Asymmetric Dual-Teacher Self-Training for Adapting Stereo Foundation Models

Zhen Liang[1,4] Feng Yang[2], Lile Cai[2], Tiankai Chen[3], Hongfeng Long[1,4], Xulei Yang[2], Enhai Liu[3], Rujin Zhao[1,4*], Xun Xu[2*]

[1] the State Key Laboratory of Optical Field Manipulation Science and Technology, Institute of Optics and Electronics, Chinese Academy of Sciences, China.

[2] Institute for Infocomm Research (I^{2}R), A*STAR, Singapore

[3] School of Computing and Artificial Intelligence, Southwest Jiaotong University, China

[4] University of Chinese Academy of Sciences, China

***Abstract*—Stereo Foundation Models have demonstrated remarkable zero-shot generalization. However, leveraging unlabeled downstream data to further boost their performance, especially under domain shifts caused by visual corruptions, remains an underexplored bottleneck. To this end, we propose Asymmetric Dual-Teacher (ADT), a novel self-training framework for the source-free domain adaptation of stereo foundation models. Unlike conventional single-teacher paradigms, which often suffer from confirmation bias when coupling pseudo-label generation with reliability estimation, ADT explicitly decouples these roles. Specifically, a slow-evolving Anchor Teacher provides stable pseudo-disparity targets to preserve robust pre-trained priors, while a fast-evolving Active Teacher captures domain-specific error patterns to generate pixel-wise reliability for gated optimization. Furthermore, we introduce Unified Consistency Reliability (UCR), which fuses complementary geometric constraints (left-right and forward-backward consistency) with photometric cues. This approach yields a conservative, soft reliability map that effectively filters noisy supervision. Extensive evaluations across five diverse benchmarks, including KITTI, Middlebury, ETH3D, and challenging scenarios like Booster and DrivingStereo, demonstrate that our method achieves state-of-the-art performance in source-free adaptation.**

***Index Terms*—Stereo Matching, Unsupervised Domain Adaptation, Self-Training.**

I. Introduction

Stereo matching has recently witnessed a paradigm shift with the emergence of Stereo Foundation Models. By leveraging million-scale synthetic training datasets and transformer-based architectures, models such as FoundationStereo [1] and Stereo Anywhere [2] have achieved impressive zero-shot generalization capabilities. Unlike traditional deep stereo networks trained from scratch [3], [4], those foundation models encapsulate rich geometric priors, providing a strong baseline across diverse scenarios.

Despite strong zero-shot generalization, stereo foundation models may still degrade noticeably when deployed in real downstream environments that exhibit unknown distribution shift, such as sensor drift and adverse weather (fog, rain, low contrast) [5]–[7]. As illustrated in Fig. 1, these corruptions can induce visible artifacts and structural inconsistencies in the predicted disparities. In practice, acquiring ground-truth disparities for each downstream domain is expensive or infeasible. Instead, only unlabeled stereo pairs are available after deployment. This naturally raises a key question: *can unlabeled downstream data be leveraged to recover performance and improve robustness of a pre-trained stereo foundation model under domain shift?*

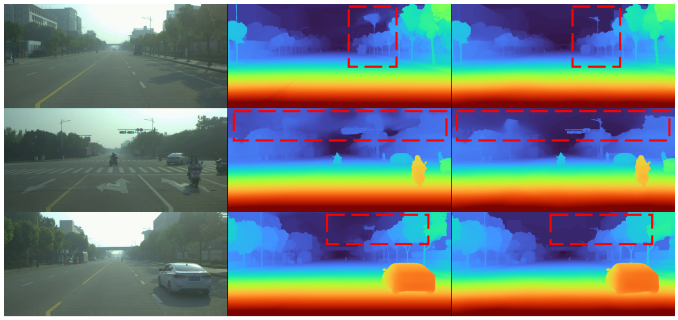

Fig. 1. Despite strong zero-shot generalization, the FoundationStereo baseline exhibits visible artifacts and structural inconsistencies under adverse weather conditions, as highlighted by the red dashed boxes.

Self-training has been proven effective for leveraging unlabeled data via a teacher-student paradigm, where a teacher generates pseudo-disparities, and the student is optimized to regress to them. A few recent works [8]–[10] attempt to apply self-training to cost-volume-based stereo matching methods, where the predicted disparities are based on the cost-volume matrix, and the reliability of teacher-predicted pseudo-labels is determined by the matching probability distribution of the cost volume. In practice, this reliability is typically used to filter or reweight pseudo-labels, e.g., via hard thresholding or loss weighting, to mitigate the impact of noisy supervision. However, existing self-training methods either simultaneously learn from both labeled and unlabeled data [8], [9] or fail to demonstrate effectiveness on foundation models [10].

We highlight two challenges when adapting self-training-based methods to enhance disparity foundation models. First, stereo matching is a dense regression problem, making pseudo-

* Correspondence to Ruijin Zhao <zhaorj@ioe.ac.cn> and Xun Xu <xu_xun@a-star.edu.sg>.

This research work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021), the Chang'e-8 Terrain Camera Project (No. YA24K306), and the Sichuan Science and Technology Program (Grant No. 2025ZNSFSC1504).

label reliability harder to calibrate than in classification. Second, existing self-training often couples pseudo-label generation and reliability estimation within a single teacher. For example, the approach [10] uses a single teacher model to predict both pseudo labels and consistency-based reliability weight for weighted self-training, where higher cross-scale or iterative consistency implies higher reliability. However, this coupled design is prone to confirmation bias: under severe corruptions, the teacher can be internally consistent yet incorrect, and the student is encouraged to reinforce these errors through self-training. Moreover, prediction consistency alone is insufficient to reliably estimate pseudo-label reliability under severe corruptions. These limitations motivate decoupling pseudo-label generation from reliability estimation and incorporating additional cues beyond internal consistency for conservative reliability calibration.

To address these challenges, we propose an Asymmetric Dual-Teacher (ADT) framework that explicitly decouples label generation from reliability estimation. Specifically, we utilize two distinct teachers: an Anchor Teacher and an Active Teacher. The Anchor Teacher provides stable regression targets through slow evolution, leveraging foundation priors to prevent the student from collapsing into degenerate solutions. Meanwhile, the Active Teacher evolves rapidly to produce reliability weights. Instead of offering direct supervision, it identifies domain-specific error patterns and highlights trustworthy regions to effectively guardrail the optimization process.

Furthermore, we build a Unified Consistency Reliability (UCR) module that combines multiple cues to score label confidence. While existing methods often rely on isolated confidence metrics [11], [12], UCR integrates geometric consistency (including left-right disparity agreement and occlusion checks [13]) with photometric consistency. By fusing these multi-modal cues into a soft-reliability map, we provide a robust, physics-based external reference for pixel-level confidence gating.

Our main contributions are summarized as follows:

- We address the challenge of adapting stereo foundation models to downstream tasks using unlabeled data only by improving pseudo-label quality in self-training, through an Asymmetric Dual-Teacher (ADT) framework that decouples label generation from reliability estimation.
- We build a Unified Consistency Reliability (UCR) mechanism that integrates geometric and photometric cues to perform fine-grained pseudo-label filtering, effectively suppressing noisy supervision under domain shift.
- Our proposed method ADT demonstrates robust and consistent performance across diverse and challenging stereo benchmarks, including adverse weather conditions and complex real-world scenes.

II. Related Work

Stereo Matching: Stereo matching has evolved rapidly with deep learning, mostly driven by cost-volume regression pipelines [4], [14] and iterative refinement architectures [8], [15]. While supervised models achieve strong in-domain performance, they often suffer notable degradation under domain shift, e.g., changes in sensor characteristics, lighting, weather, or scene composition [16]. Recent stereo foundation models [1], [2] improve zero-shot robustness by pre-training on large-scale synthetic/mixed data and distilling transferable geometric priors. Nevertheless, foundation models may still fail on challenging downstream tasks, such as adverse weather conditions, where stereo correspondence becomes ambiguous, and model predictions become error-prone [6]. This motivates adapting foundation stereo models using available target-domain unlabeled data, where the key challenge becomes how to obtain reliable pseudo-supervision without ground truth.

Domain Adaptation for Stereo Matching: To bridge the domain gap without ground truth, self-supervised adaptation has become the dominant paradigm. Early methods rely on photometric consistency losses [12], [13], but often fail in textureless or occluded regions. Later approaches, such as DSMNet [17] and AdaStereo [18], improve robustness through domain-invariant feature alignment or by leveraging auxiliary reconstruction tasks. However, they typically filter out low-confidence predictions by discarding precisely the hard samples most valuable for adaptation and thus suffer from sparse supervision. In contrast to discarding uncertain labels, our framework corrects them by fusing complementary signals from decoupled teachers.

Self-Training for Adaptation: The teacher-student framework is the most effective self-training paradigm. In the realm of stereo matching, a teacher, typically an Exponential Moving Average (EMA) of the student, generates pseudo-disparities to guide training [6], [9], [10]. However, single-teacher frameworks face a fundamental Stability–Plasticity Dilemma [19]: a slow EMA update (high stability) lags in correcting domain-specific errors, while a fast update (high plasticity) risks catastrophic forgetting and confirmation bias from overfitting to noisy labels [20]. This trade-off is especially problematic in foundation model adaptation, where preserving pre-trained knowledge is as crucial as acquiring new features. Existing multi-model methods such as Co-Teaching [21] and Dual-Student [22], use symmetric architectures that filter errors via consensus but do not explicitly balance prior retention with adaptation. In contrast, we propose an Asymmetric Dual-Teacher Framework that decouples roles. The Anchor Teacher preserves foundational stability, while the Active Teacher adapts responsively to the target domain, effectively resolving the update dilemma inherent in conventional self-training.

III. Methodology

A. Problem Definition

Given a rectified stereo pair (I_L, I_R), stereo matching aims to estimate a dense disparity map D. We consider a source-free adaptation setting for stereo foundation models, where a model pre-trained on large-scale synthetic data is adapted to an unlabeled target domain $\mathcal{D}_t$. In this setting, only unlabeled stereo pairs from $\mathcal{D}_t$ are available, while neither source data nor ground-truth target disparities are accessible. The objective

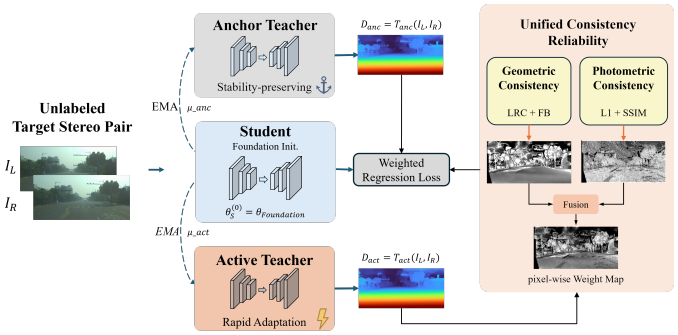

Fig. 2. Overview of the proposed Asymmetric Dual-Teacher self-training framework. A slow-updated Anchor Teacher provides stable pseudo-disparity targets, while a fast-updated Active Teacher estimates pixel-wise reliability via the Unified Consistency Reliability (UCR) module. The student foundation model is optimized using a reliability-weighted regression loss, decoupling stable supervision from adaptive reliability estimation.

is to leverage these unlabeled target data to further improve the generalization of the pre-trained foundation model via self-training.

A fundamental challenge stems from the regression nature of stereo matching: pseudo-disparity supervision must be both sufficiently stable to preserve pre-trained geometric priors, yet rigorous enough to filter out erroneous labels induced by domain shifts. Balancing these two requirements is critical for effective self-training-based domain adaptation.

B. Asymmetric Dual-Teacher Framework

Conventional self-training typically uses a single teacher to provide pseudo-disparity targets and to estimate their reliability, making pseudo-supervision prone to error reinforcement under domain shift. To mitigate this limitation, we propose an Asymmetric Dual-Teacher (ADT) framework that decouples target generation and reliability estimation (Fig. 2).

For each unlabeled target stereo pair (I_L, I_R), we maintain three networks with identical architectures: a Student model S, an Anchor Teacher T_{anc}, and an Active Teacher T_{act}. The Anchor Teacher provides pseudo-disparity targets D_{anc} for regression. It evolves slowly and serves as a stability anchor, preserving the geometric priors of the foundation model and preventing target drift during adaptation. In contrast, the Active Teacher is designed to be highly adaptive. It does not supervise the student directly via disparity regression. Instead, it estimates the pixel-wise reliability of pseudo-supervision by capturing domain-specific error patterns. Its predictions, D_{act}, are used exclusively to derive a soft confidence map, W_{ucr}, via the proposed Unified Consistency Reliability (UCR) module.

Asymmetric Update Dynamics: Let θ_S, θ_{anc}, and θ_{act} denote the parameters of the student, anchor teacher, and active teacher, respectively. At iteration t, the student is updated via gradient descent, while the two teachers are updated using asymmetric exponential moving averages (EMA):

$$\theta_{act}^{(t)} \leftarrow \mu_{act}\theta_{act}^{(t-1)} + (1-\mu_{act})\theta_S^{(t)} \tag{1}$$

$$\theta_{anc}^{(t)} \leftarrow \mu_{anc}\theta_{anc}^{(t-1)} + (1-\mu_{anc})\theta_S^{(t)} \tag{2}$$

where $\mu_{act} < \mu_{anc}$. In practice, the anchor teacher is updated every K steps (equivalently every 2 epochs in our default configuration), while the active teacher is updated every step. This asymmetric update strategy ensures that θ_{anc} maintains long-term stability, while θ_{act} remains responsive to domain shifts.

C. Unified Consistency Reliability

Because stereo self-training is sensitive to label errors, we build the UCR module to conservatively filter pseudo-labels. We observe that erroneous pseudo-labels are more detrimental to stereo matching performance than a complete absence of supervision. Consequently, our reliability estimation is designed to be highly conservative, ensuring that only the most trustworthy signals guide the student. UCR combines complementary geometric and photometric checks to compute a soft confidence map $W_{ucr}(x, y)$ for each pixel.

1) Geometric Consistency: Geometric consistency provides a strong cue for detecting occlusions and mismatches, but left-right consistency alone can be over-optimistic under repetitive textures. We therefore combine left-right consistency (LRC) with a forward–backward (FB) check to obtain a stricter geometric reliability.

LRC: For a left pixel (x_l, y) in the left image, the Active Teacher predicts $D_{act}^L(x_l, y)$ and its correspondence in the right view is $x_r' = x_l - D_{act}^L(x_l, y)$. We compute the left-right discrepancy:$e_{lrc}(x_l, y) = |D_{act}^L(x_l, y) - D_{act}^R(x_r', y)|$. This error is converted into a soft confidence:

$$C_{lrc}(x_l, y) = \exp(-\lambda_{lrc} \cdot e_{lrc}(x_l, y)) \tag{3}$$

where $\lambda_{\text{lrc}} > 0$ controls the sharpness of the decay.

FB Check: We start from a right pixel (x_r, y), and map it to the left coordinate (x_l', y) via $x_l' = x_r + D_{act}^R(x_r, y)$. The FB error is:

$$e_{fb}(x_r, y) = \left|D_{act}^R(x_r, y) - D_{act}^L(x_l', y)\right| \tag{4}$$

We compute the right-view confidence:

$$C_{fb}^R(x_r, y) = \exp\big(-\lambda_{fb} \cdot e_{fb}(x_r, y)\big) \tag{5}$$

where $\lambda_{\text{fb}} > 0$. And then warp it to the left view using D_{act}^L to align with the student's coordinate system, yielding $C_{fb}(x_l, y)$.

The final geometric reliability is the fusion of these components:

$$W_{geom}(x_l, y) = C_{lrc}(x_l, y) \cdot C_{fb}(x_l, y) \tag{6}$$

This ensures that high confidence is assigned only when a match is bi-directionally consistent and structurally valid.

2) Photometric Consistency Check: Geometric consistency alone can be over-optimistic, especially in ill-posed regions such as textureless areas where predictions may appear self-consistent. To provide an additional external cue, we incorporate an additional photometric check by warping the right image to the left view using D_{act}^L. The error combines L1 distance and SSIM, and is converted into a soft confidence score via an exponential decay.

Concretely, we reconstruct the left view $\hat{I}_L$ by warping the right image I_R with the disparity D_{act}^L predicted by the Active Teacher. The photometric error is defined as:

$$E_{photo} = \alpha \frac{1 - \text{SSIM}\left(I_L, \hat{I}_L\right)}{2} + (1-\alpha) \left\| I_L - \hat{I}_L \right\|_1 \quad (7)$$

where α balances structural similarity and pixel-wise intensity differences. This error is then mapped to a soft photometric confidence via an exponential decay: $W_{photo} = \exp(-\lambda_p \cdot E_{photo})$.

3) Unified Fusion: The final reliability map W_{ucr} is obtained by multiplicatively fusing geometric and photometric confidence scores:

$$W_{ucr}(x, y) = W_{geom}(x, y) \cdot W_{photo}(x, y) \quad (8)$$

This fusion ensures that a pixel contributes to supervision only when it is both geometrically valid and photometrically consistent, effectively suppressing noise while preserving informative hard samples.

D. Loss Formulation

The core innovation of our method lies in the functional decoupling of the supervision signals. Specifically, the regression target is provided by the stable Anchor Teacher, while the reliability weighting is estimated by the adaptive Active Teacher. We define the reliability-weighted pseudo-label regression loss as:

$$\mathcal{L}_{\text{pseudo}} = \frac{\sum_{(x,y)} W_{ucr}(x, y)\ \ell_1(D_S(x, y),\ D_{anc}(x, y))}{\sum_{(x,y)} W_{ucr}(x, y) + \epsilon} \quad (9)$$

Using D_{anc} as the target prevents target drift and preserves foundation priors, while W_{ucr} selectively attenuates unreliable regions identified by the active teacher.

IV. Experiments

A. Experimental Setup

Datasets: We evaluate on five public benchmarks covering both stereo scenes and challenging real-world corruptions: KITTI 2015 [23], Middlebury [24], ETH3D [25], Booster [26], and DrivingStereo Weather split [27] (Cloudy/Rainy/Foggy). We construct non-overlapping training and test splits for each dataset. Model adaptation is performed exclusively on the training split, and evaluation is conducted on the held-out test split.

Implementation Details: Our method, implemented in PyTorch and trained on 4 NVIDIA A5000 GPUs with a total batch size of 4, adapts models using only unlabeled target stereo pairs, without ground-truth disparities or source data. Three identical networks per backbone are initialized from official pretrained checkpoints. Optimization uses AdamW (lr $= 5 \times 10^{-4}$, weight decay $= 1 \times 10^{-4}$) with cosine annealing over up to 200 epochs, with early stopping if mean UCR confidence stagnates for 10 epochs. Key parameters are $\mu_{act} = 0.999$, $\mu_{anc} = 0.9999$, with T_{act} and T_{anc} updated per step and every 2 epochs, respectively. Images are cropped to 320×736. UCR loss weights are $\lambda_{lrc} = 2.0$, $\lambda_{fb} = 2.0$, and $\lambda_{photo} = 10.0$.

Evaluation Metrics: We report the End-point Error (EPE, i.e., the mean absolute disparity error), D1 (percentage of pixels with disparity error > 3 px and $> 5\%$), and BP-X (percentage of pixels with disparity error $> X$ px).

B. Comparisons with State-of-the-art

We compare our method against strong zero-shot stereo models and recent adaptation baselines. For all zero-shot baselines, we report results from official pretrained checkpoints without using any target data. To further evaluate the architectural universality of our framework, we instantiate our ADT strategy on three distinct backbones: two standard iterative models (ADT-IGEV [8], ADT-Selective-IGEV [28]), and a large-scale foundation model (ADT-FoundationStereo [1]). Table I summarizes results on Middlebury, KITTI 2015, ETH3D, and Booster. Our ADT-instantiated foundation model achieves consistent gains over the corresponding zero-shot foundation baseline across all settings. The improvements are especially pronounced on Booster, where reflective/transparent regions tend to produce overconfident but incorrect disparities; by training against stable pseudo targets while downweighting unreliable pixels, our method reduces the impact of noisy supervision and yields more accurate disparity estimates.

Robustness under Adverse Weather: Table II reports results on the DrivingStereo Weather subsets. Weather corruptions (fog glare, rain streaks, and low contrast) strongly degrade pseudo-label reliability, often causing self-training to deteriorate. Our method improves performance by explicitly separating pseudo-label generation from reliability estimation: the Anchor Teacher provides temporally stable regression targets, while the Active Teacher, through UCR, produces a conservative reliability map that suppresses weather-induced false matches. Concretely, on the Foggy subset, our approach reduces BP-3 from 2.02 to 1.12 and EPE from 0.90 to 0.75 compared to the zero-shot FoundationStereo baseline (Table II), indicating substantially fewer large-error pixels under severe degradation.

As illustrated in Fig. 3, we present qualitative comparisons on the DrivingStereo benchmark under adverse weather conditions. Our method suppresses weather-induced disparity artifacts and better preserves structural continuity in degraded regions.

C. Ablation Study

We conduct ablations on the DrivingStereo-Foggy subset (Table III) and make the following key insights.

Single Teacher Baselines: Standard single-teacher self-training with fast (μ=0.999) or slow (μ=0.9999) EMA teachers yields only marginal gains over the zero-shot baseline, showing that unfiltered pseudo-labels are unreliable under noise. Adding UCR improves performance but saturates quickly, indicating that reliability estimation alone is insufficient when targets and confidence are coupled, as consistent yet incorrect predictions, e.g., in foggy conditions, cannot be effectively corrected.

Asymmetric Dual-Teacher Decoupling: Decoupling target generation and reliability estimation enables significant improvement. Our dual-teacher design, where the Anchor

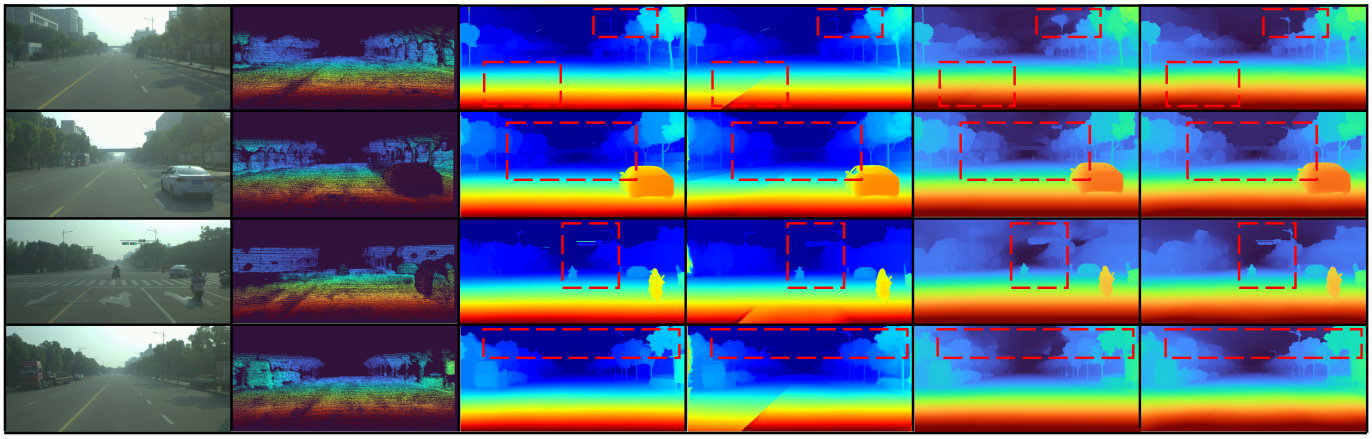

Fig. 3. Qualitative comparisons of state-of-the-art methods under adverse weather conditions on DrivingStereo dataset.

TABLE I
QUANTITATIVE COMPARISONS ON STANDARD STEREO BENCHMARKS. ALL METRICS ARE SMALLER THE BETTER.

Category	Method	Middlebury			KITTI 2015			ETH3D			Booster		
		D1	EPE	BP-2	D1	EPE	BP-2	D1	EPE	BP-2	D1	EPE	BP-2
Zero-shot	PSMNet (CVPR'18)	25.83	8.94	32.28	32.73	5.55	44.18	9.2	8.07	10.94	32.60	10.67	54.19
	AANet (CVPR'20)	17.79	4.62	26.88	8.86	1.54	14.85	3.99	0.89	7.44	29.28	7.74	53.12
	FoundationStereo (CVPR'25)	1.44	0.52	2.30	2.69	0.82	5.04	0.30	0.16	0.38	4.51	1.77	5.67
	Stereo Anywhere (CVPR'25)	3.92	1.11	7.12	3.12	0.95	6.76	0.76	0.33	1.31	5.13	1.78	9.94
	SMoE (ICCV'25)	6.22	1.60	9.96	4.69	1.04	7.92	0.58	0.27	2.44	10.86	3.15	20.40
	IGEV (CVPR'23)	8.29	1.81	18.63	5.07	1.12	8.44	1.49	0.40	2.03	12.01	4.69	21.70
	Selective-IGEV (CVPR'24)	5.45	1.22	8.75	5.19	1.33	7.54	4.01	0.33	3.36	9.67	3.04	13.73
	LightStereo (ICRA'25)	18.00	5.47	27.30	14.21	2.33	21.75	5.98	2.68	7.68	13.29	4.11	24.56
Unsupervised Adaptation	CST-RAFT-Stereo (CVPR'25)	10.59	1.39	-	4.93	1.03	-	2.35	0.25	-	-	-	-
	CST-Selective-IGEV (CVPR'25)	5.37	0.77	-	4.91	1.05	-	3.54	0.27	-	-	-	-
	ADT-IGEV (Ours)	7.06	1.48	10.40	4.28	1.03	7.40	1.11	0.28	1.54	9.26	2.82	17.82
	ADT-Selective-IGEV (Ours)	5.26	1.12	7.95	4.62	1.11	6.77	0.92	0.26	1.33	7.62	2.65	9.86
	ADT-FoundationStereo (Ours)	**1.32**	**0.51**	**2.18**	**2.37**	**0.79**	**4.54**	**0.27**	**0.15**	**0.35**	**4.30**	**1.49**	**5.44**

TABLE II
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE DRIVINGSTEREO DATASET UNDER ADVERSARIAL WEATHER CONDITIONS.

Method	Cloudy		Rainy		Foggy	
	EPE	BP-3	EPE	BP-3	EPE	BP-3
PSMNet(CVPR'18)	4.54	29.11	18.84	49.60	17.51	58.68
AANet (CVPR'20)	1.54	9.44	2.64	21.85	2.21	18.05
FoundationStereo (CVPR'25)	0.88	2.47	1.61	12.11	0.90	2.02
Stereo Anywhere (CVPR'25)	0.91	2.85	1.55	10.80	0.91	2.49
SMoE (ICCV'25)	0.88	2.43	1.33	**6.40**	1.21	6.77
Selective-IGEV (CVPR'24)	1.07	4.34	2.00	12.13	1.10	3.05
LightStereo (ICRA'25)	2.19	19.79	4.12	33.77	3.60	37.80
IGEV (CVPR'23)	0.97	3.14	1.59	9.25	1.26	7.72
ADT-IGEV	0.81	1.48	1.34	7.69	0.95	3.44
ADT-Selective-IGEV	0.80	2.06	1.40	7.59	0.81	1.17
ADT-FoundationStereo	**0.76**	**1.48**	**1.28**	7.23	**0.75**	**1.12**

Teacher provides stable pseudo-labels and the Active Teacher computes UCR-based confidence, achieves D1 = 1.10 and EPE = 0.75, outperforming the best single-teacher + UCR variant (D1 = 1.66, EPE = 0.83). This confirms that separating supervision and confidence enhances pseudo-label quality.

Effectiveness UCR & Soft Confidence: Replacing UCR with geometric-only LRC degrades performance, highlighting the importance of photometric cues for rejecting spurious matches in fog. Hard masking (0/1) also underperforms soft weighting, as it discards informative pixels and leads to sparse supervision; soft weights preserve gradients in ambiguous regions.

Role Switching: Swapping teacher roles (Anchor generates confidence, Active provides targets) results in a drop to D1 = 1.54, demonstrating that performance depends on functional specialization. The reliability estimator must remain adaptive to domain shifts, while the target generator must stay stable. Misaligned roles impair adaptation and optimization dynamics.

D. Training Efficiency Analysis

As shown in Table IV, maintaining asymmetric dual teachers incurs additional GPU memory usage and training time compared to the single-teacher baseline. However, we argue that this training-time investment is highly cost-effective, yielding a significant 39.1% error reduction on the challenging Foggy split. Importantly, the additional cost is confined to the training phase; during inference, the final student model retains the same computational footprint as the baseline, incurring no extra latency at deployment.

TABLE III
ABLATION STUDY ON THE DRIVINGSTEREO-FOGGY DATASET.

Method	Mask	Conf.	DrivingStereo - Foggy			
			D1↓	EPE↓	BP-2↓	BP-3↓
Zero-shot Baseline	–	–	2.01	0.90	7.15	2.02
Single-T (μ_{fast})	–	–	1.90	0.88	7.04	1.92
Single-T (μ_{slow})	–	–	1.84	0.88	6.70	1.86
Single-T (μ_{fast})	Soft	UCR	1.79	0.88	6.84	1.81
Single-T (μ_{slow})	Soft	UCR	1.66	0.83	5.86	1.67
Dual-Teacher (Ours)	**Soft**	**UCR**	**1.10**	**0.75**	**3.57**	**1.12**
Dual-Teacher (LRC-only)	Soft	LRC	1.53	0.79	4.86	1.54
Dual-Teacher (Hard Mask)	0/1	UCR	1.37	0.83	5.10	1.40
Dual-Teacher (Role Reversal)	Soft	UCR	1.54	0.85	6.09	1.56

TABLE IV
COST VS. PERFORMANCE ANALYSIS ON THE DRIVINGSTEREO-FOGGY SUBSET.

Method	Training Cost		Inference	BP-2 ↓
	Peak Mem (GB)	Time / Iter (s)	Latency (ms)	
Single-Teacher	21.18	2.39	74.1	5.86
Dual-Teacher (Ours)	21.31	3.02	74.2	3.57
Relative Change	*+0.6%*	*+26.3%*	*+0.1%*	*-39.1%*

E. Sensitivity to EMA Dynamics

Tab. V analyzes the EMA decay rates with the lens of our *target–reliability decoupling*. The strongest results consistently occur when the Anchor Teacher uses a larger decay (more conservative) than the Active Teacher, matching the intended design: stable pseudo targets paired with adaptive reliability estimation. Extremely fast active updates produce noisy confidence, while overly slow active updates yield stale reliability under domain shift.

TABLE V
SENSITIVITY ANALYSIS OF EMA DECAY RATES ON DRIVINGSTEREO-FOGGY.

Anchor Teacher (μ_{anc})	Metric	Active Teacher (μ_{act})			
		0.9	0.99	0.999	0.9999
0.99	EPE	1.01	0.90	0.86	0.93
	D1	2.55	2.02	1.76	2.12
0.999	EPE	0.94	0.81	0.82	0.84
	D1	2.15	1.50	1.38	1.58
0.9999	EPE	0.89	0.78	**0.75**	0.80
	D1	1.92	1.29	**1.10**	1.32

V. CONCLUSION

This work studied the generalization capability of pre-trained stereo matching foundation models in downstream tasks, particularly under scenarios where only unlabeled target-domain data is available. To address the challenges posed by domain shifts and unreliable pseudo-labels during adaptation, we propose an Asymmetric dual-teacher (ADT) framework that leverages complementary supervisory signals from stable geometric priors and adaptive reliability cues. This design effectively balances prior knowledge preservation with target-domain adaptability, enabling stable and reliable model adaptation. Extensive experiments across a wide range of stereo benchmarks validate the effectiveness of our approach, demonstrating consistent and significant improvements over existing methods, especially in challenging conditions.

REFERENCES

[1] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," in *CVPR*, 2025.

[2] L. Bartolomei, F. Tosi, M. Poggi, and S. Mattoccia, "Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail," in *CVPR*, 2025, pp. 1013–1027.

[3] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *ICCV*, 2017.

[4] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *CVPR*, 2018.

[5] Y. Wang, J. Zheng, C. Zhang, Z. Zhang, K. Li, Y. Zhang, and J. Hu, "Dualnet: Robust self-supervised stereo matching with pseudo-label supervision," in *AAAI*, 2025.

[6] Y. Wang, J. Hu, J. Hou, C. Zhang, R. Yang, and D. O. Wu, "Rose: Robust self-supervised stereo matching under adverse weather conditions," *TCSVT*, 2025.

[7] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *ICLR*, 2019.

[8] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *CVPR*, 2023.

[9] X. Yue, Z. Lu, X. Lin, W. Ren, Z. Shao, H. Hu, Y. Zhang, and Q. Liao, "Semi-stereo: A universal stereo matching framework for imperfect data via semi-supervised learning," in *CVPR*, 2024.

[10] J. Zhou, P. Ye, H. Zhang, J. Yuan, R. Qiang, L. YangChenXu, W. Cailin, F. Xu, and T. Chen, "Consistency-aware self-training for iterative-based stereo matching," in *CVPR*, 2025.

[11] A. Tonioni, M. Poggi, S. Mattoccia, and L. Di Stefano, "Unsupervised adaptation for deep stereo," in *ICCV*, 2017.

[12] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "On the uncertainty of self-supervised monocular depth estimation," in *CVPR*, 2020.

[13] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017.

[14] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *CVPR*, 2019.

[15] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *International Conference on 3D Vision*, 2021.

[16] R. Chabra, J. Straub, C. Sweeney, R. Newcombe, and H. Fuchs, "Stereodrnet: Dilated residual stereonet," in *CVPR*, 2019.

[17] F. Zhang, X. Qi, R. Yang, V. Prisacariu, B. Wah, and P. Torr, "Domain-invariant stereo matching networks," in *ECCV*, 2020.

[18] X. Song, G. Yang, X. Zhu, H. Zhou, Z. Wang, and J. Shi, "Adastereo: A simple and efficient approach for adaptive stereo matching," in *CVPR*, 2021.

[19] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. W. Lau, "Dual student: Breaking the limits of the teacher in semi-supervised learning," in *ICCV*, 2019.

[20] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *IJCNN*, 2020.

[21] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *NeuRIPS*, 2018.

[22] X. Huo, L. Xie, W. Zhou, H. Li, and Q. Tian, "Focus on your target: A dual teacher-student framework for domain-adaptive semantic segmentation," in *ICCV*, 2023.

[23] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *CVPR*, 2015.

[24] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German Conference on Pattern Recognition*, 2014.

[25] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *CVPR*, 2017.

[26] P. Z. Ramirez, F. Tosi, M. Poggi, S. Salti, S. Mattoccia, and L. Di Stefano, "Open challenges in deep stereo: the booster dataset," in *CVPR*, 2022.

[27] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, "Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios," in *CVPR*, 2019.

[28] X. Wang, G. Xu, H. Jia, and X. Yang, "Selective-stereo: Adaptive frequency information selection for stereo matching," in *CVPR*, 2024.