *long-term adaptation* using large sequences (typically each domain contains more than 2K frames) within a static domain. These approaches overlook practical scenarios in which domains evolve continuously over time. To address this gap, we introduce a continual test-time adaptation scenario for SDE that reflects temporally evolving domain distributions.

Self-supervised Learning for Stereo Depth Estimation A long-standing challenge in SDE is the low density and high acquisition cost of ground-truth labels. To overcome this, self-supervised learning has been widely adopted [23, 24] , commonly using the photometric loss between stereo pairs. However, this signal often suffers from matching ambiguities, such as occlusions and specular surfaces, leading to unreliable supervision. [24, 17] exploited traditional stereo algorithms to generate pseudo-labels, filtering the outliers with confidence measures. [25] further introduced a monocular completion network to distill hard-to-match regions from stereo matching. While effective, it requires a separate network and multiple inference steps, leading to significant computational overhead, making it impractical for online adaptation. In contrast, our approach leverages a robust teacher model with lower computational cost, offering reliable guidance even in challenging regions where handcrafted pseudo-labels are absent.

Mixture-of-Experts Mixture-of-Experts (MoE)[26, 27] has been widely used in various domains, including CTTA MoE enables dynamic selection of expert subnetworks via routing mechanism, making it effective for multi-task learning [28, 29] and continual learning [30, 31]. In CTTA, several studies have explored parameter-efficient fine-tuning (PEFT) approaches that integrate MoE modules into pre-trained backbones [32, 33], enabling domain adaptation with minimal trainable parameters. However, most of this research has been conducted on Transformer-based architectures, and the application of MoE within CNNs remains relatively unexplored. DeepMoE [34] is a common approach in CNN-based architecture to treat individual channels as experts, allowing fine-grained modulation of feature representations and improving model sparsity. Our work is motivated by this underexplored direction, proposing a CNN-compatible MoE design that supports instance-aware adaptation under continual domain shifts.

3 Preliminaries

Continual Test-time Adaptation In CTTA paradigm, we are given a model pre-trained on a source dataset $(\mathcal{X}_S, \mathcal{Y}_S)$. The goal is to adapt the model to multiple unlabeled target data distribution $\mathcal{X}_T = \{\mathcal{X}_{T_1}, \mathcal{X}_{T_2}, \ldots, \mathcal{X}_{T_n}\}$ during deployment, where n represents the number of unseen domains. When the target domain data $\mathcal{X}_{T_i}$ consists of N_i target samples for $i = 1, \ldots, n$, for simplicity, we denote (I_t^L, I_t^R) as the t^{th} target stereo image pair for $t = 0, \ldots, |\mathcal{X}_T| - 1$, where $|\mathcal{X}_T| = \Sigma_{i=1}^{n} N_i$. Similarly, in the following sections of this paper, we omit the domain notation $\mathcal{X}_{T_i}$ as all target domains are considered to be integrated into a single sequence. Therefore, at each time step t, the stereo model predicts disparity map from a single stereo image pair (I_t^L, I_t^R), and updates its parameters before proceeding to the next input.

Pseudo-supervision for SDE Since ground-truth labels are unavailable at test time, the model must be trained in a *self-supervised manner*. Following prior work [17], we obtain the handcrafted disparity map D_{proxy} as a pseudo-label using the traditional stereo matching algorithm, Semi-Global Matching (SGM) [16]. Then, D_{proxy} is filtered by the confidence threshold ε, where confidence $c_t(p)$ is calculated via left-right consistency at the corresponding pixel p. Consequently, the mask $\mathcal{M}_{\text{valid}}$, which indicates the reliable region of D_{proxy} (*i.e.* valid region), is denoted as follows:

$$\mathcal{M}_{\text{valid}}(p) = \begin{cases} 1 & \text{if } c_t(p) \geq \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$\mathcal{M}_{\text{invalid}} = 1 - \mathcal{M}_{\text{valid}}$, indicating the unreliable region of D_{proxy} (*i.e.* invalid region).

Mixture-of-Experts in CNN Prior work [34] applying Mixture-of-Experts (MoE) to convolutional neural networks (CNNs) formulates C convolutional kernels (*i.e.* the output channels of the previous layer) as an individual expert E_i for $i = 1, .., C$, and introduces a gating network G to compute a

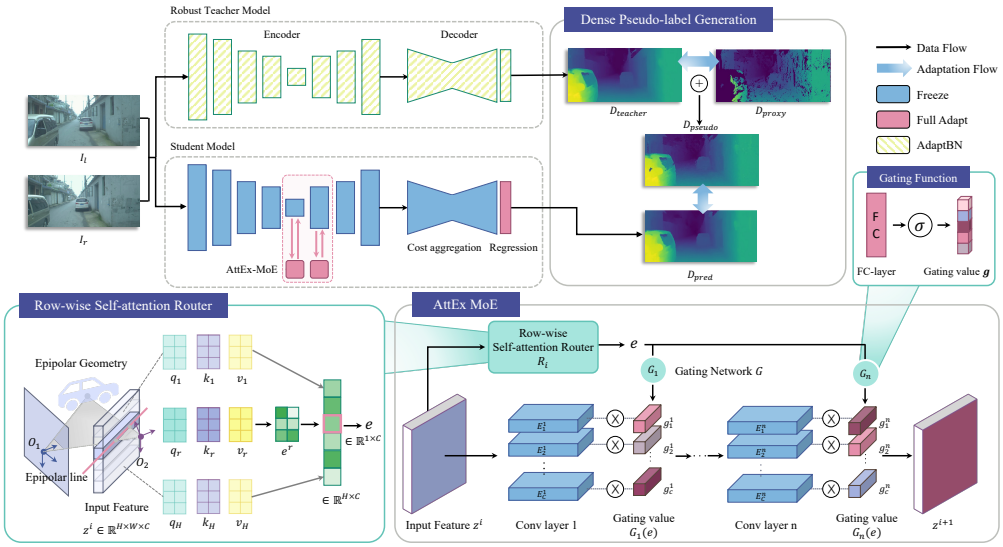


Figure 1: **The Overview of RobIA.** During test time, the student model is trained using dense pseudo-labels generated by combining sparse handcrafted proxy D_{proxy} with Robust Teacher prediction D_{teacher}, ensuring stable adaptation under dynamic conditions. AttEx-MoE integrates a row-wise self-attention router and gating network G into deep encoder blocks. The row-wise self-attention router extracts global context from an input feature map z, which is subsequently processed by a gating network. The student backbone is kept frozen, and only the router, gating network, and the regression parameters of the decoder are updated.

weighted combination of expert outputs. The output feature y is defined as:

$$y = \sum_{i=1}^{C} g_i \cdot E_i(x), \tag{2}$$

where $\boldsymbol{g} \in \mathbb{R}^C$ is the gating values obtained from the gating network G, with g_i being the weight assigned to the i^{th} expert $E_i(x)$. In this case, the gating network G is implemented as a multi-headed sparse gating network that takes a shallow embedding e of the *raw input image* and produces a channel-wise activation score. $G(e)$ is obtained by projecting the embedding e into a score vector via a learnable weight matrix W_g, followed by ReLU activation:

$$G(e) = \text{ReLU}(W_g \cdot e). \tag{3}$$

This formulation enables input-dependent dynamic selection of feature channels, improving both model sparsity and representational flexibility.

4 Proposed Method

4.1 Motivation and Overview

CTTA presents several key challenges, notably catastrophic forgetting and error accumulation. Parameter-efficient fine-tuning (PEFT) methods address these issues by freezing pretrained weights, preserving source knowledge. However, in conventional PEFT-based CTTA approaches, an efficient module that adapts to the target domain is added to the original feature, performing the uniform transformation on all inputs. This rigid transformation is insufficient for handling various input variations, particularly in dense predictions under domain shifts. Moreover, pseudo-labels from handcrafted stereo algorithms are often consistently filtered in the unreliable supervision regions due to inherent matching difficulties. Despite their domain-agnostic properties, they lead to over-reliance on reliable regions, misdirected adaptation elsewhere, and ultimately degrading overall performance.

To overcome these limitations, we propose a novel CTTA framework, **RobIA** (see Fig. 1). RobIA is designed to preserve the rich representational capacity of source knowledge while enabling stable and effective adaptation through the **Attend-and-Excite Mixture-of-Experts (AttEx-MoE)** architecture. This module extracts instance-specific features via a row-wise self-attention router and dynamically

excites frozen channel experts, allowing the model to adapt flexibly without modifying the pretrained backbone. To further address the challenge of biased and sparse pseudo-labels, we additionally introduce the **Robust AdaptBN Teacher**. It provides a hybrid supervision strategy that combines reliable handcrafted pseudo-labels with predictions from the robust teacher model.

4.2 Input-Aware Mixture-of-Experts via Attend-and-Excite

Mixture-of-Experts (MoE) in CNNs consists of a shallow embedding network and gating network with ReLU activation to treat each convolutional kernel as an individual expert and introduce sparsity into the model. However, this approach presents two key limitations: (1) shallow convolutional layers used in the embedding network lack sufficient spatial context for optimal expert selection, and (2) ReLU-induced sparsity limits representational capacity, especially when the backbone is frozen. To address these challenges, we propose **Attend-and-Excite Mixture-of-Experts (AttEx-MoE)**, a novel MoE design tailored for PEFT-based adaptation in stereo depth estimation. AttEx-MoE enables instance-aware, channel-wise modulation under continual domain shifts.

Row-wise Self-Attention Router A key component of AttEx-MoE is the row-wise self-attention router, which extracts global context to guide instance-aware expert excitation. Since convolutions are effective at capturing local patterns but struggle with modeling long-range dependencies, prior work [18] identified this limitation and proposed global average pooling (GAP) to create channel-wise features. However, GAP has the drawbacks of treating all spatial positions uniformly and ignoring spatial importance. Therefore, we employ a self-attention mechanism in the gating network to explicitly capture inter-channel dependencies and their spatial relationships. This allows Attend-and-Excite operations to adaptively excite the most relevant features for each target instance. Moreover, we apply 1D self-attention along each epipolar line (*i.e.*, row) of the feature map by leveraging stereo geometry [35], significantly lowering computational overhead while maintaining global context.

As shown in Fig. 1, the encoder of our model consists of N blocks, with the AttEx-MoE module applied to the final, deepest encoder block (*i.e.*, at $1/32$ resolutions) and the corresponding upsampling module connected via skip connections. In each block, the row-wise self-attention router R_i for $i = 1, ..., N$ computes the gating input e from the input feature map z^{i-1}, which is then passed to the gating network G in each convolutional layer.

$$e = \frac{1}{H} \sum_{r=1}^{H} e_r, \tag{4}$$

where $e_r = \text{softmax}\left(\frac{q_r k_r^\top}{\sqrt{d}}\right) v_r$ is computed for $r \in [1, H]$ using $q_r = z_r W_q$, $k_r = z_r W_k$, and $v_r = z_r W_v$. Here, $z_r \in \mathbb{R}^{W \times C}$ denotes the r-th row of the feature map $z \in \mathbb{R}^{H \times W \times C}$, and the attention is computed independently for each row. The attention score e_r obtained along the epipolar line is averaged over the height dimension H to compute the final gating input e.

Expert Excitation via Sigmoid In [34], each expert output $E_i(x)$ is scaled by a gate value $G(x)$ produced via ReLU activations, which suppresses certain experts by zeroing their outputs, as explained in (2) and (3). While these sparse expert outputs can enhance generalization [36], this sparsity may reduce expressivity when the backbone is frozen in the PEFT setting.

Therefore, we adopt a Sigmoid-based soft gating mechanism, in contrast to the commonly used ReLU. This design allows all experts to be activated to varying degrees rather than enforcing hard selection of a subset of experts, promoting richer expert combinations and diverse representations conditioned on each input instance. We find that soft gating is particularly effective in the PEFT setting (See Tab. 5), as it maximizes the utilization of all available experts with limited learning capacity. The gating mechanism is defined with the Sigmoid activation σ as:

$$G(e) = \sigma(W_g \cdot e). \tag{5}$$

4.3 Dense Pseudo-label Generation via Dual Supervision

For effective adaptation to the target domain, the design of the supervision plays a critical role. Pseudo-labels generated by traditional stereo algorithms are typically sparsified through reliability-based thresholding [17], which leads to over-reliance on sparse supervision and consequently limits performance gains in CTTA. To address this, we propose a dense pseudo-label generation strategy that

combines the domain-agnostic reliability of handcrafted pseudo-labels with the learning capability of a learnable teacher model.

AdaptBN [3, 20, 21] adapts only the affine parameters (scale γ, shift β) of batch normalization layers, enabling low-dimensional, channel-wise feature modulation and stable adaptation. We leverage the AdaptBN-based teacher model to complement unreliable regions in sparse, handcrafted pseudo-labels, producing dense pseudo-labels as supervision for student model. This mitigates the student's over-reliance on sparse pseudo labels and reduces performance degradation in unreliable regions.

Why AdaptBN? Most CTTA frameworks [9, 11] avoid stochastically-updated teacher model, as it is expected to provide stable supervision with minimal computational and memory overhead. Accordingly, common CTTA designs adopt Mean Teacher models updated via exponential moving average (EMA) [37] or fixed, source-trained models [38]. However, stereo depth estimation task presents a unique setting where handcrafted stereo algorithms can yield reliable pseudo-labels (*i.e.* proxy), reducing the need for stability-focused teachers. In this context, the teacher must not only stabilize but also generalize beyond the limitations of the sparse proxy, particularly in regions where handcrafted labels are unreliable.

Since AdaptBN performs adaptation via low-dimensional affine transformations, it offers a controlled yet expressive adaptation mechanism for test-time adaptation. This makes AdaptBN spatially robust, allowing the teacher to effectively adapt to the target domain while reducing the model's over-reliance to the sparsely provided proxy labels. In contrast, EMA-based teachers are less suitable in this context. Although they maintain stability by updating weights gradually, they are prone to error accumulation when the student produces biased predictions. In such cases, EMA teachers tend to reinforce incorrect predictions, failing to provide correct guidance to the student [39]. We further validate these claims through analysis and ablation studies, which demonstrate AdaptBN teacher's contributions to both stability and adaptive refinement, especially in regions where handcrafted labels are unreliable.

4.4 Continual Test-time Adaptation Process

Model Initialization We insert the lightweight AttEx-MoE module into a source-trained base model. Following recent CTTA studies [32], we train the AttEx-MoE module on the labeled source dataset while keeping the backbone frozen. This short *warm-up* phase allows the MoE module to learn instance-aware routing behavior without altering the core representations, providing a stable initialization for CTTA. During the supervised warm-up phase, we trained the model in the same way as DeepMoE [34]. For test-time adaptation, we freeze the base network and only update the AttEx-MoE module and the regression parameters of the decoder.

Total Loss The overall loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{proxy}} + \lambda \mathcal{L}_{\text{teacher}} \tag{6}$$

$$\mathcal{L}_{\text{proxy}} = \mathcal{M}_{\text{valid}} \cdot smooth_{L1}(D_{\text{proxy}} - D_{\text{pred}}), \tag{7}$$

$$\mathcal{L}_{\text{teacher}} = \mathcal{M}_{\text{invalid}} \cdot smooth_{L1}(D_{\text{teacher}} - D_{\text{pred}}), \tag{8}$$

where D_{proxy} and D_{teacher} denote the pseudo-labels generated by the handcrafted stereo algorithm [16] and the AdaptBN teacher model, respectively. The predicted disparity map D_{pred} is supervised by two loss terms, $\mathcal{L}_{\text{proxy}}$ and $\mathcal{L}_{\text{teacher}}$, corresponding to each pseudo-label source. Incorporating supervision from D_{teacher} helps regularize the student model, mitigating over-reliance on sparse handcrafted labels and preventing performance drops in regions lacking reliable pseudo-labels. λ is a loss weight, which controls the influence of teacher predictions in our dense pseudo-label formulation.

5 Experiments

Datasets. To simulate TTA and CTTA scenarios, following prior work [8], all experiments were conducted on well-renowned stereo benchmarks, including KITTI RAW [40], DrivingStereo [41], and DSEC [42]. These datasets cover various conditions, such as different weather scenarios and urban cityscapes in both daylight and nighttime. The synthetic Flyingthings3D, part of the synthetic SceneFlow dataset [2] was used to pretrain the stereo model before test time. Sparse pseudo-labels D_{proxy} were obtained via Semi-Global Matching (SGM) [16], followed by a left-right consistency check. The effective label density varies across datasets—roughly 92% for KITTI RAW, 72% for DrivingStereo, and 45% for DSEC, highlighting the need for robust dense supervision, especially in relatively sparser datasets such as DrivingStereo and DSEC.

Table 1: Performance comparison of Continual Test-time Adaptation on **DrivingStereo** benchmark over 10 rounds. To save space, only 1st and 10th round scores are written. **Bold** denotes best and *AT* denotes our method with dense pseudo-label D_{teacher}.

Method	Round / Condition / Adapt.	1 dusky D1-all	1 dusky EPE	1 cloudy D1-all	1 cloudy EPE	1 rainy D1-all	1 rainy EPE	10 dusky D1-all	10 dusky EPE	10 cloudy D1-all	10 cloudy EPE	10 rainy D1-all	10 rainy EPE	All ↓ Mean D1-all	All ↓ Mean EPE
MADNet 2 [45]	(a) no adapt.	13.24	1.69	6.56	1.22	11.51	2.18	13.24	1.69	6.56	1.22	11.51	2.18	10.44	1.70
	(b) FT	5.08	1.06	4.82	1.09	6.4	1.43	6.04	1.66	5.85	2.06	7.13	1.9	6.13	1.73
	(c) MAD++	6.46	1.15	4.36	1.04	6.01	1.29	5.79	1.60	6.04	2.01	7.28	1.89	5.86	1.41
CoEx [43]	(d) no adapt.	5.53	1.14	3.55	0.99	7.61	1.64	5.53	1.14	3.55	0.99	7.61	1.64	5.56	1.26
	(e) AdaptBN	5.16	1.11	3.13	0.93	6.14	1.41	2.71	0.85	2.36	0.8	3.32	1.03	3.08	0.94
	(f) FT	5.25	1.11	2.98	0.91	5.81	1.37	3.05	0.88	2.48	0.81	3.63	1.09	3.04	0.92
	(g) FT + AT	5.09	1.1	3.01	0.91	5.83	1.38	2.63	**0.84**	2.33	**0.79**	3.08	**0.99**	2.93	0.91
EcoTTA [12]	(h) MetaNet	4.61	1.05	2.89	0.89	**4.21**	1.17	3.42	0.93	2.69	0.87	4.46	1.26	3.07	0.95
RobIA (ours)	(i) AttEx-MoE	**4.01**	**1.01**	**2.4**	**0.84**	4.44	**1.13**	2.72	0.88	2.29	0.84	3.89	1.22	2.98	0.97
	(j) AttEx-MoE + AT	4.28	1.03	2.4	0.84	4.54	1.16	**2.4**	**0.84**	**2.24**	0.82	**3.02**	1.00	**2.77**	**0.91**

For CTTA, we constructed a new benchmark by sampling 500 frames per domain from existing TTA sequences, constructing a short sequence with frequent domain shifts. Each cycle consists of 3–4 domains and is repeated over 10 rounds to simulate long-term adaptation with recurring conditions. Specifically, we obtained a sequence of dusky→cloudy→rainy for DrivingStereo, night1 to night4 for DSEC, and city→residential→campus→road for KITTI RAW. Unlike prior works that simulate long-term shifts over 44K frames [17], our CTTA setting imposes more rapid domain adaptation within short sequences, reflecting real-world constraints where environmental changes occur frequently and previously seen conditions may reappear. We include TTA results for all datasets, additional results on CTTA, and ablation studies on pseudo-supervision in the supplementary material.

Implementation Details. We use CoEx [43], a compact and real-time stereo network using MobileNetV2 [44] backbone, as our base architecture. Following prior work [45], we retrained the model on the synthetic source datasets with strong data augmentations to improve generalization. All experiments were conducted on NVIDIA A6000 and RTX 3090 GPUs and further implementation details, including hyper parameters, are in the supplementary material.

no adapt. denotes the source-trained model without adaptation. We additionally evaluated two variants of MADNet2 [45]: *FT*, which updates all model parameters, and *MAD++*, which applies modular updates. All PEFT-based CTTA methods, including *AdaptBN*, *MetaNet*, and our proposed *AttEx-MoE*, were implemented on top of the CoEx for a fair comparison. AdaptBN tunes the affine parameters of batch normalization layers, MetaNet tunes the meta network from EcoTTA [12], and AttEx-MoE updates the lightweight gating module for expert selection. In all cases, the decoder's regression parameters were jointly updated. *AT* setting uses dense pseudo-labels generated from our AdaptBN teacher.

Evaluation Metrics. We reported End-Point Error (EPE) and D1-all that measures the percentage of pixels with absolute disparity error exceeding 3 pixels and 5% of the ground truth. We adopt the standard online adaptation protocol: the model predicts each frame before updating, then uses that frame to adapt before moving to the next, reflecting deployment without access to ground truth.

5.1 Main Results

CTTA Experiments. Tab. 1 presents the 10-rounds CTTA performance on the DrivingStereo. Our method (h) consistently outperforms all baselines across different weather conditions and adaptation rounds on DrivingStereo. Compared to MADNet-based experiments (a)-(c), which is a state-of-the-art stereo TTA method, our approach yields substantial improvements in all domains (dusky, cloudy, rainy), demonstrating stronger generalization and adaptability under continual shifts.

While full tuning approach (f) achieves performance gains in the early rounds of the experiment, it suffers from performance degradation due to error accumulation and forgetting over time. Parameter-efficient tuning methods including (e) and (h) tend to preserve source knowledge more effectively, leading to more stable performance across rounds. However, these approaches often exhibit limited adaptation performance due to the restricted capacity of parameter-efficient tuning. For instance, (h) EcoTTA–which adapts meta network–shows reduced effectiveness on structured prediction tasks such

Table 2: Performance comparison of Continual Test-time Adaptation on **DSEC** benchmark over 10 rounds. To save space, only 1st and 10th round scores are written. **Bold** denotes best and *AT* denotes our method with dense pseudo-label D_{teacher}.

Method	Round / Condition / Adapt.	1								10								All ↓ Mean	
		Night#1		Night#2		Night#3		Night#4		Night#1		Night#2		Night#3		Night#4			
		D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE
MADNet 2	(a) no adapt.	8.38	1.80	14.71	2.37	11.00	1.86	11.82	1.85	8.38	1.80	14.71	2.37	11.00	1.86	11.82	1.85	11.48	1.97
	(b) FT	4.7	1.24	7.79	1.49	5.97	1.31	6.33	1.31	3.57	1.11	7.34	1.4	5.62	1.25	5.8	1.24	5.71	1.27
	(c) MAD++	5.52	1.34	8.43	1.53	6.21	1.34	6.66	1.33	3.89	1.16	7.53	1.42	5.7	1.27	5.89	1.27	6.04	1.32
CoEx	(d) no adapt.	6.10	1.38	12.24	1.94	8.34	1.58	8.05	1.50	6.10	1.38	12.24	1.94	8.34	1.58	8.05	1.50	8.68	1.60
	(e) AdaptBN	4.96	1.22	8.47	1.49	4.59	1.11	**4.67**	**1.09**	2.97	1.04	6.07	1.27	4.32	1.1	4.45	1.11	4.54	1.13
	(f) FT	4.99	1.23	8.41	1.48	4.66	1.12	**4.67**	1.1	3.00	1.04	6.24	1.28	4.44	1.11	4.57	1.12	4.59	1.13
	(g) FT + AT	5.11	1.24	8.76	1.52	4.73	1.13	4.73	1.1	**2.87**	**1.01**	5.86	1.24	**4.02**	1.06	**4.15**	**1.07**	**4.38**	**1.10**
EcoTTA	(h) MetaNet	4.36	**1.17**	**7.17**	**1.39**	4.69	1.13	5.33	1.18	3.46	1.08	6.71	1.36	4.71	1.14	5.28	1.19	5.13	1.21
RobIA (ours)	(i) AttEx-MoE	**4.33**	**1.17**	7.92	1.46	**4.38**	**1.1**	4.79	1.12	3.00	1.03	5.69	1.23	4.23	1.09	4.61	1.11	4.47	1.12
	(j) AttEx-MoE + AT	4.45	**1.17**	8.18	1.48	4.75	1.15	4.97	1.12	2.96	**1.01**	**5.65**	**1.22**	4.11	**1.05**	4.34	1.08	4.46	1.11

as stereo depth estimation. In contrast, our approach (i) AttEx-MoE enables input-dependent expert routing and selective feature excitation, achieving both robust and adaptive test-time performance, leading to improved adaptation accuracy.

Incorporating dense pseudo-labels via the AdaptBN teacher (denoted *AT*) further enhances long-term stability. As shown in 10^{th}-round results, models trained with dense supervision ((g) and (j)) exhibit notably lower error rates after long-term adaptation compared to their sparse-only counterparts. This underscores the importance of broader supervision coverage, as our dual-source dense pseudo-label strategy helps prevent the model from focusing exclusively on the reliable regions of sparse handcrafted labels.

Tab. 2 reports results on the DSEC benchmark, which includes challenging nighttime conditions. Our method again outperforms MADNet2 and other baselines, maintaining strong adaptation performance. Notably, both FT and MoE variants benefit from AdaptBN-based supervision, showing improved accuracy and stability by round 10. For instance, in Night#4, D1-all error drops from 4.57 to 4.15 in (f) and (g), and from 4.61 to 4.34 in (i) and (j), respectively. Moreover, these results highlight that AttEx-MoE tuning provides greater adaptability than other PEFT-based methods ((e) and (h)), while achieving comparable performance to full model tuning with improved efficiency.

Table 3: Performance comparison of Continual Test-time Adaptation on **KITTI RAW** benchmark over 10 rounds. To save space, only 1st and 10th round scores are written. **Bold** denotes best and *AT* denotes our method with dense pseudo-label $D_{teacher}$.

Method	Round / Condition / Adapt.	1								10								All ↓ Mean	
		City		Residential		Campus$^{(\times 2)}$		Road		City		Residential		Campus$^{(\times 2)}$		Road			
		D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE
MADNet 2	(a) no adapt.	3.83	1.08	3.06	1.09	5.43	1.12	3.34	1.05	3.83	1.08	3.06	1.09	5.43	1.12	3.34	1.05	3.92	1.09
	(b) FT	1.21	0.92	0.96	0.93	2.06	0.82	1.12	0.86	0.91	0.88	0.79	0.91	1.56	0.75	1.01	0.84	1.13	0.85
	(c) MAD++	1.51	0.98	0.95	0.94	2.28	0.87	1.22	0.88	0.92	0.87	0.85	0.92	1.72	0.77	1.08	0.88	1.24	0.88
CoEx	(d) no adapt.	2.08	0.99	1.75	0.96	2.89	0.96	2.79	1.00	2.08	0.99	1.75	0.96	2.89	0.96	2.79	1.00	2.38	0.98
	(e) AdaptBN	1.09	**0.86**	0.82	**0.9**	1.5	0.74	1.05	**0.85**	0.66	**0.84**	0.72	**0.89**	1.25	**0.7**	0.86	0.82	0.91	**0.82**
	(f) FT	1.04	**0.86**	0.82	0.92	1.33	0.73	**0.97**	0.86	0.63	**0.84**	0.68	0.90	1.16	0.71	0.81	**0.81**	0.86	**0.82**
	(g) FT + AT	1.06	0.87	**0.79**	0.91	**1.27**	**0.72**	1.03	**0.85**	**0.55**	**0.84**	**0.6**	**0.89**	**1.13**	**0.7**	**0.8**	0.82	**0.82**	**0.82**
EcoTTA	(h) MetaNet	**0.94**	0.87	0.96	0.92	1.65	0.77	1.29	0.87	1.24	0.92	1.17	0.95	1.78	0.83	1.69	0.91	1.33	0.88
RobIA (ours)	(i) AttEx-MoE	1.18	0.88	0.97	0.92	1.45	0.75	1.26	0.86	0.76	0.85	0.8	**0.89**	1.32	0.74	1.07	0.83	1.06	0.84
	(j) AttEx-MoE + AT	1.21	0.89	1.04	0.92	1.51	0.75	1.41	0.86	0.78	0.86	0.77	**0.89**	1.36	0.75	1.19	0.84	1.09	0.84

Tab. 3 reports the CTTA performance on the KITTI RAW. Our method outperforms MADNet2, a state-of-the-art stereo TTA method, maintaining strong adaptation performance throughout the experiment. In the case of KITTI RAW, more than 90% of the handcrafted pseudo-labels are considered reliable, as the dataset primarily consists of daytime urban scenes that are less challenging for stereo matching compared to weather-affected or nighttime scenes. This leads to relatively stable adaptation compared to other benchmarks. However, (g) FT + AT benefit from the additional dense

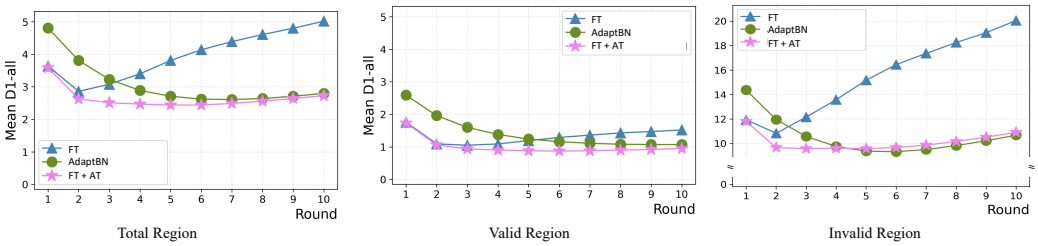


Figure 2: **D1-all error rate in different pseudo-label regions.** D1-all error rate over 10 adaptation rounds in different pseudo-label regions. We separate evaluation into (left) the entire image, (middle) regions with valid handcrafted pseudo-labels, and (right) regions without reliable supervision (invalid).

pseudo-labels, which improve model adaptability by further expanding the supervision coverage, resulting in better performance by round 10.

5.2 Analyses

Performance degradation due to Sparse Pseudo-labels. To understand the limitations of sparse handcrafted pseudo-labels, we evaluate model performance under continually shifting weather conditions by separately analyzing regions with reliable (valid) and unreliable (invalid) pseudo-labels with higher learning rate($2e{-}6$). As shown in Fig. 2, models trained solely with sparse supervision show a stable reduction in error within valid regions but suffer from sharp performance degradation in invalid regions—even when revisiting previously seen domains.

In contrast, the model adapted with AdaptBN progressively exhibit consistent improvement across both valid and invalid regions over time. Notably, when incorporating our dual-source dense pseudo-labels (denoted as *AT*), we observe that the error rate in invalid regions, which previously increased due to lack of supervision, is significantly mitigated. This indicates that the AdaptBN teacher provides reliable supervisory signals even in regions previously lacking ground-truth guidance, effectively supporting generalization across the entire image.

Table 4: Computational Cost Analysis. The average computational cost and error rate during 10 round experiments of DrivingStereo dataset.

Method	Adapt.	#_Trainable (M)	Mem. (MB)	Runtime (ms)	Mean↓ D1-all (%)	Mean↓ EPE (px)
MADNet 2	(a) no adapt.	-	179	11	10.44	1.70
	(b) FT	3.22	276	31	6.13	1.73
	(c) MAD++	3.22	398	18	5.86	1.41
CoEx	(d) no adapt.	-	694	23	5.56	1.26
	(e) AdaptBN	0.03	2704	139	3.08	0.94
	(f) FT	2.73	2744	255	3.04	0.92
	(g) FT + EMA	2.73	2835	267	2.96	0.92
	(h) FT + AT	2.79	5469	378	2.93	0.91
RobIA (ours)	(i) AttEx-MoE	1.19	1392	97.34	2.98	0.97
	(j) AttEx-MoE + AT	1.22	4096	204	2.77	0.91

Table 5: Ablation study on AttExMoE Architecture. The average D1-all and EPE error rate during 10 round experiments of DrivingStereo dataset.

Router	**Activation**	**Mean↓** D1-all	EPE
Shallow Embedding	ReLU	3.16	0.96
	Sigmoid	3.11	0.96
GAP	ReLU	4.06	1.06
	Sigmoid	3.27	0.95
Self-attention	ReLU	3.61	1.13
	Sigmoid	3.09	0.94
Column-wise Self-attention	ReLU	4.11	1.07
	Sigmoid	3.20	0.97
Row-wise Self-attention	ReLU	3.77	1.09
	Sigmoid (ours)	2.98	0.97

Computational Cost Tab. 4 provides a comparative analysis of computational cost and adaptation performance. All runtime and memory measurements were recorded on an NVIDIA RTX 3090 GPU. MADNet2, while designed for test-time efficiency, shows consistent performance degradation under continual domain shifts, indicating limited robustness in dynamic settings. In contrast, our AttEx-MoE tuning (i) achieves comparable or superior accuracy with approximately half the number of trainable parameters and reduced memory usage compared to full model tuning methods such as (f) FT and (g) FT + EMA. (j) achieves the best overall mean D1-all and EPE, while maintaining a reasonable computational cost, demonstrating the effectiveness of input-aware expert selection for resource-efficient adaptation.

Table 6: Ablation study on λ.

	1						10						All↓	
	dusky		**cloudy**		**rainy**		**dusky**		**cloudy**		**rainy**		**Mean**	
λ	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE
(a) 0.05	4.31	1.03	2.54	0.85	4.16	1.07	3.29	0.9	2.45	0.83	3.25	1.03	2.97	0.91
(b) 0.1	4.28	1.03	2.4	0.84	4.54	1.16	2.4	0.84	2.24	0.82	3.02	1.00	2.77	0.91
(c) 0.2	4.56	1.05	2.75	0.88	4.76	1.17	2.69	0.85	2.23	0.8	2.82	0.96	2.84	0.90
(d) 0.3	4.69	1.07	2.88	0.89	5.39	1.27	2.44	0.83	2.14	0.8	2.84	0.95	2.89	0.91

5.3 Ablation Studies

AttEx-MoE architecture. Tab. 5 presents an ablation study on the architectural components of AttEx-MoE. We compare different gating network designs, including the type of router and the activation function (ReLU vs. Sigmoid), under the same training setting.

Shallow embedding and GAP-based routing result in higher error rates, supporting our observation that these methods lack sufficient spatial context for precise expert selection. Self-attention routers improve performance, but full 2D attention introduces additional computation without significant gains. Although column-wise self-attention does not leverage epipolar geometry, it partially preserves spatial structure and maintains comparable efficiency. However, its higher error rates further validate our design choice of row-wise attention. ReLU activation—commonly used to enforce sparsity—consistently underperforms across all router types, likely due to reduced expressivity under frozen backbones in the PEFT setting.

Our final design, combining row-wise self-attention with sigmoid-based expert excitation, achieves the best overall results. The router effectively captures structured global context along epipolar lines, and sigmoid activation enables more expressive and stable expert modulation. These results confirm the effectiveness of our Attend-and-Excite design for robust and efficient instance-aware adaptation.

Loss weight λ for $L_{teacher}$. We ablated the loss weight λ in Tab. 6. As shown in Tab. 6, we evaluated a range of λ values over 10 CTTA rounds on the DrivingStereo sequence, using AttEx-MoE as the student model. The best performance is achieved with (b) $\lambda = 0.1$, which is also adopted in the main experimental results. Since the teacher model needs to newly adapt to the target domain and the effect of proxy supervision is relatively more important in the early stages, the larger λ values tend to limit the model's adaptability at the early rounds of the adaptation. However, as teacher predictions become more accurate over time, the larger λ leads to better performance in later rounds. In contrast, smaller values behave similarly to single-source supervision, resulting in higher error rates at the end of adaptation.

6 Conclusion

We presented **RobIA**, a robust and instance-aware framework for continual test-time adaptation (CTTA) in stereo depth estimation. RobIA addresses key challenges posed by dynamic domain shifts and sparse supervision through two core components: AttEx-MoE, a lightweight Mixture-of-Experts module guided by epipolar-aware self-attention, and a Robust AdaptBN Teacher that complements handcrafted pseudo-labels for generating dense supervision. This design enables flexible, input-specific adaptation while maintaining computational efficiency. Extensive experiments across dynamically shifting target domains demonstrate that RobIA consistently outperforms existing methods, highlighting the importance of instance-aware adaptation and hybrid supervision strategies for reliable deployment of stereo depth models in real-world settings.

Limitations and Future Work. While RobIA demonstrates strong performance in CTTA, it has certain limitations. Although AttEx-MoE offers input-aware adaptation, its reliance on predefined expert structures may limit flexibility in highly heterogeneous environments. Future work includes online expert refinement to further improve adaptation performance in long-term deployment scenarios.

Acknowledgements

This work was supported by IITP grant funded by MSIT (No. RS-2022-00155966: AI Convergence Innovation Human Resources Development (Ewha Womans University) and No. RS-2021-II212068: AI Innovation Hub).

References

[1] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017.

[2] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

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

[4] 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.

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

[6] Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. *arXiv preprint arXiv:2403.07366*, 2024.

[7] Alessio Tonioni, Oscar Rahnama, Thomas Joy, Luigi Di Stefano, Thalaiyasingam Ajanthan, and Philip HS Torr. Learning to adapt for stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9661–9670, 2019.

[8] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi DI Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 195–204, 2019.

[9] 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.

[10] Yulu Gan, Yan Bai, Yihang Lou, Xianzheng Ma, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7595–7603, 2023.

[11] Jiaming Liu, Senqiao Yang, Peidong Jia, Renrui Zhang, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. Vida: Homeostatic visual domain adapter for continual test time adaptation. *arXiv preprint arXiv:2306.04344*, 2023.

[12] Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11920–11929, 2023.

[13] Hyewon Park, Hyejin Park, Jueun Ko, and Dongbo Min. Hybrid-tta: Continual test-time adaptation via dynamic domain shift detection, 2024. URL `https://arxiv.org/abs/2409.08566`.

[14] Dahuin Jung, Dongyoon Han, Jihwan Bang, and Hwanjun Song. Generating instance-level prompts for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11847–11857, 2023.

[15] Hayeon Jo, Hyesong Choi, Minhee Cho, and Dongbo Min. iconformer: Dynamic parameter-efficient tuning with input-conditioned adaptation. *arXiv preprint arXiv:2409.02838*, 2024.

[16] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pages 807–814. IEEE, 2005.

[17] Matteo Poggi, Alessio Tonioni, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Continual adaptation for deep stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4713–4729, 2021.

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[19] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. *Advances in Neural Information Processing Systems*, 35:27253–27266, 2022.

[20] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551, 2020.

[21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[22] Kwonyoung Kim, Jungin Park, Jiyoung Lee, Dongbo Min, and Kwanghoon Sohn. Pointfix: Learning to fix domain bias for robust online stereo adaptation. In *European Conference on Computer Vision*, pages 568–585. Springer, 2022.

[23] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[24] Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Unsupervised adaptation for deep stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1605–1613, 2017.

[25] Filippo Aleotti, Fabio Tosi, Li Zhang, Matteo Poggi, and Stefano Mattoccia. Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 614–632. Springer, 2020.

[26] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[27] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[28] Hanrong Ye and Dan Xu. Taskexpert: Dynamically assembling multi-task representations with memorial mixture-of-experts. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21828–21837, 2023.

[29] Yuqi Yang, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, and Bo Li. Multi-task dense prediction via mixture of low-rank experts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 27927–27937, 2024.

[30] Minh Le, Huy Nguyen, Trang Nguyen, Trang Pham, Linh Ngo, Nhat Ho, et al. Mixture of experts meets prompt-based continual learning. *Advances in Neural Information Processing Systems*, 37:119025–119062, 2024.

[31] Sen Lin. Theory on mixture-of-experts in continual learning. In *2024 Fall Central Sectional Meeting*. AMS.

[32] Daeun Lee, Jaehong Yoon, and Sung Ju Hwang. Becotta: Input-dependent online blending of experts for continual test-time adaptation. *arXiv preprint arXiv:2402.08712*, 2024.

[33] Hang Guo, Tao Dai, Yuanchao Bai, Bin Chen, Xudong Ren, Zexuan Zhu, and Shu-Tao Xia. Parameter efficient adaptation for image restoration with heterogeneous mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:13522–13547, 2024.

[34] Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E Gonzalez. Deep mixture of experts via shallow embedding. In *Uncertainty in artificial intelligence*, pages 552–562. PMLR, 2020.

[35] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6197–6206, 2021.

[36] Bo Li, Yifei Shen, Jingkang Yang, Yezhen Wang, Jiawei Ren, Tong Che, Jun Zhang, and Ziwei Liu. Sparse mixture-of-experts are domain generalizable learners. *arXiv preprint arXiv:2206.04046*, 2022.

[37] 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.

[38] Marc Botet Colomer, Pier Luigi Dovesi, Theodoros Panagiotakopoulos, Joao Frederico Carvalho, Linus Härenstam-Nielsen, Hossein Azizpour, Hedvig Kjellström, Daniel Cremers, and Matteo Poggi. To adapt or not to adapt? real-time adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16548–16559, 2023.

[39] Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson W.H. Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[40] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

[41] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 899–908, 2019.

[42] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. doi: 10.1109/LRA.2021.3068942.

[43] Antyanta Bangunharcana, Jae Won Cho, Seokju Lee, In So Kweon, Kyung-Soo Kim, and Soohyun Kim. Correlate-and-excite: Real-time stereo matching via guided cost volume excitation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3542–3548. IEEE, 2021.

[44] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[45] Matteo Poggi and Fabio Tosi. Federated online adaptation for deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20165–20175, 2024.

[46] Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Unsupervised domain adaptation for depth prediction from images. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2396–2409, 2019.

[47] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21919–21928, 2023.

[48] Xianda Guo, Chenming Zhang, Youmin Zhang, Wenzhao Zheng, Dujun Nie, Matteo Poggi, and Long Chen. Lightstereo: Channel boost is all you need for efficient 2d cost aggregation. *arXiv preprint arXiv:2406.19833*, 2024.

[49] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021.

[50] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16263–16272, 2022.

[51] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13941–13958, 2023.

A Implementation Details

For the warm-up process, we trained the model using the Adam optimizer for 10 epochs with a fixed learning rate of 5e-4. During test-time adaptation (TTA), we also used Adam optimizer across all methods. The learning rate was set to 5e-6 for training the AdaptBN teacher model and 1e-5 for MADNet2 [45]. Tab. 7 reports the learning rates of student models used in experiments. For full-tuning methods, we used relatively smaller learning rates—approximately 10 times lower than those used in efficient tuning methods such as AdaptBN, MetaNet [12], and our AttEx-MoE. This choice is motivated by the observation that large learning rates in full-tuning settings significantly impair generalization performance.

Our baseline model is CoEx [43], a compact and real-time stereo matching network. The feature extractor in CoEx is composed of four scale levels, with upsampling modules built using long skip connections at each scale. To implement EcoTTA [12] in the CTTA setting for stereo matching , we inserted its meta network into the MobileNetV2 backbone of CoEx at four scale-level blocks (K=4).

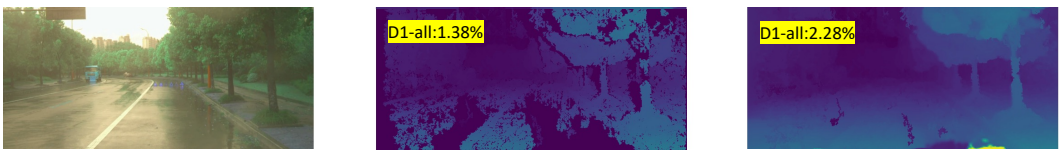
Table 7: Learning rates.

Dataset	Exp.	Learning Rate
Full tuning		
DrivingStereo	CTTA	5e-7
DSEC	CTTA	2e-6
KITTI RAW	CTTA	1e-5
Efficient tuning		
DrivingStereo	CTTA	5e-6
DSEC	CTTA	3e-5
KITTI RAW	CTTA	1e-4

B Additional Analysis and Ablation Studies

Pseudo-label visualization. Fig. 3 visualizes the pseudo-labels generated by different methods after ten rounds of adaptation from the rainy sequences of the DrivingStereo dataset. On each, we reported disparity maps of the student model and the corresponding error rate. Note that, for sparse pseudo-labels, we only measured error for regions where both the pseudo-label and the ground truth are valid.

The handcrafted sparse pseudo-labels often lack supervisory signals in challenging regions for stereo matching, such as reflective surface, low-texture regions, or occlusions. Furthermore, this issue extends to image borders, where the hand-crafted stereo algorithms (*e.g.*, SGM) often struggles due



(a) Left image (b) Sparse pseudo-label (c) Dense pseudo-label

(d) AttEx-MoE (e) AttEx-MoE + AT

Figure 3: Pseudo-labels (top row) and predictions (bottom row) after ten adaptation rounds. We visualize the sparse handcrafted pseudo-label (b) and the dense pseudo-label using the AdaptBN teacher (c), and the student predictions of AttEx-MoE trained with sparse (d) and dense (e) supervision. (a) shows the input left image.

Table 8: Ablation study on Pseudo Supervision. Average D1-all and EPE over 10 adaptation rounds on DrivingStereo. We use single- or dual-source supervision based on whether only one or both of handcrafted and learned signals are used.

		1						**5**						**10**						All↓	
		dusky		**cloudy**		**rainy**		**dusky**		**cloudy**		**rainy**		**dusky**		**cloudy**		**rainy**		**Mean**	
Source	**Method**	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE
Single	(a) proxy	4.01	1.01	2.4	0.84	4.44	1.13	2.54	0.86	2.23	0.83	3.89	1.2	2.72	0.88	2.29	0.84	3.89	1.22	2.98	0.97
	(b) photometric	4.3	1.04	2.45	0.85	4.76	1.18	3.21	0.95	2.75	0.89	4.74	1.25	3.67	1.01	2.91	0.91	4.45	1.22	3.55	1.02
	(c) AdaptBN	5.5	1.13	3.48	0.98	6.91	1.5	3.73	0.97	2.83	0.88	5.65	1.3	3.15	0.92	2.55	0.85	5.03	1.21	4.16	1.06
Dual	proxy																				
	+ (d) Source	4.28	1.03	2.4	0.84	4.62	1.18	2.75	0.88	2.3	0.82	4.0	1.13	2.78	0.87	2.31	0.82	3.82	1.1	3.09	0.95
	+ (e) EMA	4.27	1.03	2.38	0.84	4.56	1.16	2.84	0.88	2.26	0.82	4.09	1.16	3.11	0.91	2.32	0.82	4.03	1.15	3.16	0.96
	+ (f) photometric [46]	4.02	1.01	2.4	0.84	4.44	1.13	2.54	0.86	2.23	0.83	3.87	1.18	2.69	0.87	2.29	0.84	3.87	1.22	2.98	0.96
	+ (g) **AT (ours)**	4.28	1.03	2.4	0.84	4.54	1.16	2.46	0.85	2.2	0.81	3.27	1.01	2.4	0.84	2.24	0.82	3.02	1.0	2.77	0.91

to the lack of neighboring pixels for reliable cost aggregation. As a result, models relying solely on these handcrafted labels tend to exhibit reduced adaptability in CTTA. In contrast, our dense pseudo-labels provide complete spatial coverage, allowing the model to adapt effectively even in unreliable regions. This confirms the role of dense, teacher-guided pseudo-supervision in enhancing spatial robustness under continual adaptation settings.

Pseudo-supervision ablation. Tab. 8 presents an ablation study comparing various pseudo-supervision strategies using our AttEx-MoE based Parameter-efficient tuning method. Single-source supervision methods (a–c) perform worse overall. (a) Proxy supervises the model effectively, but the limited guidance of sparse handcrafted pseudo-supervision means that the model's performance degradation later in the round. (b) photometric loss leads to increasingly higher error rates due to noisy guidance. (c) AdaptBN supervision alone shows more stable error reduction but underperforms when used without proxy supervision, as the teacher model itself requires time to adapt to the target domain.

Dual-source variants (d–g) combine proxy labels with additional supervision to improve generalization across the entire input space. Although using a fixed source model (d) enhances stability, it lacks adaptability under distribution shift. (e) EMA initially helps, but tends to propagate student errors. Following [46], (f) augments proxy label with photometric and smoothing losses to compensate for its sparsity, but does not yield meaningful improvement over (a) proxy-only. In contrast, our method (g) leverages a robust AdaptBN teacher to provide dense supervision, achieving the best overall results (D1-all 2.77% , EPE 0.91), maintaining the performance in long-term and continuously changing conditions. These findings underscore the importance of dense, complementary supervision in overcoming the limitations of sparse pseudo-labels during continual adaptation.

C Additional Experiments Results

Table 9: Performance comparison of Continual Test-time Adaptation on DrivingStereo benchmark over 10 rounds for IGEV-Stereo and LightStereo backbone. **Bold** denotes best and *AT* denotes our method with dense pseudo-label D_{teacher}.

		Mean↓ DrivingStereo	
Backbone	**Adapt.**	D1-all (%)	EPE (px)
IGEV-Stereo	(a) no adapt.	7.26	1.32
	(b) AdaptBN	3.87	0.98
	(c) FT	3.56	0.96
	(d) AttEx-MoE	3.15	0.92
	(e) AttEx-MoE + AT	**2.76**	**0.87**
LightStereo	(a) no adapt.	18.59	3.11
	(b) AdaptBN	5.95	1.74
	(c) FT	5.91	1.72
	(d) AttEx-MoE	5.76	1.28
	(f) AttEx-MoE + AT	**4.95**	**1.11**

Table 10: Performance comparison of Continual Test-time Adaptation on DrivingStereo and KITTI RAW benchmark over 10 rounds for MADNet2 backbone. **Bold** denotes best and *AT* denotes our method with dense pseudo-label D_{teacher}.

		Mean↓				
		DrivingStereo		KITTI		Runtime
Backbone	**Adapt.**	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)	(ms)
MADNet2	(a) no adapt.	10.44	1.70	3.92	1.09	11
	(b) FT	6.13	1.73	1.13	**0.85**	31
	(c) MAD++	5.86	1.41	1.24	0.88	18
	(d) AttEx-MoE	5.53	1.20	1.12	0.86	23
	(e) AttEx-MoE + AT	**4.83**	**1.09**	**1.10**	**0.85**	49

As shown in Tab. 9, beyond the 3D cost aggregation model CoEx, we also evaluated our method on *IGEV-Stereo* [47], a widely used iterative-refinement backbone, and on *LightStereo* [48], a lightweight real-time network with 2D cost aggregation. On all three architectures, our RobIA with AttEx-MoE and AdaptBN-Teacher consistently improves accuracy, demonstrating that our approach generalizes well to various stereo architectures.

Our components remain consistently effective when applied to IGEV-Stereo. Without adaptation, (a) IGEV-Source suffers from significant domain shift, yielding high D1-all errors 7.26% and EPE 1.32. While tuning only BN parameters ((b) AdaptBN) or full fine-tuning ((c) FULL++) reduces error, both approaches still struggle (Da-all (b) 3.87%, (c) 3.56%). In contrast, (d) our AttEx-MoE based PEFT lowers D1-all to 3.15%, and our RobIA implementation ((e) AttEx-MoE + AT) further improves it to 2.76%. This demonstrates that our plug-and-play modules effectively enhance generalization even on top-performing backbones.

Tab. 9 also shows that RobIA generalizes well to LightStereo. Without adaptation, (a) LightStereo-Source performs poorly due to severe domain shift (18.59%). (b) AdaptBN alone improves results to 5.95%, yet falls short on dynamic scenes. Our AttEx-MoE module (e) further reduces the D1-all error to 5.76%, and RobIA (f) achieves the best performance (D1-all 4.95%, EPE 1.11), confirming the robustness and plug-and-play nature of our method even on compact backbones. These results with IGEV-Stereo and LightStereo demonstrate that RobIA is effective across both strong and lightweight backbones, and maintains robustness under significant domain shifts.

Furthermore, we re-implemented our method, including AttEx-MoE and the AdaptBN teacher, on top of the *MADNet2* [8] encoder, which is also used by MAD++. Since MADNet2 lacks normalization layers, we inserted BatchNorm layers after each convolution to enable AdaptBN and maintain architectural consistency.

As shown in Table 10, on KITTI, which presents relatively mild domain shifts, (b) FULL++ achieves D1-all 1.13% by updating all layers, while (d) our AttEx-MoE method achieves a better result (1.12%) with fewer parameters. (e) RobIA, which combines AttEx-MoE with the AdaptBN teacher, maintains a strong KITTI score of 1.1%. On DrivingStereo, which exhibits stronger domain shifts, (b) FULL++ struggles with large domain shifts (6.13%), and (c) MAD++ provides only minor improvement (5.86%). In contrast, (d) our AttEx-MoE method further reduces the error to 5.53%, and (e) RobIA achieves the best result at 4.83%. These results show that our input-aware AttEx-MoE gate, combined with dual-source supervision, matches full tuning on stable domains, and significantly outperforms both full and modular adaptation approaches under large distribution shifts.

Table 11: Performance comparison for Sequential Continual Test-time Adaptation across different datasets, from KITTI RAW (rounds 1–5) to DrivingStereo (rounds 6–10). **Bold** denotes best and *AT* denotes our method with dense pseudo-label D_{teacher}.

		KITTI Round 1 → 5		**DrivingStereo (after KITTI)** Round 6 → 10		ALL	
Backbone	**Adapt.**	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)
CoEx	(a) AdaptBN	1.11	0.84	5.90	1.70	3.16	1.21
	(b) AttEx-MoE	1.11	0.85	4.22	1.08	2.44	0.95
	(c) AttEx-MoE + AT	1.15	0.84	2.66	0.90	**1.80**	**0.87**

Table 12: Performance comparison of Test-time Adaptation on **KITTI RAW** benchmarks. **Bold** denotes best and *AT* denotes our method with dense pseudo-label D_{teacher}.

Condition		(8027 frames) City		(28067 frames) Residential		(1149 × 2 frames) Campus$^{(\times 2)}$		(5674 frames) Road		Mean	
Method	**Adapt.**	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE	D1-all	EPE
RAFT-Stereo [49]	(a) no adapt.	1.55	0.89	1.77	0.82	2.53	0.89	1.77	0.85	1.90	0.86
CREStereo [50]	(b) no adapt.	1.87	0.99	1.71	0.89	3.21	1.07	2.00	0.89	2.20	0.96
IGEV-Stereo [47]	(c) no adapt.	2.26	1.00	2.56	0.94	3.01	0.99	2.52	0.96	2.58	0.97
UniMatch [51]	(d) no adapt.	2.66	1.13	3.20	1.10	3.10	1.13	2.26	1.08	2.81	1.11
MADNet 2 [45]	(e) no adapt.	4.04	1.10	4.05	1.03	6.07	1.29	4.01	1.08	4.54	1.13
	(f) FT	1.23	0.90	1.05	0.80	2.39	0.92	1.02	0.83	1.42	0.86
	(g) MAD++	1.39	0.93	1.16	0.83	2.88	1.00	1.14	0.85	1.64	0.90
CoEx [43]	(h) no adapt.	2.66	1.07	2.66	0.99	3.65	1.11	2.46	0.98	2.86	1.04
	(i) FT	0.83	**0.84**	0.75	0.76	1.49	**0.8**	**0.75**	**0.79**	0.96	0.80
	(j) FT + AT	**0.8**	**0.84**	**0.66**	**0.74**	**1.45**	**0.8**	0.79	0.8	**0.93**	**0.80**
RobIA (ours)	(k) AttEx-MoE	0.99	0.87	0.96	0.78	1.6	0.82	0.9	0.81	1.11	0.82
	(l) AttEx-MoE + AT	1.05	0.87	0.91	0.78	1.56	0.81	0.94	0.83	1.12	0.82

We simulated a sequential CTTA scenario where the model first adapts to KITTI for 5 rounds, followed by adaptation to DrivingStereo for another 5 rounds, mimicking a realistic progression from a stable to a more challenging domain. Results are shown in Tab. 11. (b) and (c) show similar error rates with (a) on KITTI, but they achieve substantially lower error on DrivingStereo. (a), despite strong performance of 1.11% on KITTI, fails to adapt in the second phase (5.90%). These results highlight the robustness of our AttEx-MoE with AdaptBN teacher, which generalizes better under sequential, cross-domain conditions, mirroring realistic deployment settings.

TTA Experiments. Previous studies have demonstrated effective performance under TTA, which motivates us to assess whether our approach, designed for CTTA, also generalizes well in this setting. To evaluate the effectiveness of our method under standard test-time adaptation (TTA) settings, we conducted experiments on three real-world stereo datasets: KITTI RAW, DrivingStereo, and DSEC. The results are reported in Tab. 12, Tab. 13, and Tab. 14 for KITTI RAW, DrivingStereo, and DSEC, respectively. Following prior work [45], we compared against MADNet2 [45], and several state-of-the-art stereo models known for strong generalization performance. The stereo models only trained on synthetic source datasets (a–d) exhibit significant performance degradation on target domains due to domain shifts. While adaptation-based methods generally improve accuracy, performance gains remain limited for efficiency-oriented state-of-the-art TTA methods (e–g).

Our efficient tuning method (k), which employs AttEx-MoE with input-aware feature excitation, achieves higher accuracy than prior TTA baselines, while remaining comparable to full model tuning and reducing computational cost, as discussed in the main paper. Results supervised by our dense pseudo-labels are shown in (j) and (l). Notably, substantial improvements by dense supervisions with AdaptBN Teacher are observed on DrivingStereo and DSEC, where pseudo-label sparsity is higher due to challenging conditions including adverse weather and nighttime imagery.

These findings demonstrate that our method is not only effective in the proposed continual adaptation scenario but also exhibits effective adaptability in standard TTA settings involving long-term adaptation within individual domains.

Qualitative Comparison of CTTA Results. We report qualitative comparisons of disparity maps predicted by various models and supervision strategies evaluated in our CTTA experiments. Figs. 4 and 5 show examples from the cloudy and rainy sequences of the DrivingStereo dataset. For each example, predictions are visualized at round 1, 5, and 10 to highlight temporal adaptation behaviors.

We observed that our method consistently improves predictions across rounds, particularly in challenging regions such as reflective surfaces, low-texture areas, and image borders. Models trained with only sparse supervision show limited generalization beyond the confident regions of the handcrafted pseudo-labels, which limits their adaptability in less confident areas over time. In contrast, dense pseudo-supervision enables broader coverage and leads to stable improvements across the entire image, demonstrating stronger generalization under continual domain shifts.

Table 13: Performance comparison of Test-time Adaptation on **DrivingStereo** benchmarks. **Bold** denotes best and *AT* denotes our method with dense pseudo-label D_{teacher}.

Method	Condition Adapt.	(1667 frames) dusky D1-all	dusky EPE	(1119 frames) cloudy D1-all	cloudy EPE	(4950 frames) rainy D1-all	rainy EPE	Mean D1-all	Mean EPE
RAFT-Stereo [49]	(a) no adapt.	11.52	1.59	3.08	0.88	4.18	1.02	6.26	1.16
CREStereo [50]	(b) no adapt.	17.43	3.61	7.08	1.23	4.08	1.07	9.53	1.97
IGEV-Stereo [47]	(c) no adapt.	11.70	1.85	3.57	0.95	5.27	1.26	6.95	1.35
UniMatch [51]	(d) no adapt.	14.84	2.69	7.51	1.27	5.78	1.25	9.38	1.74
MADNet 2 [45]	(e) no adapt.	16.47	3.03	13.16	1.66	6.72	1.35	12.12	2.01
	(f) FT	10.34	2.27	4.41	1.04	5.20	1.63	6.65	1.65
	(g) MAD++	10.06	**2.01**	5.25	1.09	4.34	1.09	6.55	1.40
CoEx [43]	(h) no adapt.	13.55	3.02	5.24	1.12	4.12	1.15	7.64	1.76
CoEx	(i) FT	8.85	2.29	3.04	0.9	3.7	1.27	5.20	1.49
	(j) FT + AT	**7.93**	2.09	2.63	**0.88**	**2.29**	**0.84**	**4.28**	**1.27**
RobIA (ours)	(k) AttEx-MoE	9.05	2.28	3.21	0.96	2.8	0.91	5.02	1.38
	(l) AttEx-MoE + AT	8.27	2.18	**2.61**	0.91	2.77	0.92	4.55	1.34

Table 14: Performance comparison of Test-time Adaptation on **DSEC** benchmarks. **Bold** denotes best and *AT* denotes our method with dense pseudo-label D_{teacher}.

Method	Condition Adapt.	(883 frames) Night#1 D1-all	Night#1 EPE	(1813 frames) Night#2 D1-all	Night#2 EPE	(2315 frames) Night#3 D1-all	Night#3 EPE	(2405 frames) Night#4 D1-all	Night#4 EPE	Mean D1-all	Mean EPE
RAFT-Stereo [49]	(a) no adapt.	13.04	3.41	21.64	4.26	10.91	1.91	10.07	1.68	13.92	2.82
CREStereo [50]	(b) no adapt.	11.34	2.38	23.48	3.19	15.37	2.39	12.42	1.75	15.65	2.43
IGEV-Stereo [47]	(c) no adapt.	9.14	1.85	11.97	1.96	12.65	2.01	10.01	1.66	10.94	1.87
UniMatch [51]	(d) no adapt.	34.29	5.43	39.80	5.32	26.75	3.29	26.29	3.28	31.78	4.33
MADNet 2 [45]	(e) no adapt.	8.94	1.97	13.86	2.32	10.63	1.83	10.55	1.69	11.00	1.95
	(f) FT	4.69	1.28	7.13	1.43	6.20	1.35	6.06	1.27	6.02	1.33
	(g) MAD++	5.66	1.43	8.39	1.53	7.91	1.50	7.79	1.39	7.44	1.46
CoEx [43]	(h) no adapt.	6.14	1.52	10.27	1.77	7.82	1.58	7.64	1.45	7.97	1.58
	(i) FT	3.55	1.1	5.14	1.2	4.39	1.11	4.41	1.07	4.37	1.12
	(j) FT + AT	3.65	1.1	**4.98**	**1.18**	**4.19**	**1.1**	**4.12**	**1.04**	**4.24**	**1.11**
RobIA (ours)	(k) AttEx-MoE	3.66	1.13	5.32	1.23	4.59	1.15	4.85	1.13	4.61	1.16
	(l) AttEx-MoE + AT	**3.51**	**1.09**	5.38	1.23	4.77	1.18	4.58	1.1	4.56	1.15

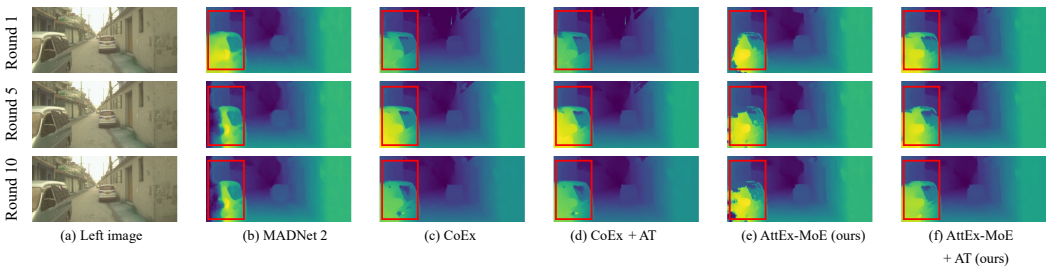


Figure 4: Qualitative results for cloudy sequences in the DrivingStereo dataset.

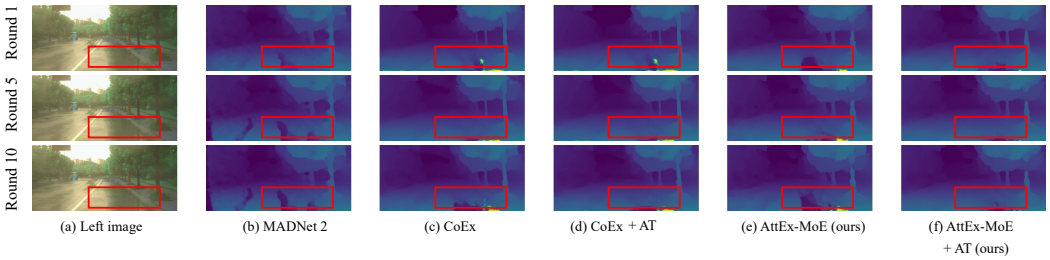


Figure 5: Qualitative results for rainy sequences in the DrivingStereo dataset.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: The claims regarding our proposed continual test-time adaptation framework for stereo depth estimation (RobIA), including AttEx-MoE and AdaptBN-based dense supervision, are clearly stated in both the abstract and introduction. These contributions are consistently supported by methodological details and experiments in Sections 4 and 5.

   Guidelines:

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

2. **Limitations**

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

Answer: [Yes]

Justification: We explicitly discuss the limitations of our method in the final paragraph of the Conclusion. In particular, we note that AttEx-MoE relies on predefined expert structures, which may limit flexibility in highly heterogeneous environments. We also outline potential future directions such as incorporating uncertainty estimation or online expert refinement to improve long-term adaptability.

Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [No]

   Justification: Our paper does not include any formal theoretical results or proofs.

   Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: Our paper describes all necessary components to reproduce the main experimental results, including model architecture (Sec. 4,5), training settings and hyperparameters (Appendix), evaluation metrics, and datasets used (Sec. 5). All experiments are conducted on publicly available datasets. We also plan to release our code to facilitate reproducibility.

Guidelines:

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

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

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

Answer: [Yes]

Justification: All datasets used in our experiments are publicly available (see Sec. 5). We will release the code and reproduction instructions.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We describe the experimental setting in Section 5. Training details such as optimizer type, learning rate, and other hyperparameters are provided in the appendix.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: We do not report error bars or confidence intervals, as each method was evaluated using a single run with fixed random seed and consistent data splits. Reporting statistical variability is a direction we plan to explore in future work.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

Answer: [Yes]

Justification: We report the GPU types (NVIDIA A6000 and RTX 3090), along with runtime and memory usage in our computational cost analysis (see Table 3).

Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: Yes–Our work fully conforms with the NeurIPS Code of Ethics.

Guidelines:

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

10. **Broader impacts**

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

Answer: [NA]

Justification: Not applicable–Our work is theoritical and does not directly affect society.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: Not applicable–Our work does not involve any data or models that have a high risk for misuse.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Yes–Our work properly credits the original owners of assets used in the paper.

    Guidelines:

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

13. **New assets**

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

    Answer: [NA]

    Justification: Not applicable–Our work does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

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

    Answer: [NA]

    Justification: Not applicable–Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

    Answer: [NA]

    Justification: Not applicable–Our work does not involve human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: This paper does not decribe the usage of LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.