Consistency-aware Self-Training for Iterative-based Stereo Matching

Jingyi Zhou[1*], Peng Ye[1,3,4*], Haoyu Zhang[1*], Jiakang Yuan[1]
Rao Qiang[2], Liu YangChenXu[2], Wu Cailin[2], Feng Xu[1], Tao Chen[1†]

[1]Fudan University [2]Xiaomi Inc [3]Shanghai AI Laboratory [4]The Chinese University of Hong Kong

Abstract

Iterative-based methods have become mainstream in stereo matching due to their high performance. However, these methods heavily rely on labeled data and face challenges with unlabeled real-world data. To this end, we propose a consistency-aware self-training framework for iterative-based stereo matching for the first time, leveraging real-world unlabeled data in a teacher-student manner. We first observe that regions with larger errors tend to exhibit more pronounced oscillation characteristics during model prediction. Based on this, we introduce a novel consistency-aware soft filtering module to evaluate the reliability of teacher-predicted pseudo-labels, which consists of a multi-resolution prediction consistency filter and an iterative prediction consistency filter to assess the prediction fluctuations of multiple resolutions and iterative optimization respectively. Further, we introduce a consistency-aware soft-weighted loss to adjust the weight of pseudo-labels accordingly, relieving the error accumulation and performance degradation problem due to incorrect pseudo-labels. Extensive experiments demonstrate that our method can improve the performance of various iterative-based stereo matching approaches in various scenarios. In particular, our method can achieve further enhancements over the current SOTA methods on several benchmark datasets.

1. Introduction

Stereo matching aims to predict the pixel-wise disparity between a pair of rectified images [18]. With the rapid development of 3D vision tasks and their applications such as robotics and autonomous driving, stereo matching has become a fundamental vision task to provide depth information in the real 3D world. Although existing stereo matching methods have achieved remarkable performance on certain benchmark datasets [12, 22], they heavily rely on the number and quality of labeled data. However, due to hardware and environmental constraints, obtaining high-quality

---

*Equal Contribution
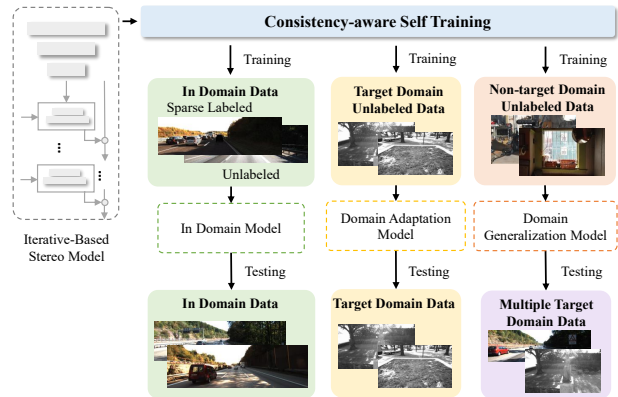†Corresponding author.

Figure 1. Illustration of our Consistency-aware Self-Training (CST) method, which can leverage unlabeled data in multiple scenarios to boost the iterative-based stereo-matching model.

labels for stereo data is known to be costly and sometimes even infeasible. Thus, current labeled stereo data is predominantly composed of synthetic datasets, leading to suboptimal performance when applied to unlabeled real-world data. Unlike the scarcity of labeled real-world stereo data, unlabeled data is relatively easier to acquire while often ignored by previous works.

Self-training has been proven to be an effective approach to leveraging unlabeled data to enhance model performance and generalization ability via a teacher-student manner. As a result, a few recent works [23, 29] attempt to apply self-training on 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 obtained via the matching probability distribution of cost volume. However, these cost-volume-based self-training methods still encounter the following challenges: 1) As the iterative-based methods led by RAFT-Stereo [14] gradually become mainstream, these methods may fail to adapt effectively since the iterative-based methods employ recurrent units to update the disparity with multi-source information and sometimes without calculating the whole cost volume. 2) The reliability computation of these methods resembles classification, treating the pixels along the epipolar line as independent discrete labels. It focuses on single-point con-

centration while neglecting the spatial concentration along the epipolar axis, which may lead to a significant discrepancy between the predicted and true reliability. 3) These methods follow the all-or-nothing threshold filtering approach, where pseudo-labels are either selected or discarded with no intermediate state. Considering that pseudo-labels within the threshold are not equally reliable and those not selected still hold potential training value, such a rigid binary selection approach poses a bottleneck for performance improvement, particularly in challenging areas.

To address these issues, we propose a novel consistency-aware self-training framework for iterative-based stereo matching for the first time. Our framework is based on the insightful observation that regions with larger errors tend to exhibit more pronounced oscillation characteristics during model prediction. Motivated by this, we propose a consistency-aware soft filtering module that evaluates the reliability of teacher-predicted pseudo-labels for soft filtering unreliable pseudo-labels to mitigate error accumulation and performance degradation. It consists of a multi-resolution prediction consistency filter and an iterative prediction consistency filter. The former is used to assess the prediction fluctuations of multiple resolutions since error-prone pixels often display instability and inconsistency with the changing of image resolution. The latter is used to assess the prediction fluctuations of iterative optimization since error-prone areas strongly correlate with fluctuating regions of iterative predictions. Further, we introduce a consistency-aware soft-weighted loss to assign higher weights to predictions with greater reliability, reducing the impact of teacher-predicted incorrect pseudo-labels while keeping the student's possibility to learn from diverse and challenging areas. Our method: 1) introduces an effective self-training strategy for iterative-based stereo matching without relying on the cost volume; 2) computes the reliability suitable for regression by accounting for the concentration along the epipolar axis rather than isolated peak values, which is more accurate. 3) employs soft filtering to avoid absolute binary selection, ensuring a more nuanced pseudo-label selection process.

To demonstrate the effectiveness of the proposed method, comprehensive experiments are conducted across multiple challenging scenarios, including in domain, domain adaptation and domain generalization, as shown in Fig. 1. The experimental results verify that our method significantly enhances the performance and generalization capability of various iterative-based stereo matching models, enabling them to perform well not only on unlabeled datasets but also on real-world datasets that have not been encountered during training. Especially, our method can achieve further improvements over state-of-the-art methods and on the Middlebury, KITTI2015, and ETH3D datasets,

Our contributions are briefly summarized as follows:

- We propose a novel consistency-aware self-training framework for iterative-based stereo matching for the first time, based on the insightful observation that regions with larger errors tend to exhibit more pronounced oscillation characteristics during model prediction.
- We design a consistency-aware soft filtering module to assess the pixel-level reliability of teacher-predicted pseudo-labels from spatial and temporal dimensions, which includes a multi-resolution prediction consistency filter and an iterative prediction consistency filter.
- We introduce a consistency-aware soft-weighted loss for the training of the student model, which reduces the impact of the teacher model's errors on the student model while maintaining the student model's possibility to learn from diverse and challenging areas.
- Our method achieves impressive results in various scenarios, including in domain, domain adaptive and domain generalization, achieving new SOTA results among the published works.

2. Related Work

2.1. Stereo Matching

Cost-volume-based methods. Since Zbontar and LeCun first applied machine learning methods to the feature matching stage of stereo matching, traditional matching algorithms [2, 10] have gradually replaced by learning-based methods [7, 15, 25]. Initially, cost-volume-based methods led the development. PSMNet [5] utilized feature concatenation to build a 4D cost volume. HITNet [21] introduced a swift multi-resolution initialization phase, differentiable 2D geometric propagation, and warping mechanisms that balance speed and accuracy. Although these methods have made great progress compared to traditional methods, the huge memory consumption of these methods for inferring high-resolution images limits their further improvement.

Iterative-based methods. In recent years, more and more works focused on iterative-based stereo matching. RAFT-Stereo [14] firstly explored the use of iterative multi-scale update blocks composed of GRUs [6] to produce the final disparity map by progressing from coarse to fine. CREStereo [12] employed an adaptive group correlation layer to mitigate the effects of cameras not being perfectly aligned in reality. IGEV-Stereo [24] introduced an additional Geometry Encoding Volume to compensate for the missing non-local geometry information in all-pairs correlations. EAI-Stereo [31] and DLNR [32] explored the use of LSTM [8] in iterative module. Selective-Stereo [22] proposes a Selective Recurrent Unit to relieve the loss of detailed information. Even though these iterative-based methods performed well in public benchmarks, they still face challenges in real-world scenes due to the scarcity of labeled data. This drives us to think about how to leverage

the unlabeled data to improve not only the benchmark performance but also the generalization ability of these models.

2.2. Self Training

Self-training is a classical method [11] and has been widely used in various fields, such as image classification [4], object detection [13], and semantic segmentation [27]. The typical approach to self-training involves using a teacher-student framework with pseudo-labeling to leverage unlabeled data during training. However, self-training often encounters confirmation bias, where the student inherits and reinforces the teacher's errors. To mitigate this, two strategies are commonly employed. First, applying strong perturbations to the student's input can help decouple predictions [27, 28]. Second, filtering pseudo-labels enhances their reliability, preventing the propagation of errors from teacher to student. In stereo matching, self-training has been applied to cost-volume-based methods [23, 29], which use cost-volume to assess label reliability. However, these methods are not effectively applicable to iterative-based approaches that predict prediction maps with recurrent networks receiving multi-source information. This research gap encourages us to explore self-training methods applicable to iterative-based approaches.

3. Method

The overview of CST-Stereo is shown in Fig. 2, given an unlabeled stereo pair, we first obtain the pseudo-labels using the teacher model. Then, the multi-resolution consistency score w_{rc} and iterative consistency score w_{ic} can be obtained by the consistency-aware soft filtering module (CSF) which is composed of a multi-resolution prediction consistency filter (MRPCF) and an iterative prediction consistency filter (IPCF). CSF assesses the reliability of the pseudo-labels for soft filtering unreliable pseudo-labels to mitigate error accumulation caused by the inaccurate predictions of the teacher model. Besides, the final soft filtering weights can be calculated according to w_{rc} and w_{ic} and can be utilized as the weight of pseudo-label supervision for the student model to update the student model. Finally, the teacher model is updated using exponential moving average (EMA) after several training iterations, inspired by the DINO [3]. Below we first give the problem definition and then detail each module of our design.

3.1. Problem Definition

The workflow of our framework is illustrated in Algorithm 1. Given a labeled dataset $\mathcal{D}^l$ and an unlabeled dataset $\mathcal{D}^u$ (*e.g.*, labeled synthetic data and unlabeled real-world data), we first train a model on $\mathcal{D}^l$ to obtain the initial weights for both the teacher model T and the student model S. Then, given an unlabeled stereo pair $X = (I_l, I_r)$ sampled from $\mathcal{D}^u$, our goal is to improve the performance of

---

Algorithm 1: Workflow of CST-Stereo

Input: Labeled dataset $\mathcal{D}^l$, Unlabeled dataset $\mathcal{D}^u$,
  Strong augmentations $\mathcal{A}^s$, Upscale $\mathcal{F}^u$,
  Downscale $\mathcal{F}^d$, Teacher/student model T/S,
  The update interval K of the teacher model
Output: Self-trained student model S
Train a model on $\mathcal{D}^l$ with weight $\mathcal{W}$.
Initialize T and S with $\mathcal{W}$ and set current iteration
  $k = 0$.
for $X \subset \mathcal{D}^u$ **do**
  $X^H, X^L = \mathcal{F}^u(X), \mathcal{F}^d(X)$
  $P^H, P^O, P^L = T(X^H), T(X), T(X^L)$
  Intermediate predictions of $T(X)$: $\{P^i\}_{i=\lceil n/2 \rceil}^n$.
  Calculate w_{rc} and w_{ic} using Eq. (2) and Eq. (4)
    based on P^H, P^O, P^L, and $\{P^i\}_{i=\lceil n/2 \rceil}^n$.
  Calculate w_{soft} using Eq. (5) based on w_{rc}, w_{ic}.
  Calculate the prediction of student model
    $\hat{P} = S(\mathcal{A}^s(X))$.
  Update S to minimize L_{ST} calculated by Eq. (6)
    based on $w_{soft}, \hat{P}$, and P^O.
  $k \leftarrow k + 1$
  if $k\%K == 0$ **then**
    | $\theta_T \leftarrow \lambda\theta_T + (1 - \lambda)\theta_S$
  else
    | pass
return S

---

iterative-based stereo matching models by leveraging unlabeled data in a self-training manner.

3.2. Consistency-Aware Soft Filtering Module

Self-training has been proven to effectively utilize unlabeled data to improve the model performance and generalization ability. However, the effectiveness of self-training heavily relies on the quality of the pseudo-labels. Since pseudo-labels often contain inaccurate predictions, directly training on the pseudo-labels without filtering may result in poor model performance as shown in Tab. 3. To obtain more accurate pseudo-labels and avoid performance degradation caused by noisy supervision, we propose a consistency-aware soft filtering module (CSF) that filters pseudo-labels based on the estimated reliability. Instead of employing the commonly used binary filtering strategy, we adopt a soft thresholding approach by assigning each pixel an independent weight according to calculated reliability. This design reduces the potential impact of errors, thereby ensuring the stability of model training while still allowing the model to learn from challenging regions. Otherwise, some difficult areas such as edges and occluded regions might never be adequately learned.
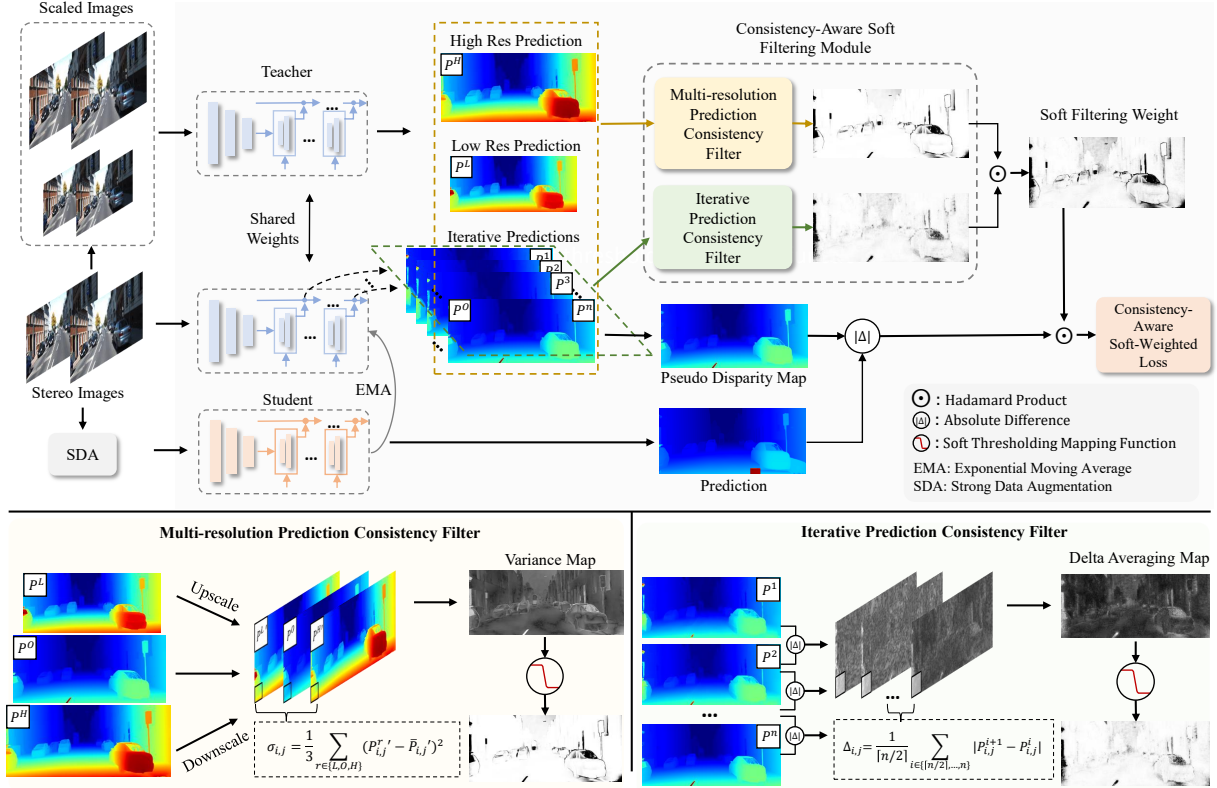
Figure 2. Overview of our proposed CST-Stereo. Our framework leverages unlabeled stereo data in a teacher-student manner. In detail, the student receives strongly augmented images and learns from the teacher's predictions, with model parameters updated in a delayed manner with EMA for cyclic enhancement. Further, the consistency-aware soft filtering module is applied to evaluate the reliability of teacher-predicted pseudo-labels, which includes a multi-resolution prediction consistency filter and an iterative prediction consistency filter. Finally, the consistency-aware soft-weighted loss is calculated for optimization.

3.2.1. Multi-resolution Prediction Consistency Filter

To filter noisy prediction with low reliability, we first develop a multi-resolution prediction consistency filter (MR-PCF) based on our observation. **Observations:** As a pixel-level dense disparity regression task, stereo matching is particularly sensitive to changes in spatial scale information. Generally, at lower resolutions, the model demonstrates better global semantic capture and is less affected by noise, while at higher resolutions, it focuses more on detailed information. Due to the transitions, the predictions for some pixels become inconsistent. According to the experiments and analysis, we identify a significant correlation between the consistency of pixel predictions across different resolutions and their reliability, as shown in Fig. 3. Specifically, pixels that maintain consistent predictions across varying resolutions are more likely to be reliable. This discovery provides a robust mechanism for identifying and addressing areas of uncertainty within the model's output.

Design: Based on the observation, we first evaluate the accuracy of the pseudo disparity map by calculating the consistency of the multi-resolution prediction. Specifically, we calculate the pixel-wise variance among predic-

tions from the high, original, and low resolutions. In detail, the teacher model takes upscaled, original, and downscaled stereo images: X^H, X, and X^L as inputs to generate disparity predictions: P^H, P^O, and P^L respectively. The variance map σ is then calculated as:

$$\sigma_{i,j} = \frac{1}{3} \sum_{r \in \{H,O,L\}} (P_{i,j}^{r}{}' - \bar{P}_{i,j}')^2, \quad (1)$$

where $P_{i,j}^{r}{}'$ represents the resized disparity map of $P_{i,j}^r$ and $\bar{P}'$ is the average of $P^{H'}$, $P^{O'}$, and $P^{L'}$. Pixels with lower variance indicate that their predicted values are more consistent across spatial transformations without significant deviations, and vice versa. Subsequently, a soft thresholding mapping function is employed to derive the reliability weight for each pixel, converting the variance into a reliability score within the 0-1 range. The function can be written as follows:

$$w_{rc} = \frac{1}{1 + e^{-\varepsilon_1(\sigma - \tau_1)}}, \quad (2)$$

where τ_1 denotes the soft threshold and ε_1 is a scale factor. In this scheme, the greater the variance, the lower the

reliability weight assigned to that pixel, with weights approaching zero for pixels with excessively high variance. The mapping function not only assigns varying weights to pixels based on their reliability but also ensures that pixels with excessive errors have minimal impact on the training process, providing a robust mechanism to weigh pixel predictions based on multi-resolution prediction consistency.

3.2.2. Iterative Prediction Consistency Filter

Although MRPCF performs well for resolution-sensitive errors, which often occur at edges and discontinuous regions, it still falls short in handling complex scenarios and ambiguous areas. Thus, we propose a iterative prediction consistency filter (IPCF) to complement the spatial perspective. **Observations:** Iterative-based stereo matching model uses a recurrent neural network, often composed of GRUs, to incrementally refine predictions from either scratch or a rough estimate, aiming to approximate the true disparity values. During this process, the iterative update module integrates multi-source information, such as local cost values, contextual feature, and geometric information, to produce updated predictions. Generally, predictions initially undergo significant changes but gradually converge to a relatively stable value. However, we observed that some pixels exhibit oscillatory behavior in the later stages of iteration, which strongly correlates with error-prone regions, as shown in Fig. 4. The oscillations likely indicate inconsistencies among the multi-source information in those areas, preventing the model from confidently converging to a single value. Such results also provide us with valuable insights into assessing the reliability of the model predictions.

Design: According to the observation, we propose an iterative prediction consistency filter that aims at finding inaccurate areas in the pseudo label through unstable disparity prediction in iterations. Specifically, we calculate the average of the difference map between the adjacent iteration rounds as follows:

$$\Delta_{i,j} = \frac{1}{\lceil n/2 \rceil} \sum_{k \in \{\lceil n/2 \rceil, \ldots, n\}} |P_{i,j}^{k+1} - P_{i,j}^{k}|, \qquad (3)$$

where n is the total number of iterations, P^k is the prediction of the k-th iteration. $\Delta_{i,j}$ can serve as an indicator of the prediction's stability over iterations. A smaller $\Delta_{i,j}$ implies that the prediction of the current pixel has stabilized after balancing the diverse information sources and thus is more reliable. Similar to the former section, a soft thresholding mapping function is used to shift the delta averaging map to reliability weight:

$$w_{ic} = \frac{1}{1 + e^{-\varepsilon_2(\Delta - \tau_2)}}, \qquad (4)$$

where τ_2 and ε_2 represent the soft threshold and scale factor, respectively.
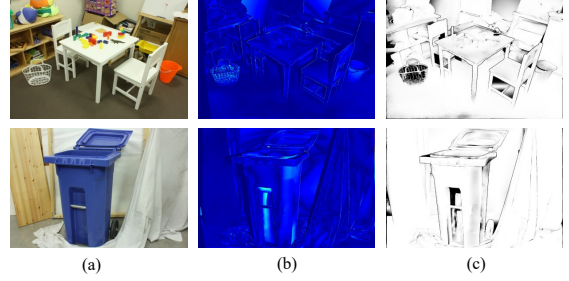

Figure 3. Correlations between error regions and multi-resolution prediction consistency. (a) Referenced image. (b) Error map. (c) Multi-resolution prediction consistency map (Darker areas denote lower consistency).
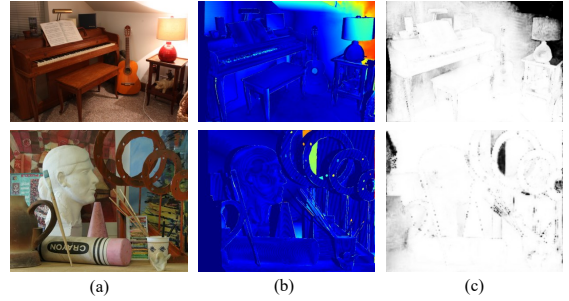

Figure 4. Correlations between error regions and iterative prediction consistency. (a) Referenced image. (b) Error map. (c) Iterative prediction consistency map (Darker areas denote lower consistency).

3.3. Consistency-Aware Soft-Weighted Loss

To take advantage of both filters and reduce the impact of inaccurate pseudo-labels, we propose consistency-aware soft-weighted loss to supervise the student model. In detail, given the weight w_{rc} and w_{ic} from MRPCF and IPCF, the soft filtering weight is obtained by multiplying the weights, effectively combining the strengths of both filters:

$$w_{soft} = w_{rc} \odot w_{ic}. \qquad (5)$$

The multiplication ensures that only pixels deemed reliable by both filters receive substantial weight, thus enforcing a stricter criterion for reliability. The soft filtering weights are then applied to the self-training loss function, yielding the consistency-aware soft-weighted loss:

$$L_{st} = w_{soft} \odot |\hat{P} - P^O|, \qquad (6)$$

where $\hat{P}$ is the prediction of the student model and P^O denotes the pseudo disparity map generated by the teacher model. By employing the loss, our approach leverages more unlabeled training data by incorporating less confident samples in a weighted manner, rather than discarding potentially informative examples as in hard thresholding. It not only retains more valuable data but also mitigates the risk of overfitting to highly confident yet potentially noisy labels, significantly enhancing the model's robustness and overall

Method	KITTI 2015			Middlebury				ETH3D		
	D1-bg	D1-fg	D1-all	bad 1.0	bad 2.0	bad 4.0	Avgerr	bad 1.0	bad 2.0	Avgerr
PSMNet [5]	1.86	4.62	2.32	63.9	42.1	23.5	6.68	-	-	-
RAFT-Stereo [14]	1.75	2.89	1.96	9.37	4.74	2.75	1.27	2.44	0.44	0.18
CREStereo [12]	1.45	2.86	1.69	8.25	3.71	2.04	1.15	**0.98**	0.22	<u>0.13</u>
EAIStereo [31]	1.59	2.92	1.81	7.81	3.68	2.14	1.09	2.31	1.14	0.21
DLNR [32]	1.60	<u>2.59</u>	1.76	6.82	3.20	1.89	1.06	-	-	-
GMStereo [26]	1.49	3.14	1.77	23.6	7.14	2.96	1.31	1.83	0.25	0.19
IGEV-Stereo [24]	1.38	2.67	1.59	9.41	4.83	3.33	2.89	1.12	<u>0.21</u>	0.14
Selective-IGEV [22]	<u>1.33</u>	2.61	<u>1.55</u>	<u>6.53</u>	<u>2.51</u>	<u>1.36</u>	<u>0.91</u>	1.23	0.22	**0.12**
CST-Stereo (Ours)	**1.30**	**2.48**	**1.50**	**6.23**	**2.40**	**1.27**	**0.87**	<u>1.02</u>	**0.17**	**0.12**

Table 1. Quantitative evaluation on KITTI 2015 [17], Middlebury [19] and ETH3D [20] leaderboard. **Bold**: Best. <u>Underline</u>: Second best.

performance. Finally, after several optimization steps for the student, the parameters of the teacher are updated by the exponential moving average (EMA) as follows:

$$\theta_T \leftarrow \lambda\theta_T + (1-\lambda)\theta_S, \tag{7}$$

where θ_T and θ_S denote weights of the teacher and student models respectively.

4. Experiments

4.1. Datasets

Sceneflow [16] is a synthetic dataset with over 39000 dense labeled stereo pairs. **Middlebury 2014** [19] is a real-world high-resolution indoor dataset with 23 training pairs and 10 testing pairs. **KITTI 2015** [17] contains 200 training pairs and 200 testing pairs with sparse disparity maps which is collected for real-world driving scenes. **ETH3D** [20] is a collection of gray-scale stereo pairs with both indoor and outdoor scenes, consisting of 27 training pairs and 20 testing pairs. **Instereo2k** [1] is a real-world indoor dataset with 2050 stereo pairs. **Holopix50k** [9] is a large-scale in-the-wild stereo dataset with nearly 50k unlabeled image pairs.

4.2. Implementation Details

We implement our CST-Stereo with Pytorch framework using NVIDIA A100 GPUs. Before self-training, all the models are pre-trained on the Sceneflow dataset for 200k steps with a learning rate of 2e-4. And the learning rate is set to be 1e-4 during self-training. In the training process, We use AdamW optimizer and the one-cycle learning rate schedule. For the consistency-aware soft filtering module, the ε_1 and ε_2 are set to be 5 and 10 respectively, while τ_1 and τ_2 are 2 and 0.5 respectively. For the EMA, the λ is set to be 0.99 and the decay iteration number is 100.

4.3. Comparisons with State-of-the-art

Following [29], we evaluate the effectiveness of our method by in domain, domain adaptation, and domain generalization performance.
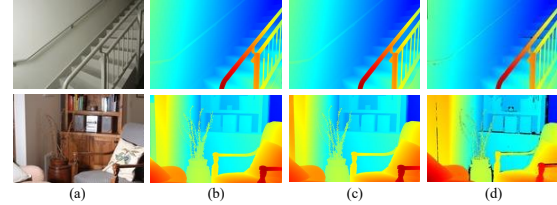

Figure 5. Visualization of our method boosting the in domain performance on the Middlebury datasets. (a) Referenced images. (b) Selective-IGEV [22]. (c) CST-Selective-IGEV. (d) Ground Truth

4.3.1. In Domain Performance

Based on the current SOTA method Selective-IGEV [22], our method makes further improvements on in domain performance by leveraging both the labeled and unlabeled data, as shown in Tab. 1, with visualization results presented in Fig. 5. We conduct extensive experiments following the same training pipeline as Selective-IGEV, which first pre-train a model on the synthetic datasets and fine-tune the model on a mixed dataset comprising both synthetic and real data. Differently, we also leverage unlabeled data of the real-world datasets including unannotated pixels in sparsely labeled data and some unlabeled images. The results are submitted to the Middlebury, KITTI2015, and ETH3D leaderboard and show evident improvements over previous SOTA methods. At the time of writing, our method ranks 3rd on the Middlebury leaderboard. The results indicate that our approach not only matches but further enhances the capabilities of the current SOTA methods.

4.3.2. Domain Adaptation Performance

To further verify the effectiveness of our proposed method, we conducted extensive experiments on kinds of datasets, including Middlebury, KITTI 2015, ETH3D, and Instereo2k, to assess the domain adaptation performance of CST-Stereo. With models pre-trained on the synthetic Sceneflow dataset, our method exclusively leverages the unlabeled data from the target datasets through self-training. As illustrated in Tab. 2, existing stereo matching methods often struggle to perform well when they do not have access to labeled data from the target datasets. Our approach significantly enhances the generalization performance of sev-

Table 2. Generalization evaluation on KITTI 2015 [17], Middlebury [19], ETH3D [20] and Instereo2k dataset.

Method	Middlebury F EPE	Middlebury F D1	Middlebury H EPE	Middlebury H D1	KITTI 2015 EPE	KITTI 2015 D1	ETH3D EPE	ETH3D D1	Instereo2k EPE	Instereo2k bad 1.0	Instereo2k bad 2.0
PSMNet [5]	40.51	57.93	9.79	32.18	4.09	28.64	2.58	66.34	70.40	95.95	93.68
RAFT-Stereo [14]	3.84	15.64	1.44	11.21	1.13	5.68	0.27	2.61	1.01	11.30	5.82
CREStereo [12]	6.59	21.05	1.99	13.37	2.20	7.39	0.77	5.10	1.36	17.94	11.03
EAIStereo [31]	6.16	18.25	2.15	11.74	1.87	6.99	16.30	23.87	1.42	17.79	10.89
DLNR [32]	6.57	14.46	1.45	9.46	3.05	18.91	9.91	23.00	0.99	13.20	6.78
GMStereo [26]	4.10	29.15	1.92	15.68	1.21	5.76	0.42	6.23	1.90	26.12	15.80
IGEV-Stereo [24]	5.87	11.85	1.36	7.21	1.21	6.03	0.33	4.06	1.82	12.11	5.72
Selective-IGEV [22]	5.28	12.07	1.35	7.31	1.25	6.05	0.7	5.27	2.08	17.35	9.37
CST-RAFT-Stereo	3.12	12.60	1.21	10.74	**1.03**	4.93	**0.25**	**2.35**	**0.60**	**7.61**	**3.36**
CST-CREStereo	4.24	17.68	1.96	12.95	1.24	**4.86**	0.36	3.68	1.34	15.08	8.63
CST-Selective-IGEV	**1.32**	**7.75**	**0.77**	**5.37**	1.05	4.91	0.27	3.54	0.84	10.16	5.40
CST-RAFT-Stereo†	**3.78**	14.62	**1.39**	10.59	**1.05**	5.23	**0.28**	**2.40**	**0.83**	**10.70**	**5.28**
CST-CREStereo†	4.81	20.53	2.01	13.15	1.10	**5.12**	0.30	3.38	1.08	16.65	9.25
CST-Selective-IGEV†	4.11	**9.86**	1.48	**6.47**	1.11	5.21	0.32	4.12	0.94	14.62	6.12

Table 2. Generalization evaluation on KITTI 2015 [17], Middlebury [19], ETH3D [20] and Instereo2k dataset. **Bold**: Best. †: Without using any data from the target dataset. CST-RAFT, CST-CRE and CST-Selective represent CST-Stereo with RAFT-Stereo, CREStereo and Selective-IGEV as the baselines respectively.
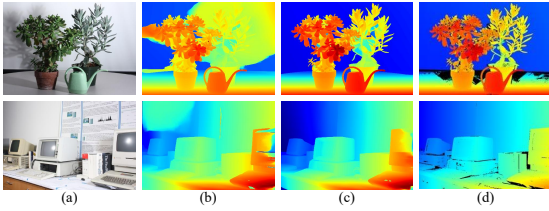

Figure 6. Visualizations for the domain adaptation performance on the Middlebury datasets. (a) Referenced images. (b) Selective-IGEV [22]. (c) CST-Selective-IGEV. (d) Ground Truth
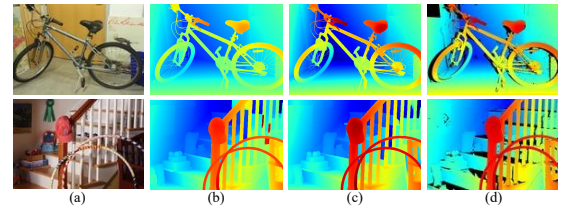

Figure 7. Visualizations for the domain generalization performance on the Middlebury datasets. (a) Referenced images. (b) Selective-IGEV [22]. (c) CST-Selective-IGEV. (d) Ground Truth

eral iterative-based stereo matching methods across multiple datasets without using any labeled data from the target real-world dataset. Visualization results in Fig. 6 also demonstrate the effectiveness our method.

4.3.3. Domain Generalization Performance

We also evaluate the generalization ability to unseen real-world data in Tab. 2. Unlike the previous setups, we did not utilize any data from the target dataset, including labeled and unlabeled. Instead, we applied self-training on unlabeled data from an independent dataset, Holopix50k. Remarkably, our method still managed to improve performance on the target dataset, underscoring its effectiveness in generalizing to unseen domains without direct access to the target data. Visualization results are presented in Fig. 7, showing our effectiveness for domain generalization task.

4.4. Ablation Study

In this section, we ablate our method on the Middlebury evaluation set. All the models are pre-trained on the Sceneflow dataset, with only unlabeled data from the Middlebury dataset used in the self-training process.

Model	Self Training	Filter	Soft Design	EPE	D1
Baseline				5.28	12.07
ST	√			6.15	11.76
ST+Filter	√	√		2.85	9.17
Full Model	√	√	√	**1.32**	**7.75**

Table 3. Ablation studies for soft filtering module on Middlebury Eval dataset based on Selective-IGEV.

4.4.1. Ablations for consistency-aware soft filtering

To demonstrate the effectiveness of our proposed CSF, we conduct experiments by removing the filtering step and allowing all pseudo-labels to participate in the training process, as shown in Tab. 3. Results show that without filtering exhibits a 48% decrease in EPE compared to the result with filtering, and even shows a performance drop of 22% compared to the baseline. This degradation underscores the importance of the soft filtering module. Besides, We also ablate our soft mechanism. With hard filtering, pseudo-labels are selected based on a binary decision: if the reliability score exceeds a fixed threshold, the sample is selected; otherwise, it is discarded. Conversely, our soft filtering module applies a weighted selection mechanism, where samples

Filter	EPE	D1
Baseline	5.28	12.07
DDE[23, 29]	7.47	15.44
LRC[29]	4.19	13.11
MRPCF (Ours)	1.58	8.52
IPCF (Ours)	1.70	9.18
MRPCF + IPCF (Ours)	**1.32**	**7.75**

Table 4. Ablation studies for filters on Middlebury Eval dataset.

Decay	1	10	100	1000	No decay
EPE	4.15	1.51	**1.32**	2.88	3.46
D1	9.69	7.93	**7.75**	9.13	10.54

Table 5. Ablation studies for EMA on Middlebury Eval dataset.

with higher reliability scores contribute more to the training process. Results show that soft filtering obtains further improvements compared with hard filtering, demonstrating the superiority of our soft filtering method.

4.4.2. Ablations for filters

To evaluate the effectiveness of the proposed multi-resolution prediction consistency filter (MRPCF) and iterative prediction consistency filter (IPCF), we ablate the two filters respectively based on Selective-IGEV. As shown in Tab. 4, the inclusion of each filter independently leads to noticeable improvements in model performance, demonstrating their effectiveness. Furthermore, when both filters are combined, the model achieves the best performance, indicating that the two filters complement each other to enhance the self-training process. Besides, we compared our filtering mechanisms with those proposed in previous works [23, 29]. LRC is a filter used to remove occluded regions, and DDE is a filter that measures confidence based on the probability distribution of the cost volume. The results reveal that the previous filtering techniques are less effective for iterative-based methods, highlighting the necessity of our tailored filtering strategies for optimal performance.

4.4.3. Ablations for EMA

We also ablate the impact of EMA based on Selective-IGEV as shown in Tab. 5. The results indicate that incorporating EMA significantly improves performance. Besides, as we varied the number of delay steps for parameter updates, we observed an interesting trend: the performance initially improves as the delay increases, but eventually declines when the delay becomes too large. This phenomenon can be attributed to the stability of gradient updates being transferred to the teacher model. With too few delay steps, the updated weights may become unstable, leading to suboptimal teacher updates. Conversely, with too large a decay, the teacher's improvement becomes sluggish, failing to keep pace with the student model's learning.

Method	Middlebury			
	F		H	
Selective-IGEV	5.28	12.07	1.35	7.31
Ada-Selective-IGEV	3.65	11.44	1.56	6.59
ZOLE-Selective-IGEV	5.65	11.87	1.16	6.76
CST-Selective-IGEV	**1.32**	**7.75**	**0.77**	**5.37**
Ada-Selective-IGEV[†]	4.20	12.58	1.84	7.51
ZOLE-Selective-IGEV[†]	6.11	13.05	1.30	6.98
CST-Selective-IGEV[†]	**4.11**	**9.86**	**1.48**	**6.47**

Table 6. Comparision with several cross domain methods in domain adaptaion and domain generalization on Middlebury dataset.

4.5. Discussion

Although some unsupervised cross domain methods also enhances the model performance on out-of-domain data, we emphasize that our approach fundamentally differs from existing domain adaptation and domain generalization methods. First, while many of these methods primarily aim to enhance out-of-domain performance, they often provide limited insights into improvements in in-domain performance. In contrast, we propose a more universal strategy for leveraging unlabeled data that not only enhances out-of-domain capabilities—including domain adaptation and domain generalization—but also effectively boosts in-domain performance through the strategic use of unlabeled samples. Second, the relationship between these methods and ours is orthogonal rather than mutually exclusive. This orthogonality allows for the potential integration and exploration of our self-training approach alongside existing techniques, thereby creating opportunities for more comprehensive frameworks that can fully capitalize on unlabeled data in various contexts. Furthermore, we demonstrate a comparison between our method and some unsupervised domain adaptation techniques in Tab. 6 The results indicate that our approach exhibits more effective improvements, especially in terms of domain generalization.

5. Conclusion

In this paper, we propose CST-Stereo, a novel self-training method for iterative-based stereo matching models by effectively leveraging the unlabeled data. In detail, we introduce a consistency-aware soft filtering module that can dynamically adjust the reliability of pseudo-labels to mitigate the risks associated with incorrect labels. Extensive experimental results validate the efficacy of our method, showcasing substantial improvements over current SOTA methods on multiple benchmark datasets. Although our proposed method effectively mitigates the negative impact of incorrect pseudo-labels, the performance of the model after self-training remains constrained by the original model's capabilities and architecture. In the future, we will focus on developing more robust and adaptive model architectures.

Acknowledgments

References

[1] Wei Bao, Wen Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. Instereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63, 2020. 6

[2] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239, 2001. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[4] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *AAAI Conference on Artificial Intelligence*, 2020. 3

[5] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, pages 5410–5418, 2018. 2, 6, 7

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 2

[7] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019. 2

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[9] Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. Holopix50k: A large-scale in-the-wild stereo image dataset. In *CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Seattle, WA, 2020.*, 2020. 6

[10] Andreas Klaus, Mario Sormann, and Konrad Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *18th International Conference on Pattern Recognition (ICPR'06)*, pages 15–18. IEEE, 2006. 2

[11] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. 2013. 3

[12] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *CVPR*, pages 16263–16272, 2022. 1, 2, 6, 7

[13] Yandong Li, Di Huang, Danfeng Qin, Liqiang Wang, and Boqing Gong. Improving object detection with selective self-supervised self-training. *ArXiv*, abs/2007.09162, 2020. 3

[14] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *3DV*, pages 218–227. IEEE, 2021. 1, 2, 6, 7

[15] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, 2022. 2

[16] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016. 6

[17] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, pages 3061–3070, 2015. 6, 7

[18] D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, pages 131–140, 2001. 1

[19] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, pages 31–42. Springer, 2014. 6, 7

[20] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, pages 3260–3269, 2017. 6, 7

[21] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14362–14372, 2021. 2

[22] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19701–19710, 2024. 1, 2, 6, 7

[23] Fudong Xu, Lin Wang, and Huibin Li. A unified and efficient semi-supervised learning framework for stereo matching. *Pattern Recognit.*, 147:110128, 2024. 1, 3, 8

[24] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023. 2, 6, 7

[25] Gangwei Xu, Yun Wang, Junda Cheng, Jinhui Tang, and Xin Yang. Accurate and efficient stereo matching via attention concatenation volume. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2461–2474, 2024. 2

[26] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 6, 7

[27] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-trainingwork better for semi-supervised semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4258–4267, 2022. 3

[28] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *ArXiv*, abs/2401.10891, 2024. 3

[29] Xin Yue, Zongqing Lu, Xiangru Lin, Wenjia Ren, Zhijing Shao, Haonan Hu, Yu Zhang, and Qingmin Liao. Semi-stereo: A universal stereo matching framework for imperfect data via semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 646–655, 2024. 1, 3, 6, 8

[30] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1592–1599, 2015. 2

[31] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Yong Zhao, Yitong Yang, and Ting Ouyang. Eai-stereo: Error aware iterative network for stereo matching. In *Proceedings of the Asian Conference on Computer Vision*, pages 315–332, 2022. 2, 6, 7

[32] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao. High-frequency stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1327–1336, 2023. 2, 6, 7