Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MAD: Mask-Guided and Dual-Model Self-Training for Unsupervised Domain Adaptation in Stereo Matching

Jie Tang
South China University of Technology
Gaofeng Peng
2534490495@qq.com

South China University of Technology
Jialu Liu
Shenzhen Institute of Artificial Intelligence and Robotics for Society
Shaoshan Liu
Shenzhen Institute of Artificial Intelligence and Robotics for Society

Research Article

Keywords: Stereo matching, Unsupervised domain adaptation, Pseudo-label self-training, Consistency verification, Dual-model synergy

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11190306/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MAD: Mask-Guided and Dual-Model Self-Training for Unsupervised Domain Adaptation in Stereo Matching

Jie Tang[1], Gaofeng Peng[1*], Jialu Liu[2], Shaoshan Liu[2]

[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou, 510006, Guangdong, China.
[2]Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, 518129, Guangdong, China.

*Corresponding author(s). E-mail(s): 2534490495@qq.com;
Contributing authors: cstangjie@scut.edu.cn; liujialu@cuhk.edu.cn; shaoshanliu@cuhk.edu.cn;

Abstract

Stereo matching models trained on synthetic data often degrade when transferred to real scenes due to the domain gap. Unsupervised domain adaptation provides a practical way to exploit unlabeled target-domain stereo images without requiring additional disparity annotations. However, existing adaptation approaches often introduce architecture-specific components or additional learnable modules, which can limit their applicability across different stereo matching networks. To address this issue, we present MAD, a simple, architecture-agnostic adaptation framework that combines mask-guided self-supervised fine-tuning with dual-model pseudo-label self-training. In the first stage, left-right consistency of the source-domain pretrained model is used to identify reliable non-occluded regions and reduce the influence of unreliable photometric supervision during target-domain fine-tuning. In the second stage, the predictions of the pretrained and fine-tuned models are jointly considered, and geometric and photometric consistency checks are applied to select filtered pseudo-labels for further self-training. The framework does not require modifications to the internal architecture of the underlying stereo network or additional trainable adaptation modules. We evaluate MAD on multiple real-world datasets and across several stereo matching backbones, including PSMNet, GwcNet, ACVNet, CoEx, and CFNet. The results show consistent improvements over conventional unsupervised fine-tuning across a range of target domains and environmental conditions. Cross-domain experiments further indicate that the adapted models can retain useful transferability

across different target domains. Overall, combining simple geometric filtering with complementary predictions provides a practical, architecture-independent strategy for synthetic-to-real stereo adaptation.

Keywords: Stereo matching, Unsupervised domain adaptation, Pseudo-label self-training, Consistency verification, Dual-model synergy

1 Introduction

Stereo matching estimates scene depth by finding correspondences between the two views of a binocular image pair, and it is widely used in applications such as autonomous driving and robotic navigation. Recent deep stereo matching methods have achieved strong performance on standard benchmarks, but their accuracy typically depends on large-scale dense disparity annotations. A common strategy is therefore to pre-train stereo networks on synthetic datasets such as SceneFlow [1] and then transfer them to real-world scenes. However, differences in appearance, illumination, object distribution, and image characteristics introduce a substantial domain gap, which can lead to a marked performance drop when the source-trained model is directly deployed. Although fine-tuning with labeled real-world data can alleviate this problem, obtaining dense disparity annotations in real scenes is costly. In contrast, collecting unlabeled stereo image pairs is relatively practical, motivating the study of unsupervised domain adaptation (UDA) for stereo matching.

Existing UDA approaches for stereo matching adopt different mechanisms and consequently make different trade-offs among adaptation performance, training complexity, architectural dependence, and supervision reliability. Input-level methods, such as StereoGAN [2], reduce the appearance gap through image translation, but their optimization pipelines are relatively involved. Feature-level alignment methods, including DaStereo [3] and AdaStereo [4], align source and target representations using adversarial or statistical objectives, with the corresponding design choices introducing additional assumptions or components. Output-level approaches, such as TS-Nets [5] and UCFNet [6], exploit pseudo-labels or teacher–student mechanisms, but often depend on architecture-specific outputs or auxiliary modules. Self-supervised fine-tuning methods, exemplified by MADNet [7], directly optimize photometric and geometric consistency on target-domain images and are relatively simple to integrate, but the supervision can become unreliable in occluded regions.

These observations motivate a practical question: can existing geometric and image-consistency constraints be organized into an adaptation procedure that is largely independent of the internal architecture of the stereo network, while still providing consistent improvements over conventional self-supervised fine-tuning? Rather than introducing a new stereo matching architecture, we investigate how useful information already available from a source-domain pretrained model and a target-adapted model can be combined for this purpose.

We first perform a systematic analysis of pretrained and target-domain fine-tuned models on several stereo matching networks. Two observations motivate our design.

First, although the source-domain pretrained model can produce inaccurate disparities after domain transfer, its left-right consistency can still identify a high-precision subset of non-occluded pixels. This subset can be used as a spatial prior to reduce unreliable supervision during self-supervised fine-tuning. Second, the pretrained model and the target-domain fine-tuned model often exhibit complementary prediction characteristics: the former tends to preserve sharper local structures, while the latter better captures target-domain appearance and larger-scale structures. These complementary predictions can therefore be jointly filtered and used to construct pseudo-labels for further adaptation.

Based on these observations, we propose **MAD**, a two-stage unsupervised domain adaptation framework for stereo matching. The first stage performs mask-guided self-supervised fine-tuning using non-occlusion masks derived from the source model. The second stage jointly uses the source pretrained model and the stage-I adapted model to generate filtered pseudo-labels through geometric and photometric consistency checks, which are then used for supervised self-training of the original model. MAD operates on the output disparity maps and original input images, and does not require modifications to the internal architecture of the underlying stereo network or additional trainable adaptation modules.

The main contributions of this paper are summarized as follows:

1. We present MAD, a simple, architecture-agnostic unsupervised domain adaptation framework for stereo matching that operates on existing pretrained models without modifying their internal architectures or introducing additional trainable adaptation modules.
2. We introduce a mask-guided self-supervised fine-tuning strategy that exploits left-right consistency from the source-domain pretrained model to reduce unreliable supervision from occluded regions. We further develop a dual-model pseudo-label generation scheme that combines predictions from the pretrained and target-adapted models through geometric and photometric consistency verification.
3. We conduct experiments across multiple target datasets and stereo matching backbones to examine the effectiveness and transferability of the proposed adaptation strategy. The results show consistent improvements over conventional unsupervised fine-tuning and illustrate the practical applicability of the framework across different model architectures and target conditions.

2 Related Work

2.1 Cross-Domain Stereo Matching

Deep stereo matching models are commonly trained on synthetic datasets with dense disparity annotations. Although such models can achieve high accuracy in the source domain, their performance may decrease substantially in real-world scenes because of changes in appearance, illumination, sensor characteristics, and scene content. Two common directions for improving cross-domain performance are domain generalization and unsupervised domain adaptation. Domain generalization aims to improve robustness to unseen domains without using target-domain data, whereas UDA explicitly

exploits unlabeled target-domain images during adaptation. This work focuses on the latter setting.

2.2 Domain Generalization

Domain generalization methods seek to learn representations or matching mechanisms that transfer to unseen domains. DSMNet [8] reduces domain-specific statistics through feature whitening, while MS-Nets [9] constructs cost volumes from handcrafted matching features. More recent approaches introduce additional priors or constraints, including ImageNet [10] pretrained features in GraftNet [11], shortcut-feature suppression in ITSA [12], and geometric feature consistency in FC-Stereo [13]. These methods can improve transferability without accessing target-domain images, but they address a different setting from the target-domain adaptation considered in this paper.

2.3 Unsupervised Domain Adaptation

Unsupervised domain adaptation exploits unlabeled target-domain images to reduce the source-to-target performance gap. Existing methods can be broadly grouped into input-level alignment, feature-level alignment, output-level alignment, and self-supervised fine-tuning.

Input-level alignment methods, such as StereoGAN [2], translate source-domain images toward the target appearance. Feature-level alignment methods, including DaStereo [3] and AdaStereo [4], align representations across domains using adversarial learning or statistical matching. These approaches provide effective alternatives to direct fine-tuning, but their implementation typically involves additional training objectives or components.

Output-level alignment approaches mainly rely on self-training or teacher–student paradigms. Representative methods include TS-Nets [5], DKT-Nets [14], UCFNet [6], and NeRF-Supervised [15], which explore pseudo-label supervision, knowledge distillation, uncertainty estimation, or auxiliary scene reconstruction. These methods demonstrate the usefulness of target-domain pseudo-labels, while some of their mechanisms are tailored to particular model characteristics or require additional components.

Self-supervised fine-tuning provides a comparatively simple alternative by directly using target-domain image and geometric constraints. Methods such as MADNet [7] optimize reconstruction, smoothness, and left-right consistency losses without requiring target-domain disparity annotations. However, photometric supervision is less reliable in occluded and otherwise difficult regions, which can limit the benefit of direct fine-tuning.

2.4 Discussion

The above approaches demonstrate several effective ways to exploit unlabeled target-domain data, but they also involve different assumptions and implementation requirements. Motivated by this trade-off, we focus on a more practical adaptation setting: how to improve a pretrained stereo model using only its existing predictions

and the original target-domain image pairs, without introducing architecture-specific adaptation modules.

MAD addresses this setting with two complementary operations. First, left-right consistency from the source model is used to identify a high-precision subset of non-occluded regions and to reduce unreliable supervision during self-supervised fine-tuning. Second, predictions from the source and adapted models are jointly filtered using geometric and photometric consistency checks, providing pseudo-labels for further self-training. The resulting framework reuses standard stereo constraints and can be applied to several existing stereo matching networks without changing their internal architectures.

3 Method

3.1 Problem Analysis and Design Motivation

Given a stereo matching model pre-trained on a source domain and a set of unlabeled target-domain stereo image pairs, the goal of unsupervised domain adaptation is to obtain a model that performs well on the target domain. To better understand how source-domain priors can facilitate adaptation, we conduct a series of visualization and quantitative analyses on several representative stereo matching networks, including PSMNet [16], CFNet [17], and ACVNet [18]. Specifically, we compare the predictions of their source pre-trained models and their target-domain fine-tuned models on KITTI and CityScapes from two quantitative aspects, namely left-right consistency and disparity error distribution, complemented by a qualitative comparison of the predicted disparity maps. Based on these analyses, we obtain two key observations that directly motivate the proposed framework.

Observation 1: Source-model predictions provide high-precision non-occlusion masks. To investigate whether the source pre-trained model can provide reliable supervision for target-domain adaptation, we use a PSMNet model pre-trained on SceneFlow to estimate disparity on the target domain and generate non-occlusion masks through left-right consistency checking under different consistency thresholds. We then compare the predicted masks with the ground-truth non-occluded regions.

The quantitative results are reported in Table 1. As the consistency threshold becomes stricter, the precision remains consistently high across all settings. Even when the source model exhibits relatively poor disparity accuracy (e.g., D1-all reaches 19.1%), the precision of the predicted non-occluded regions still remains above 95%. Fig. 1 further confirms this observation.

These results indicate that, despite imperfect disparity predictions on the target domain, the source model can still accurately identify reliable non-occluded regions through left-right consistency checking. Therefore, the resulting masks can serve as useful spatial priors to weight the self-supervised loss, reducing unreliable supervision caused by occluded regions and improving the reliability of target-domain fine-tuning.

Observation 2: Source and adapted models exhibit complementary prediction characteristics. To further investigate the effect of target-domain adaptation, we compare the disparity maps produced by several source pre-trained models

Table 1 Precision of non-occlusion masks generated by left-right consistency checking under different disparity error levels and mask thresholds.

Sample	Pre-trained Model Error		Mask Threshold (Left-Right Disparity Difference)			
	D1-all (%)	EPE (px)	1px	2px	3px	1%
Image 1	1.41	0.48	0.989	0.987	0.985	0.995
Image 2	1.38	0.51	0.988	0.986	0.985	0.994
Image 3	18.2	9.50	0.971	0.961	0.956	0.969
Image 4	19.1	5.91	0.974	0.967	0.964	0.969

Note: "1px" means disparity difference < 1 px; "1%" means disparity difference $< 1\%$ of the predicted disparity.

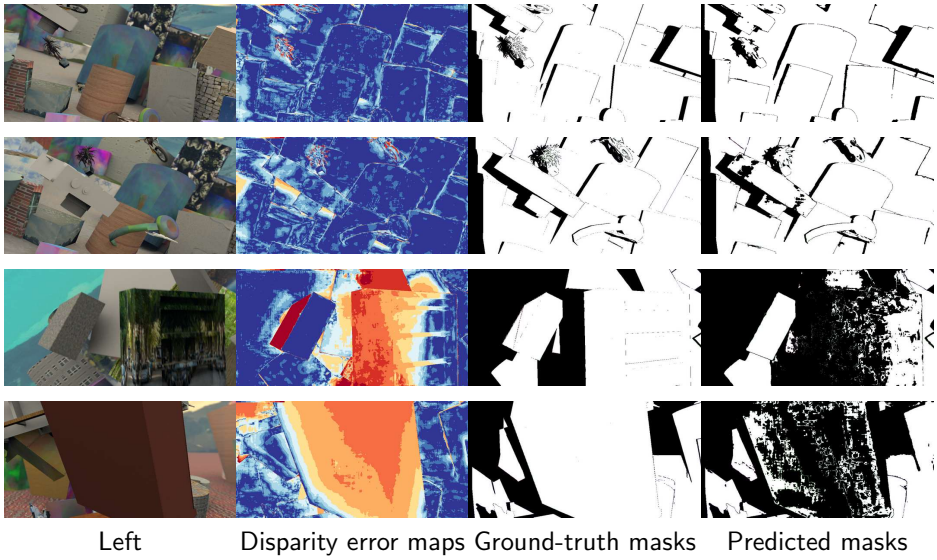


Fig. 1 Disparity prediction error maps and mask prediction visualizations.

and their corresponding models after self-supervised fine-tuning on the target domain. Fig. 2 presents representative examples.

Qualitatively, as illustrated in Fig. 2, the source pre-trained models generally preserve sharper high-frequency structures, such as object boundaries and fine geometric details, but often produce mismatches in large homogeneous regions. In contrast, the fine-tuned models better capture the overall appearance and low-frequency characteristics of the target domain, yielding smoother and more globally consistent disparity maps. However, this adaptation is often accompanied by the loss of high-frequency details, leading to blurred object boundaries and over-smoothed predictions.

These observations reveal complementary prediction characteristics between the two models. The source pre-trained model provides more accurate high-frequency structural information, whereas the adapted model offers better low-frequency consistency with the target domain. Consequently, combining their predictions can provide complementary supervision that is not available from either model alone.

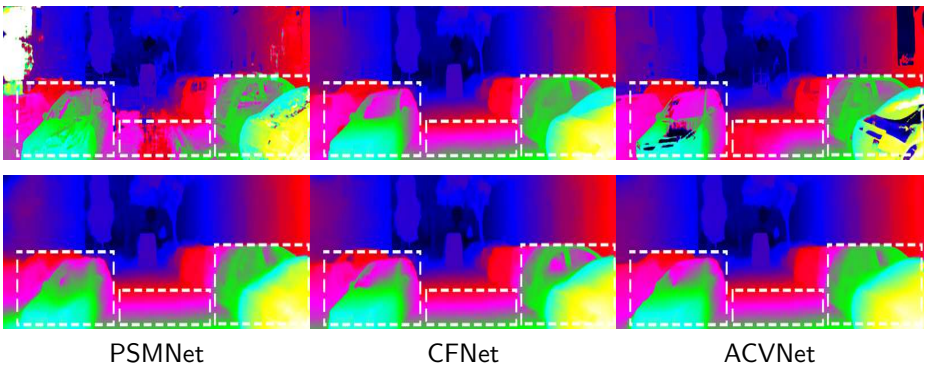


Fig. 2 Disparity maps from pre-trained models (top) and unsupervised fine-tuned models (bottom) for different methods.

3.2 Overview of MAD Framework

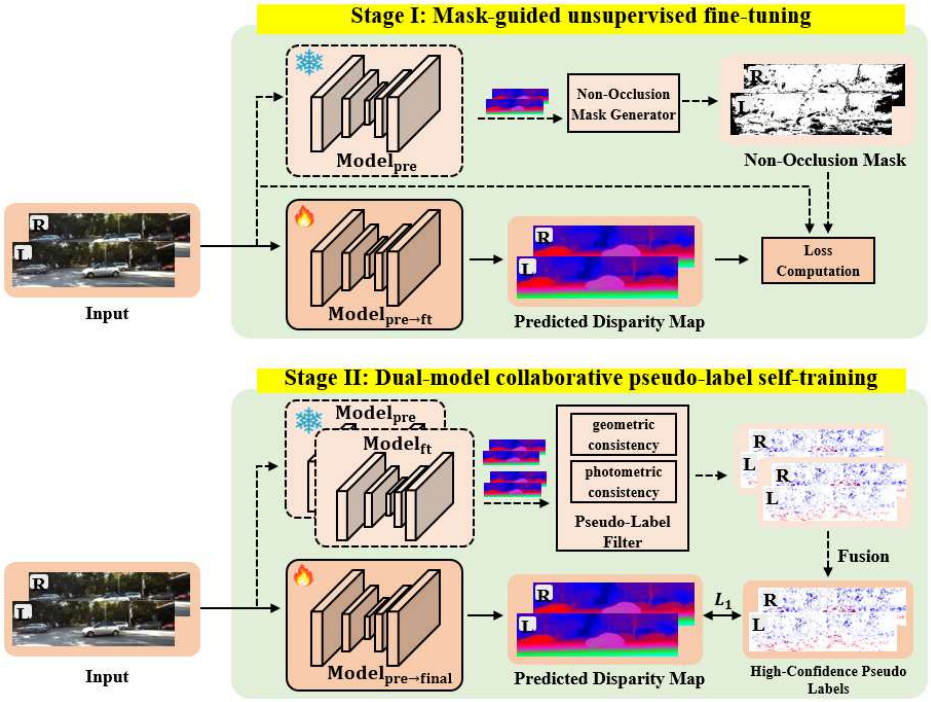


Fig. 3 Overall framework of MAD.

Our framework requires no modifications to the internal structure of any stereo matching network; it enables architecture-independent domain adaptation for existing pretrained models using their output disparity maps and the original input images.

The overall pipeline consists of two stages: mask-guided unsupervised fine-tuning and dual-model collaborative pseudo-label self-training, as illustrated in Fig. 3.

Stage I: Mask-guided unsupervised fine-tuning. The source pre-trained model $Model_{pre}$ predicts left and right disparity maps for the target-domain image pairs, and left-right consistency checking is applied to generate high-confidence non-occlusion region masks. These masks are then used to weight the photometric reconstruction loss and left-right consistency loss, enabling unsupervised fine-tuning on the target data to obtain a preliminary adapted model $Model_{ft}$.

Stage II: Dual-model collaborative pseudo-label self-training. Both $Model_{pre}$ and $Model_{ft}$ are used to predict disparities on the target images. A multi-level verification mechanism that combines geometric consistency and photometric consistency is designed to generate filtered masks for each model, selecting high-confidence disparity values that satisfy multiple constraints simultaneously as pseudo-labels. Finally, these pseudo-labels are used for supervised self-training of the original $Model_{pre}$, yielding the final model $Model_{final}$.

3.3 Stage I: Mask-Guided Unsupervised Fine-Tuning

Traditional unsupervised fine-tuning for stereo matching [19] usually employs a combination of reconstruction loss, smoothness loss, and left-right consistency loss. However, due to the absence of ground-truth disparity on the target domain, these loss functions produce unreliable gradient signals in occluded regions, which directly limits the fine-tuning effectiveness. To address this, we propose to leverage the prior information from the source pre-trained model by generating high-precision non-occlusion masks via left-right consistency checking, and then use these masks to weight the unsupervised losses, guiding the model to focus on trustworthy regions.

3.3.1 Non-Occlusion Mask Generation

Let $Model_{pre}$ denote the stereo matching model pre-trained on a source domain (e.g., SceneFlow). For any pair of binocular images (I^l, I^r) in the target domain, we first use $Model_{pre}$ to compute the left disparity map D^l and the right disparity map D^r. The reconstructed left disparity map $D^{r\to l}$ is obtained via disparity mapping as shown in (1).

$$D^{r\to l}(x,y) = D^r\left(x - D^l(x,y), y\right) \tag{1}$$

In an ideal non-occluded scenario with perfect disparity estimation, we should have $D^{r\to l}(x,y) = D^l(x,y)$. For occluded points, since the target pixel is not visible in the right view, the reconstructed disparity value deviates significantly, resulting in a large $|D^{r\to l} - D^l|$. Therefore, we can preliminarily filter non-occluded pixels by setting a threshold τ to generate an initial non-occlusion mask as shown in (2).

$$M^{l,\mathrm{noc}} = \mathbb{I}\left(\left|D^{r\to l} - D^l\right| < \tau\right) \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Notably, although $Model_{pre}$ exhibits prediction errors on the target domain, our experiments show that the resulting masks still achieve high precision — the proportion of pixels predicted as non-occluded that

are truly non-occluded remains above 95% in Table 1. While the masks are somewhat sparse, they serve as a reliable subset of non-occluded regions for weighting the subsequent unsupervised losses.

3.3.2 Mask-Weighted Loss Function

The overall loss for unsupervised fine-tuning consists of three components: reconstruction loss L_{rec}, smoothness loss L_{smooth}, and left-right consistency loss L_{cons}. To focus on reliable regions, we introduce multiple masks into the reconstruction and consistency losses.

Reconstruction loss enforces photometric consistency between the input image and the warped reconstructed image. Taking the left view as an example, as shown in (3).

$$L_{rec}^{l} = \frac{1}{\sum_{i,j} M_{ij}^{l,rec}} \sum_{i,j} M_{ij}^{l,rec} \left[\alpha_{ssim} \frac{1 - \mathrm{SSIM}(I_{ij}^{l}, \tilde{I}_{ij}^{l})}{2} + (1 - \alpha_{ssim}) \| I_{ij}^{l} - \tilde{I}_{ij}^{l} \|_1 \right] \tag{3}$$

where $\tilde{I}^{l}$ is the left view reconstructed from the right view; $\mathrm{SSIM}(\cdot)$ is the structural similarity index; and α_{ssim} is a balancing coefficient. The mask M^{rec} used in the reconstruction loss is obtained by element-wise multiplication of three masks as shown in (4).

$$M^{rec} = \neg M^{clip} \odot M^{noc} \odot M^{hard} \tag{4}$$

where M^{clip} is a boundary clipping mask to exclude invalid pixels at image edges; M^{noc} is the non-occlusion mask generated by (2), used to exclude occluded regions; and M^{hard} is a hard example mining mask, as shown in (5).

$$M_{ij}^{l,hard} = \mathbb{I}\left(\frac{1 - \mathrm{SSIM}(I_{ij}^{l}, \tilde{I}_{ij}^{l})}{2} > \epsilon \right) \cdot \mathbb{I}\left(\| I_{ij}^{l} - \tilde{I}_{ij}^{l} \|_1 > \epsilon \right) \tag{5}$$

This mask takes value 1 only when both the SSIM loss and L1 loss exceed the threshold ϵ, guiding the model to focus on pixels that are currently poorly reconstructed.

Left-right consistency loss enforces geometric consistency between the left and right disparity maps, as shown in (6).

$$L_{cons}^{l} = \frac{1}{\sum_{i,j} M_{ij}^{l,cons}} \sum_{i,j} M_{ij}^{l,cons} \left| d_{ij}^{l} - d_{i,j+d_{ij}^{l}}^{r} \right| \tag{6}$$

where M^{cons} is defined as (7).

$$M^{cons} = \neg M^{clip} \odot M^{noc} \tag{7}$$

Only boundaries and occluded regions are excluded; hard example mining is not introduced to maintain stability of the geometric constraint.

Smoothness loss encourages local smoothness in the disparity map while preserving true depth discontinuities at image edges as shown in (8).

$$L^l_{smooth} = \frac{1}{N}\sum_{i,j}\left(|\partial_x d^l_{ij}|e^{-|\partial_x I^l_{ij}|} + |\partial_y d^l_{ij}|e^{-|\partial_y I^l_{ij}|}\right) \tag{8}$$

where N denotes the number of pixels over which the loss is averaged. This loss does not involve cross-view reconstruction, thus no mask is needed.

The total Stage-I loss is the weighted sum of the three terms as shown in (9).

$$L_{\text{stage1}} = \lambda_{rec}(L^l_{rec} + L^r_{rec}) + \lambda_{cons}(L^l_{cons} + L^r_{cons}) + \lambda_{smooth}(L^l_{smooth} + L^r_{smooth}) \tag{9}$$

Through the above mask-weighted strategy, the model can be stably fine-tuned on the target domain, yielding a preliminary adapted model $Model_{ft}$, which lays the foundation for pseudo-label generation in Stage II.

3.4 Stage II: Dual-Model Collaborative Pseudo-Label Self-Training

Based on Stage I, we obtain a preliminary adapted model $Model_{ft}$ on the target domain. To further improve performance, Stage II proposes a dual-model collaborative pseudo-label self-training framework. This framework simultaneously exploits the complementary strengths of the source pre-trained model $Model_{pre}$ and the Stage-I fine-tuned model $Model_{ft}$ through multi-level consistency verification to generate filtered pseudo-labels, and then performs supervised self-training on the initial model with an adaptive weighting strategy.

3.4.1 Multi-Level Consistency Pseudo-Label Filtering

To select highly trustworthy disparity predictions from both models, we design a multi-level verification mechanism that combines geometric consistency and photometric consistency. Taking the left view as an example, for any disparity map D^l predicted by a model, a pixel is considered trustworthy only if all the following conditions are satisfied:

- **Left-right consistency**: $|D^{r\to l} - D^l| < \tau_c$, where $D^{r\to l}$ is the left disparity map reconstructed from the right disparity map.
- **Cyclic consistency**: $|D^{l\to r\to l} - D^l| < \tau_c$ and $|D^{l\to r\to l} - D^{r\to l}| < \tau_c$.
- **Photometric consistency**: reconstruction errors $|\mathcal{L}_{rec}(I^{r\to l}, I^l)| < \epsilon_c$ and $|\mathcal{L}_{rec}(I^{l\to r\to l}, I^l)| < \epsilon_c$.

Here τ_c and ϵ_c are predefined thresholds. Combining the above conditions, the pseudo-label filtering mask for a single model is as shown in (10).

$$\begin{aligned} M^l =& \mathbb{I}\big(\mathcal{L}_{rec}(I^{r\to l}, I^l) < \epsilon_c \wedge \mathcal{L}_{rec}(I^{l\to r\to l}, I^l) < \epsilon_c\big) \\ &\cdot \mathbb{I}\big(|D^{r\to l} - D^l| < \tau_c \wedge |D^{l\to r\to l} - D^l| < \tau_c \wedge |D^{l\to r\to l} - D^{r\to l}| < \tau_c\big) \end{aligned} \tag{10}$$

Applying the above filtering to $Model_{pre}$ and $Model_{ft}$ respectively yields masks M^{pre} and M^{ft}.

3.4.2 Dual-Model Pseudo-Label Fusion

For any pixel, if at least one model passes the verification, it is included in the pseudo-label candidate set. For the specific disparity value, we adopt a max fusion strategy as shown in (11).

$$D_{\text{pseudo}} = \max\left(D^{pre} \odot M^{pre}, D^{ft} \odot M^{ft}\right), \text{where } M^{pre} \vee M^{ft} > 0 \tag{11}$$

That is, for pixels that pass verification in both models, we take the larger disparity value; for pixels that pass only one model, we directly use that model's prediction. In our experiments, replacing the max operator with the min or the mean operator led to comparable accuracy, which suggests that the two candidate values are usually close and that the fusion operator itself is not the dominant factor. We therefore adopt the max operator for a practical reason: a larger disparity corresponds to a nearer surface, and in downstream applications such as obstacle avoidance and collision warning, underestimating proximity (i.e., predicting a farther depth) is considerably more risky than overestimating it. Taking the larger of the two verified values thus yields a conservative estimate with respect to collision risk.

3.4.3 Self-Training with High-Confidence Pseudo-Labels

After obtaining filtered pseudo-labels, we use them to perform supervised self-training on the initial source pre-trained model $Model_{pre}$. The loss function is the Smooth L1 loss, consistent with training on ground-truth annotations.

However, since the pseudo-labels become sparse and non-uniform after multi-level filtering, directly using all of them would cause large variations in the number of labels per batch, destabilizing training. To mitigate this, we introduce two strategies:

Adaptive random masking. Let the original density of pseudo-labels in the current batch be $\rho = \frac{1}{N}\sum_{i,j}(M_{ij}^{pre} \vee M_{ij}^{ft})$, and set the target upper density limit as $\sigma = 0.05$. We generate a random mask M^{rand} where each pixel is set to 1 with probability $\min(1, \sigma/\rho)$, thus controlling the effective density around σ.

Local density weighting. We assign a weight W_{ij} to each pseudo-label pixel, determined by the label density in its surrounding 3×3 neighborhood as shown in (12).

$$W_{ij} = \frac{1}{9}\sum_{u=-1}^{1}\sum_{v=-1}^{1} M_{i+u,j+v} \tag{12}$$

Higher density indicates higher reliability and thus larger weight. The final self-training loss is the weighted Smooth L1 loss:

$$Loss_{\text{stage2}} = \frac{1}{\sum_{ij} M_{ij}W_{ij}} \sum_{ij} M_{ij}W_{ij} \cdot \text{SmoothL1}\left(\hat{D}_{ij} - D_{ij}^{\text{pseudo}}\right) \tag{13}$$

where $M = (M^{pre} \vee M^{ft}) \odot M^{rand}$ is the final training mask, and $\hat{D}$ is the disparity map predicted by the current model. Through the above self-training process, the model can further learn the geometric structure of the target domain from filtered pseudo-labels, eventually obtaining the final adapted model $Model_{final}$.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate the effectiveness and generality of the proposed MAD framework on multiple public datasets. The source domain is the SceneFlow [1] synthetic dataset, on which all models are pre-trained. For adaptation we use unlabeled target-domain stereo pairs from KITTI 2012 [20], KITTI 2015 [21], KITTI Raw, CityScapes [22], and DrivingStereo [23].

It should be noted that, due to the limited amount of data available in KITTI 2012 and KITTI 2015, the same data split is used for both adaptation and evaluation, which is also a common practice in the performance comparisons presented in Table 2. In contrast, for KITTI Raw, Cityscapes, and DrivingStereo, different splits are used for adaptation and evaluation. Since KITTI Raw does not provide ground-truth disparity annotations, the effectiveness of adaptation on KITTI Raw is evaluated using the KITTI 2015 dataset. Similarly, Cityscapes does not provide true disparity values, but only reference disparity maps generated using SGM [24]. Therefore, the evaluation results on Cityscapes should be regarded as reference results rather than evaluations against ground-truth disparity. Among them, DrivingStereo additionally provides four weather-condition subsets (sunny, cloudy, foggy, and rainy) to evaluate adaptation capability under extreme environments.

Base Models. To verify the model-agnostic nature of MAD, we select five stereo matching networks with different architectures as base models: PSMNet [16], GwcNet [25], ACVNet [18], CoEx [26], and CFNet [17]. All models are pre-trained on SceneFlow and then adapted using unlabeled image pairs from the target domains.

Evaluation Metrics. We adopt standard metrics in the stereo matching field: End-Point Error (EPE, in pixels) of disparity maps, and the mismatch rate (D1-all), i.e., the proportion of pixels whose disparity error exceeds 3 pixels or the relative error exceeds 5%.

Implementation Details. The MAD framework consists of two stages with uniformly set hyperparameters as follows. Stage I: learning rate 0.0002, batch size 4, non-occlusion mask threshold $\tau = 1\%$ of predicted disparity, hard example threshold $\epsilon = 0.1$, loss weights $\lambda_{rec} = 1$, $\lambda_{smooth} = 10$, $\lambda_{cons} = 0.1$, and training for 5 epochs. Stage II: learning rate 0.001, batch size 4, consistency verification threshold $\tau_c = \min(1\text{px}, 1\% \text{ of predicted disparity})$, $\epsilon_c = 0.5$, pseudo-label upper density limit $\sigma = 0.05$, and training for 10 epochs. These thresholds and coefficients were selected empirically during exploratory experiments rather than through an exhaustive hyperparameter search; we did not observe the results to be sensitive to small perturbations around the reported values, although a dedicated sensitivity study is left as future work. All experiments are conducted on a single NVIDIA A800 GPU.

Table 2 Comparison with domain generalization (DG) and unsupervised domain adaptation (UDA) methods on the KITTI 2012 and KITTI 2015 datasets.

	Method	Extra Modules	KITTI 2012 D1-all (%)	KITTI 2015 D1-all (%)
DG	DSMNet [8]	✓	6.2	6.5
	CFNet [17]	×	4.7	5.8
	UCFNet-pretrain [6]	×	4.5	5.2
	Graft-PSMNet [11]	✓	5.0	5.3
	Graft-GANet [11]	✓	4.6	5.4
	ITSA-PSMNet [12]	✓	5.2	5.8
	ITSA-GwcNet [12]	✓	4.9	5.4
	ITSA-CFNet [12]	✓	4.2	4.7
	FC-PSMNet [13]	✓	5.3	5.8
	FC-GANet [13]	✓	4.6	5.3
UDA	MADNet [7]	×	9.3	8.5
	DaStereo [3]	✓	3.9	3.8
	StereoGAN-GwcNet [2]	✓	–	5.7
	HAG-GwcNet [27]	✓	–	4.65
	Ada-PSMNet [4]	✓	3.6	<u>3.5</u>
	Ada-ResNetCorr [4]	✓	5.1	5.0
	TS-PSMNet [5]	✓	3.5	4.1
	UCFNet-adapt [6]	✓	**2.8**	**3.1**
	MAD-PSMNet(Ours)	×	4.6	4.5
	MAD-GwcNet(Ours)	×	3.8	3.9
	MAD-ACVNet(Ours)	×	3.4	3.6
	MAD-CFNet(Ours)	×	<u>3.3</u>	3.9
	MAD-CoEx(Ours)	×	4.7	4.2

Note: DG methods are trained only on the synthetic SceneFlow dataset, while UDA methods are further adapted using unlabeled KITTI training images. "Extra Modules" indicates whether additional learnable modules or network modifications are introduced during adaptation. The best and second-best results are highlighted in **bold** and <u>underline</u>, respectively.

4.2 Performance Comparison

To examine the performance of MAD relative to representative domain generalization and unsupervised domain adaptation approaches, we pre-train several stereo matching networks on SceneFlow and adapt them using unlabeled KITTI image pairs. The evaluation is conducted on the KITTI training sets using the available sparse disparity annotations. The results are summarized in Table 2.

Comparison with domain generalization methods. Domain generalization methods such as Graft [11], ITSA [12], and FC-Stereo [13] do not access target-domain images during training, whereas MAD uses unlabeled target-domain image pairs for adaptation. This difference makes the two settings complementary rather than directly interchangeable. Under the target-domain adaptation setting, MAD improves the SceneFlow-pretrained models substantially on KITTI. For example, on KITTI 2015, MAD-PSMNet, MAD-GwcNet, and MAD-ACVNet obtain D1-all values of 4.5%, 3.9%, and 3.6%, respectively.

Comparison with domain adaptation methods. Compared with MADNet [7], which also uses unsupervised target-domain fine-tuning without additional learnable adaptation parameters, MAD obtains consistently lower D1-all values in the tested settings. Compared with methods that employ adversarial learning or additional

adaptation modules, such as StereoGAN [2], HAG [27], DaStereo [3], AdaStereo [4], TS-Nets [5], and UCFNet [6], MAD does not consistently achieve the best absolute accuracy. For example, UCFNet obtains lower D1-all on both KITTI 2012 and KITTI 2015, while AdaStereo and TS-Nets are also competitive. The main distinction of MAD is therefore not absolute benchmark leadership, but the use of a relatively simple adaptation procedure without architecture-specific learnable modules.

Comparison across base models. Within the MAD framework, the relative performance of the adapted models largely follows the capability of their underlying pretrained backbones. On KITTI 2015, MAD-ACVNet (3.6%) performs better than MAD-GwcNet (3.9%) and MAD-PSMNet (4.5%), while all five backbones benefit from the same adaptation procedure. This consistency across heterogeneous architectures supports the practical applicability of the proposed framework.

4.3 Generalization Experiments

To examine the transferability of the proposed adaptation strategy across different datasets and scenarios, we conduct experiments on multiple public datasets based on GwcNet. All experiments adopt the conventional unsupervised fine-tuning approach as the baseline, and compare it with our proposed MAD-GwcNet. It should be noted that existing unsupervised domain adaptation methods for stereo matching mostly report results on the KITTI benchmark, and to the best of our knowledge they do not provide comparable numbers on CityScapes or DrivingStereo. On these two datasets we therefore compare against the SceneFlow-pretrained model and the conventional unsupervised fine-tuning baseline, which are the references that can be reproduced under exactly the same setting. The results are elaborated from three aspects below.

Transfer performance across different datasets. We conduct experiments on three datasets with different resolutions and diverse urban complex scenes: KITTI Raw, CityScapes, and DrivingStereo, where the models adapted on KITTI Raw are evaluated on the KITTI 2015 split. The results are shown in Table 3. Directly testing the pre-trained GwcNet model on all three datasets yields poor performance, especially on DrivingStereo where D1-all reaches as high as 44.1%. After conventional unsupervised fine-tuning (baseline), the model performance improves significantly; for example, on DrivingStereo, D1-all drops from 44.1% to 9.8%. After applying our MAD method for domain adaptation, the improvements are better than the baseline: on KITTI Raw, D1-all decreases from 5.9% to 4.1%; on DrivingStereo, it drops sharply from 9.8% to 3.0%; and on CityScapes, it slightly decreases from 2.5% to 2.4%. An exception occurs on the CityScapes EPE, where MAD-GwcNet (0.83 px) is slightly worse than the baseline (0.78 px). We attribute this to the reference labels of CityScapes being produced by SGM [24] rather than being true disparity: SGM itself yields over-smoothed disparity maps, so this metric favors the smoother predictions of the baseline, whereas MAD deliberately preserves sharper object boundaries. Because these labels are pseudo-labels, the CityScapes results are reported for reference only. Fig. 4 presents the visual comparisons of different methods on the three datasets. It can be observed that the pre-trained model suffers from severe performance degradation in cross-domain scenarios. In contrast, MAD-GwcNet outperforms the baseline method in handling occluded regions and object edges.

Table 3 Comparison with the pre-trained model and baseline on multiple datasets. Best results are in bold.

Method	KITTI Raw		CityScapes		DrivingStereo	
	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)
Pre-trained	12.83	2.24	11.1	3.07	44.1	14.4
Baseline	5.9	1.14	2.5	**0.78**	9.8	2.81
Ours	**4.1**	**0.99**	**2.4**	0.83	**3.0**	**1.29**

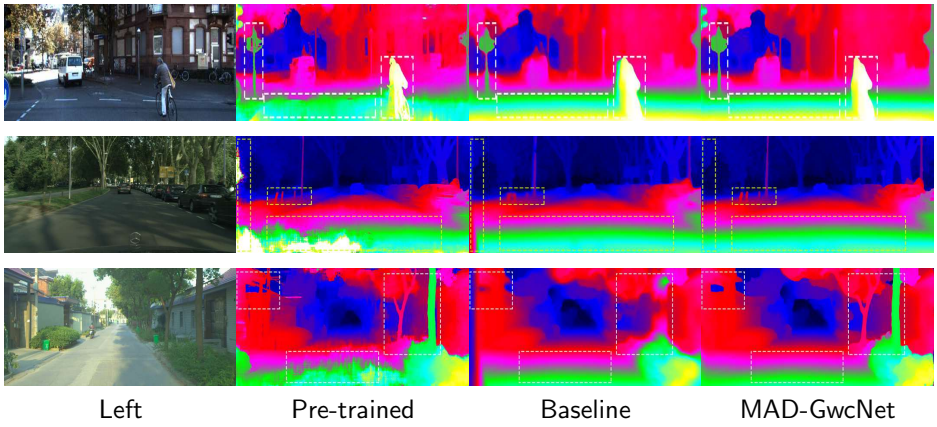


Fig. 4 Visualization of comparison with the pre-trained model and the baseline on multiple datasets.

Performance under adverse weather conditions. We further conduct comparative experiments on four extreme weather scenarios (sunny, cloudy, foggy, and rainy) from the DrivingStereo dataset to evaluate the method's generality under different environmental conditions. The results are shown in Table 4. The pre-trained model performs poorly under all extreme weather conditions, especially in foggy and rainy scenes where D1-all reaches as high as 60.8% and 77.7%. After conventional unsupervised fine-tuning, performance improves significantly. MAD-GwcNet consistently outperforms the baseline across all weather scenarios: compared with the baseline, D1-all drops from 12.5% to 6.1% on foggy days, and from 22.1% to 8.3% on rainy days. The improvements are substantial even under severe weather conditions such as fog and rain, indicating that the proposed adaptation strategy remains effective under the tested adverse weather conditions. The visual results in Fig. 5 provide qualitative evidence consistent with these quantitative results.

Cross-domain transfer performance. Recent studies [14] have pointed out that stereo matching models suffer from degraded generalization performance on unseen domains after supervised domain adaptation with target-domain annotated data. We further examine the cross-domain transferability of the adapted models. We perform domain adaptation on MAD-GwcNet using three datasets respectively, and then apply each adapted model to the other datasets for testing. The results are shown in Table 5. Although the best performance is always achieved by the model adapted on the corresponding target dataset, models adapted on other datasets also retain useful

Table 4 Comparison with the pre-trained model and baseline on multiple extreme weather datasets. Best results are in bold.

Method	Sunny		Cloudy		Foggy		Rainy	
	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)
Pre-trained	36.1	7.99	34.2	8.19	60.8	20.35	77.7	41.8
Baseline	9.5	2.86	7.6	2.52	12.5	2.47	22.1	4.07
Ours	**3.0**	**1.19**	**2.5**	**1.20**	**6.1**	**1.47**	**8.3**	**1.84**

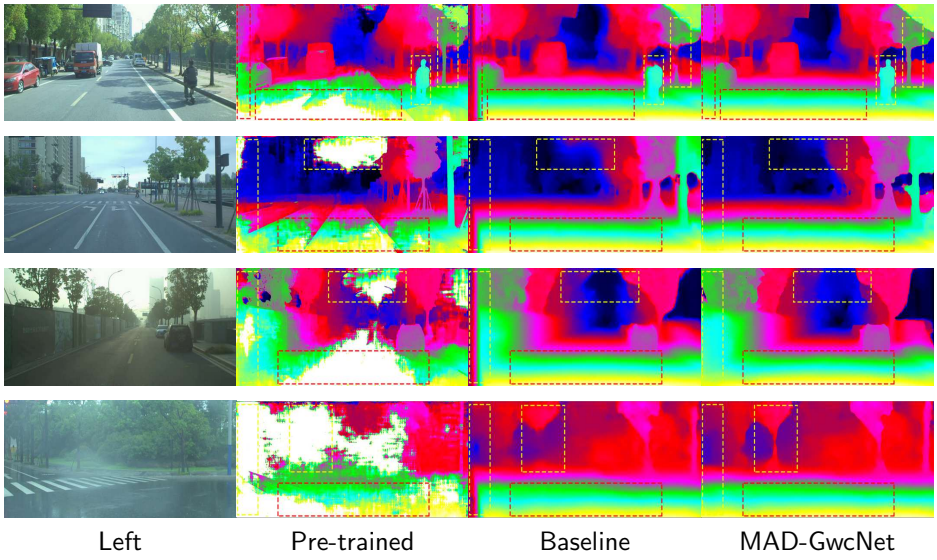


Fig. 5 Visualization of comparison with the pre-trained model and the baseline under extreme weather conditions (DrivingStereo subsets).

transferability. Taking DrivingStereo as an example, models adapted on KITTI Raw and CityScapes achieve D1-all of 6.0% and 6.1%, respectively. These values are higher than the 3.0% obtained by the model adapted on DrivingStereo itself, but remain lower than the pre-trained model (44.1%) and the baseline fine-tuned on DrivingStereo (9.8%). These results suggest that adaptation on one target domain does not necessarily remove the model's transferability to other target domains.

4.4 Ablation Studies

To examine the contribution of the main components in the proposed framework, we conduct ablation experiments on the KITTI 2015 set: as in the performance comparison, each base model is adapted on the unlabeled KITTI 2015 training images and evaluated on the same split with its sparse disparity annotations. We select five stereo matching networks with different architectures and performance levels as base models, including PSMNet, GwcNet, ACVNet, CoEx, and CFNet. All models are initialized with SceneFlow pre-trained weights. The experimental setup and results are

Table 5 Cross-domain experiments on multiple datasets.

Method	KITTI Raw		CityScapes		DrivingStereo	
	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)
Pre-trained	12.83	2.24	11.1	3.07	44.1	14.4
Baseline	5.9	1.14	2.5	**0.78**	9.8	2.81
MAD-GwcNet (K)	**4.1**	**0.99**	2.6	0.87	6.0	1.62
MAD-GwcNet (C)	4.9	1.09	**2.4**	0.83	6.1	1.76
MAD-GwcNet (D)	5.5	1.15	2.5	0.89	**3.0**	**1.29**

Note:Best results are highlighted in **bold**. K, C, and D denote models adapted on KITTI Raw, CityScapes, and DrivingStereo, respectively, and each adapted model is evaluated on all three datasets.

Table 6 Ablation experiments with multiple base models. Following the KITTI protocol of Section 4.2, all models are adapted on the unlabeled KITTI 2015 training images and evaluated on the same split using its sparse disparity annotations. Best results are in bold.

Base Model	PSMNet		GwcNet		ACVNet		CoEx		CFNet	
	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)
Pre-trained	28.20	4.05	12.83	2.24	11.44	2.50	14.57	2.47	5.76	2.23
Baseline	5.54	1.21	5.12	1.09	4.50	1.06	5.13	1.20	4.27	0.96
+ Non-occ mask	4.69	1.07	4.16	1.03	3.70	**0.98**	4.42	**1.00**	4.20	0.95
+ Single-model	4.52	**1.04**	3.94	1.01	3.66	**0.98**	4.38	1.02	4.01	**0.93**
+ Dual-model	**4.46**	1.05	**3.93**	**0.98**	**3.59**	0.99	**4.24**	1.01	**3.88**	**0.93**

shown in Table 6. The baseline method is conventional unsupervised fine-tuning using target-domain image pairs with traditional unsupervised losses. On this basis, we sequentially introduce three key components: (1) unsupervised fine-tuning weighted by non-occlusion masks; (2) single-model self-training using the unsupervised fine-tuned model; and (3) dual-model collaborative self-training that combines both the pre-trained model and the unsupervised fine-tuned model. We verify the effectiveness of each component below.

Effectiveness of non-occlusion mask weighting. Compared with direct testing of the pre-trained model, conventional unsupervised fine-tuning already significantly improves performance on KITTI. After introducing the non-occlusion mask-based weighting strategy for unsupervised fine-tuning, all models achieve further improvements: compared with the baseline, PSMNet's D1-all decreases from 5.54% to 4.69%, GwcNet from 5.12% to 4.16%, and ACVNet from 4.50% to 3.70%. This indicates that excluding occluded regions from the corresponding unsupervised losses can provide a cleaner supervision signal for target-domain fine-tuning. Notably, even though

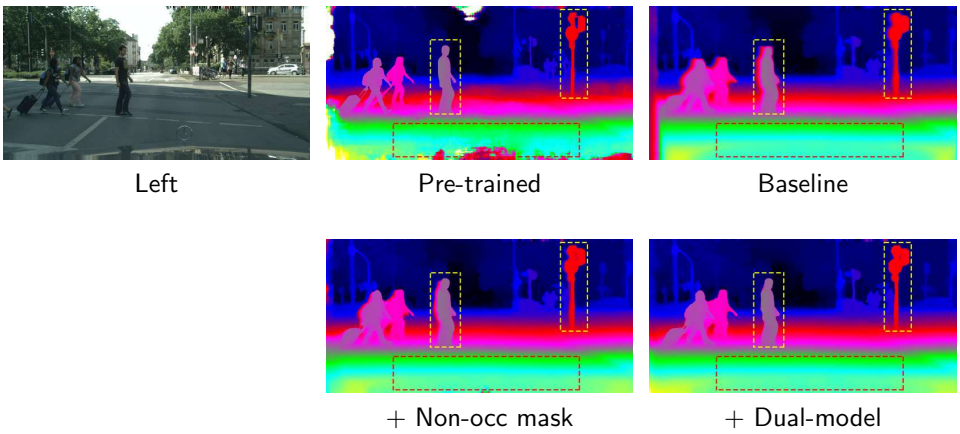


Fig. 6 Visualization of disparity maps in ablation experiments.

PSMNet's pre-trained model has a high error rate of 28.20%, resulting in relatively poor and sparse non-occlusion masks, introducing these masks still yields significant improvements. The same trend across different base models suggests that the mask weighting strategy is not strongly tied to a particular backbone.

Effectiveness of self-training strategies. After obtaining the unsupervised fine-tuned model, we further investigate pseudo-label-based self-training strategies. First, using the pseudo-labels generated by the fine-tuned model for single-model self-training on the initial model yields consistent improvements over the fine-tuned model itself: PSMNet's D1-all drops from 4.69% to 4.52%, GwcNet from 4.16% to 3.94%, and CoEx from 4.42% to 4.38%. Furthermore, by simultaneously leveraging both the pre-trained model and the unsupervised fine-tuned model for dual-model collaborative pseudo-label self-training, performance improves again over single-model self-training. Generally, the stronger the generalization capability of the pre-trained model, the more significant the incremental supervision signal it provides; for instance, CFNet (with better generalization) improves from 4.01% to 3.88%, while PSMNet (with weaker generalization) only improves from 4.52% to 4.46%. In addition, for models with similar generalization performance such as GwcNet and CoEx, the improvement magnitudes differ (GwcNet from 3.94% to 3.93%, CoEx from 4.38% to 4.24%). This may be attributed to different degrees of complementarity between the pre-trained and fine-tuned model predictions. Despite these differences, dual-model collaborative self-training provides a further improvement for all five tested backbones.

Furthermore, we visualize the experimental results using GwcNet on CityScapes as an example. As shown in Fig. 6, the pre-trained model exhibits severe performance degradation in cross-domain scenarios, with large-area recognition errors on sky and ground regions, yet it maintains relatively accurate edge predictions for some human figures and objects. The baseline method (conventional unsupervised fine-tuning) achieves some performance improvement, but due to its failure to distinguish occluded regions, its simple cropping and occlusion handling leads to recognition errors on the left side of the image, incorrect object occlusion ordering, and edge over-expansion.

After introducing non-occlusion mask weighting, although some errors in occluded regions remain, the object edges in non-occluded areas become clearer, and the errors on the left side disappear. In the final dual-model collaborative self-training stage, the model essentially combines the advantages of both the pre-trained and fine-tuned models: it achieves accurate recognition of large-area regions such as sky and ground while retaining the ability to handle occluded areas with sharp edges, reaching the best overall performance.

Overall, the ablation experiments show the individual and combined contributions of the main components in MAD: the non-occlusion mask weighting strategy effectively removes reconstruction noise, single-model self-training successfully inherits the capabilities of the unsupervised fine-tuned model, and dual-model collaborative self-training further integrates the generalization advantages of the pre-trained model.

4.5 Summary of Experiments

Through extensive experiments on multiple datasets and various base models, this chapter validates the effectiveness and generality of the proposed MAD framework. MAD consistently improves the performance of pretrained stereo matching models over conventional unsupervised fine-tuning in the tested target domains, including adverse-weather and cross-domain transfer settings. The ablation studies further confirm the individual contributions of the non-occlusion mask, single-model self-training, and dual-model collaborative self-training components.

5 Conclusion

In this paper, we presented MAD, an unsupervised domain adaptation framework for stereo matching that combines mask-guided self-supervised fine-tuning with dual-model pseudo-label self-training. The first stage uses left-right consistency of the source pretrained model to identify a high-precision subset of non-occluded regions and to reduce unreliable photometric supervision during target-domain fine-tuning. The second stage jointly considers predictions from the pretrained and fine-tuned models, filters candidate disparity values using geometric and photometric consistency, and uses the resulting pseudo-labels for further self-training.

Experiments on multiple target datasets and across five stereo matching backbones show that MAD consistently improves over conventional unsupervised fine-tuning in the tested settings. The results also indicate that combining the complementary predictions of the source pretrained model and the target-adapted model can provide additional supervision for self-training. Cross-domain transfer experiments suggest that adaptation on one target domain can retain useful transferability to other domains.

The proposed framework is not designed to replace architecture-specific UDA methods in terms of absolute disparity accuracy. Instead, its main value lies in providing a relatively simple and architecture-independent adaptation strategy with limited implementation requirements. Its performance still depends on the quality of the source pretrained model, the coverage of the non-occlusion masks, and the manually selected consistency thresholds. These limitations provide directions for future work

on more adaptive confidence estimation and threshold selection while preserving the simplicity and architecture independence of the framework.

Statements and Declarations

Funding This work was supported in part by the National Natural Science Foundation of China under Grant 62372188, in part by the Guang Dong Natural Science Foundation under Grant 2024A1515010100, in part by the Key Core and Pivotal Technology Research Project of the Hengqin Guangdong-Macao In-Depth Cooperation Zone under Grant 2530004025492.

Competing Interests The authors declare that they have no competing interests.

Data availability The public datasets used in this study (SceneFlow, KITTI 2012/2015, Cityscapes, DrivingStereo) are available from their respective sources. The derived experimental results are available from the corresponding author upon reasonable request.

Code availability The code used in this study is available from the corresponding author upon reasonable request.

Author contributions Jie Tang: supervision, methodology, writing—review and editing. Gaofeng Peng: methodology, experiments, writing—original draft. Jialu Liu: investigation, validation. Shaoshan Liu: resources, supervision.

References

[1] Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: 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, pp. 4040–4048 (2016)

[2] Liu, R., Yang, C., Sun, W., Wang, X., Li, H.: StereoGAN: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12754–12763 (2020). https://doi.org/10.1109/CVPR42600.2020.01277

[3] Zhang, C., Meng, G., Xu, R.Y.D., Xiang, S., Pan, C.: Learning adversarial point-wise domain alignment for stereo matching. Neurocomput. **491**(C), 564–574 (2022) https://doi.org/10.1016/j.neucom.2021.12.034 . Number of pages: 11 tex.issue_date: Jun 2022

[4] Song, X., Yang, G., Zhu, X., Zhou, H., Wang, Z., Shi, J.: AdaStereo: a simple and efficient approach for adaptive stereo matching. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10323–10332 (2021). https://doi.org/10.1109/CVPR46437.2021.01019

[5] Yue, X., Lu, Z., Lin, X., Ren, W., Shao, Z., Hu, H., Zhang, Y., Liao, Q.:

Semi-stereo: a universal stereo matching framework for imperfect data via semi-supervised learning. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 646–655 (2024). https://doi.org/10.1109/CVPRW63382.2024.00069

[6] Shen, Z., Song, X., Dai, Y., Zhou, D., Rao, Z., Zhang, L.: Digging into uncertainty-based pseudo-label for robust stereo matching. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(12), 14301–14320 (2023) https://doi.org/10.1109/TPAMI.2023.3300976

[7] Tonioni, A., Tosi, F., Poggi, M., Mattoccia, S., Stefano, L.D.: Real-time self-adaptive deep stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 195–204 (2019)

[8] Zhang, F., Qi, X., Yang, R., Prisacariu, V., Wah, B., Torr, P.: Domain-invariant stereo matching networks. In: Europe Conference on Computer Vision (ECCV) (2020)

[9] Cai, C., Poggi, M., Mattoccia, S., Mordohai, P.: Matching-space stereo networks for cross-domain generalization. In: 2020 International Conference on 3D Vision (3DV), pp. 364–373 (2020). https://doi.org/10.1109/3DV50981.2020.00046

[10] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE, Piscataway, NJ (2009)

[11] Liu, B., Yu, H., Qi, G.: GraftNet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13002–13011 (2022). https://doi.org/10.1109/CVPR52688.2022.01267

[12] Chuah, W., Tennakoon, R., Hoseinnezhad, R., Bab-Hadiashar, A., Suter, D.: ITSA: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13012–13022 (2022). https://doi.org/10.1109/CVPR52688.2022.01268

[13] Zhang, J., Wang, X., Bai, X., Wang, C., Huang, L., Chen, Y., Gu, L., Zhou, J., Harada, T., Hancock, E.R.: Revisiting domain generalized stereo matching networks from a feature consistency perspective. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12991–13001 (2022). https://doi.org/10.1109/CVPR52688.2022.01266

[14] Zhang, J., Li, J., Huang, L., Yu, X., Gu, L., Zheng, J., Bai, X.: Robust synthetic-to-real transfer for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20247–20257 (2024)

[15] Tosi, F., Tonioni, A., De Gregorio, D., Poggi, M.: NeRF-supervised deep stereo. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 855–866 (2023)

[16] Chang, J.-R., Chen, Y.-S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5410–5418 (2018)

[17] Shen, Z., Dai, Y., Rao, Z.: Cfnet: Cascade and fused cost volume for robust stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13906–13915 (2021)

[18] Xu, G., Wang, Y., Cheng, J., Tang, J., Yang, X.: Accurate and efficient stereo matching via attention concatenation volume. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)

[19] Godard, C., Aodha, O.M., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6602–6611 (2017). https://doi.org/10.1109/CVPR.2017.699

[20] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361. IEEE, Piscataway, NJ (2012)

[21] Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3061–3070 (2015)

[22] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)

[23] Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., Zhou, B.: 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, pp. 899–908 (2019)

[24] Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on pattern analysis and machine intelligence **30**(2), 328–341 (2007)

[25] Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3273–3282 (2019)

[26] Bangunharcana, A., Cho, J.W., Lee, S., Kweon, I.S., Kim, K.-S., Kim, S.: Correlate-and-excite: Real-time stereo matching via guided cost volume excitation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3542–3548. IEEE, Piscataway, NJ (2021)

[27] Wang, S., Seo, J., Jeon, H., Lim, S., Park, S., Lim, Y.: Horizontal attention based generation module for unsupervised domain adaptive stereo matching. IEEE Robotics and Automation Letters **8**(10), 6779–6786 (2023) https://doi.org/10.1109/LRA.2023.3313009