applied sciences

MDPI

Article

Matchability and Uncertainty-Aware Iterative Disparity Refinement for Stereo Matching

Junwei Wang, Wei Zhou, Yujun Tang and Hanming Guo *

Engineering Research Center of Optical Instrument and System, Ministry of Education, Shanghai Key Lab of Modern Optical System, School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, 516 Jungong Rd, Shanghai 200093, China; 191380032@st.usst.edu.cn (J.W.); wade_weizhou@outlook.com (W.Z.); yujtang@usst.edu.cn (Y.T.)
* Correspondence: hmguo@usst.edu.cn

Abstract: After significant progress in stereo matching, the pursuit of robust and efficient ill-posed-region disparity refinement methods remains challenging. To further improve the performance of disparity refinement, in this paper, we propose the matchability and uncertainty-aware iterative disparity refinement neural network. Firstly, a new matchability and uncertainty decoder (MUD) is proposed to decode the matchability mask and disparity uncertainties, which are used to evaluate the reliability of feature matching and estimated disparity, thereby reducing the susceptibility to mismatched pixels. Then, based on the proposed MUD, we present two modules: the uncertainty-preferred disparity field initialization (UFI) and the masked hidden state global aggregation (MGA) modules. In the UFI, a multi-disparity window scan-and-select method is employed to provide a further initialized disparity field and more accurate initial disparity. In the MGA, the adaptive masked disparity field hidden state is globally aggregated to extend the propagation range per iteration, improving the refinement efficiency. Finally, the experimental results on public datasets show that the proposed model achieves a reduction up to 17.9% in disparity average error and 16.9% in occluded outlier proportion, respectively, demonstrating its more practical handling of ill-posed regions.

Keywords: stereo matching; disparity refinement; convolutional neural networks; deep learning

Citation: Wang, J.; Zhou, W.; Tang, Y.; Guo, H. Matchability and Uncertainty-Aware Iterative Disparity Refinement for Stereo Matching. *Appl. Sci.* **2024**, *14*, 8457. https://doi.org/10.3390/app14188457

Academic Editor: Pedro Couto

Received: 30 April 2024
Revised: 10 September 2024
Accepted: 18 September 2024
Published: 19 September 2024

1. Introduction

Stereo matching is one of the most important topics in the field of photogrammetry and computer vision, with broad applications in 3D scene reconstruction, augmented reality, and automated driving. In stereo matching, the pixel disparities are obtained by dense pixel-wise matching between two rectified images along the epipolar line. Traditional stereo matching methods [1,2] have been developed over time and form a four-step framework [3]: cost computation, cost aggregation, disparity computation and disparity refinement. Most of the deep learning-based stereo matching methods [4–7], which have rapidly evolved in recent years, also adhere to this framework [8]. Disparity refinement, as the final step of stereo matching, is mainly used to improve the disparity accuracy in matched regions, handle the mismatched pixels, and determine the model's final output disparity.

Disparity refinement methods for stereo matching have been the subject of extensive and sustained research [9–11]. However, it remains challenging to explore refinement methods capable of identifying and handling ill-posed regions. These regions, present between the reference and target images, fail to meet the one-to-one pixel matching conditions due to various factors such as occlusion, poor exposure, and texturelessness. Traditional disparity refinement for stereo matching is usually based on the assumption of local color-disparity consistency [3], using the pre-calculated disparity from other sources [9,11] with 2D image texture or 3D spatial geometric guidance to detect and handle the mismatched pixels. With the advancements of stereo-matching neural networks, this processing has evolved into end-to-end implementations. Recent studies [12–14] combine implicit cost aggregation with

disparity refinement, enabling iterative disparity refinement starting from zero without reliance on pre-computed disparity. These approaches not only achieve high accuracy, but also offer flexibility in balancing output delay and accuracy by simply adjusting the number of iterations. However, existing iteration-based methods lack an explicit construction of the matching cost volume and are difficult to follow [15,16], leading to susceptibility to mismatched regions.

To mitigate the above problems without explicitly modeling the matching cost volume, based on the RAFT-Stereo [12], we propose a new model that can estimate the feature matchability and the disparity uncertainty concurrently. This model is capable of discerning mismatched pixels for more appropriate processing without additional explicit supervision, thus enhancing the model's robustness. In this paper, pixel matchability is defined as an appraisal of the reliability in feature matching, regarded as an intrinsic attribute of input image pixel pairs. Disparity uncertainty, on the other hand, is defined as an assessment of the predicted disparity's reliability, capturing the margin of error in disparity estimation [17]. Uncertainty can be utilized to filter out disparities that exceed a specified error threshold. In regions of low matchability, pixels that are properly refined exhibit lower uncertainty, whereas those lacking reliable matches in their vicinity show higher uncertainty. The model's susceptibility to mismatched pixels can be reduced by incorporating an awareness of feature matchability and disparity uncertainty.

The proposed model, termed the matchability and uncertainty-aware iterative disparity refinement (MUIR) neural network, integrates a novel matchability mask and disparity uncertainty decoder (MUD) into the iterative disparity refinement framework. The MUD is trained to decode the feature matchability mask and disparity uncertainty jointly from the disparity field hidden state, enhancing the model's ability to detect and handle ill-posed regions. Following the introduction of the MUD module, we further present the uncertainty-preferred disparity field initialization (UFI) and the masked hidden state global aggregation (MGA) modules. The UFI module employs a multi-disparity window scan-and-select method to expand the perceptual range of the initial disparity field, and improve the initial disparity accuracy. The MGA module propagates the adaptive masked reliable hidden state globally, thus introducing long-range dependencies to enhance the disparity refinement efficiency. The advantages of proposed MUIR are as follows:

1. The proposed MUIR requires no additional mask supervision or dedicated training for uncertainty prediction. It can jointly predict disparity, matchability mask, and disparity uncertainty using only the disparity ground truth for single-stage joint training, without compromising disparity accuracy.
2. The MUD module is integrated into the iterative framework to decode the matchability mask and disparity uncertainty from hidden state at any iteration, ensuring the model's scaling flexibility.
3. The model with the UFI and MGA modules achieves more efficient disparity refinement per iteration, which is more important for refinement in regions with large disparity.

2. Related Works

In the task of stereo matching for rectified image pairs, deep-learning-based methods have been intensively studied since the MC-CNN proposed by Zbontar and LeCun [18]. Most of the end-to-end methods [4–7,19–22], that focus on improving the cost aggregation process, can be considered as neural network approximations and improvements of the classical methods [1]. Among these methods, most of them have outperformed the classical ones [1,2] on the commonly used datasets [23–26]. There are also works [7,27,28] that attempt to exploit lightweight feature-guided disparity refinement to achieve higher accuracies. Recently, methods [12,14] based on the recurrent all-pairs transforms [29] in optical flow task have brought disparity refinement to a higher plateau.

2.1. Disparity Refinement Neural Networks Based on the RAFT

The RAFT-Stereo [12] is a variant of RAFT [29] for stereo matching using only 2D convolutional neural networks (2D CNNs), where the sampled slices of the image correlation volume and the image contextual feature are fused to iteratively update the disparity field's hidden state via a multi-scale iterator constructed with ConvGRUs [30]. This iterative framework can be considered as a combination of implicit cost aggregation with a monocular depth estimation pipeline, which avoids the explicit modeling and aggregation of the matching cost volumes, thus achieving higher disparity refinement efficiency and generalization performance compared to 3D CNN-based methods [4,20]. In addition, this RNN-like framework is theoretically capable of predicting arbitrary disparities within the co-visible regions of image pairs. Recently, some RAFT-based works [13,14,16,31,32] have been reported with further improvements. Incorporating the vision transformer techniques [28,33], information interaction between feature map pairs is introduced [13,16] before the iterative refinement process to augment the matching features, thus improving the accuracy of the initial disparity, optical flow or the handling of occluded regions. On the other hand, a lightweight cost aggregation using 3D CNNs is reintroduced by [32], where the cost volume with limited disparity range is used as an additional look-up table as well as to compute more accurate initial disparity, thus improving the geometric awareness of the model. Furthermore, [14,31] achieve gains in generalization and accuracy by replacing the iterator with ConvLSTMs [14] and using additional disparity adjustments after upsampling. In contrast, the MUIR proposed in this paper is trained to evaluate the reliability of feature matching and estimated disparity, which enhances the detection and handling of ill-posed regions, thus improving accuracy and efficiency. Our model takes RAFT-Stereo [12] as the baseline to ensure that the improvements can be applied to most RAFT-based models.

2.2. Matchability and Uncertainty

Some works train the occlusion prediction subnetwork using occlusion labels as explicit supervision [28]. While this approach easily models occluded pixels, it still fails to model mismatched pixels due to abnormal exposures, texturelessness, etc., and is difficult to train on sparse stereo datasets such as KITTI [23,24], which lack occlusion labels. Differently, the training of the proposed MUIR does not rely on occlusion labels. The predicted matchability mask models most of the mismatched pixels and is a by-product of the supervised training using only disparity ground truth.

Confidence or uncertainty is used to assess disparity reliability in learning-based methods. In confidence-based methods, the disparity confidence is predicted with patch-wise [34–36] or image-wise [37,38] input to a CNN-based subnetwork trained with binary cross-entropy loss (BCE loss). Unlike confidence $\in [0, 1]$, uncertainty $\in [0, +\infty]$, based on [17], is used in multiple regression tasks [39,40] to model the error in the predicted values. In [41], uncertainty-guided refinement is performed after cost aggregation to improve disparity accuracy. The CVA-Net by [42] uses matching cost volume to predict aleatoric uncertainty. It is then combined with modified GC-Net [20,43] to improve disparity accuracy. Ref. [44] exploits the intermediate multi-scale disparity maps from [5] to predict the disparity uncertainty. In addition, the KL divergence loss is proposed to match the distribution of uncertainty with disparity error, thus improving the uncertainty accuracy. Differently, the proposed MUIR prioritizes the accuracy of disparity over uncertainty. It aims to improve the robustness by enabling the model to become aware of the uncertainty. Furthermore, the disparity uncertainty of each iteration is predictable, which ensures the scaling flexibility of the model.

3. Methods

The architecture of the MUIR proposed in this paper is shown in Figure 1, which employs the recurrent disparity refinement framework based on the RAFT-Stereo [12]. The stereo image pairs $I_{\mathrm{l}}, I_{\mathrm{r}} \in [0, 255]^{3\times h^{\mathrm{up}}\times w^{\mathrm{up}}}$ are scaled to $[-1, 1]$ and then fed to the matching

feature Extractor$^{\text{mtc}}$ and the contextual feature Extractor$^{\text{ctx}}$ to encode the matching features $F_{\text{l}}^{\text{mtc}}, F_{\text{r}}^{\text{mtc}} \in \mathbb{R}^{c^{\text{mtc}} \times h \times w}$ and the contextual feature map $F^{\text{ctx}} \in \mathbb{R}^{c^{\text{ctx}} \times h \times w}$, where $c^{\text{mtc}}, c^{\text{ctx}}$ are the number of channels for the matching features and the contextual feature, respectively, and $h^{\text{up}}, w^{\text{up}}$ and h, w are the size of the input images and feature maps, respectively. The measure of visual similarity between matching features $F_{\text{l}}^{\text{mtc}}$ and $F_{\text{r}}^{\text{mtc}}$ is obtained using dot product to construct the matching correlation volume:

$$C^{\text{mtc}} = \frac{1}{\sqrt{c^{\text{mtc}}}} \left\langle F_{\text{l}}^{\text{mtc}}(j), F_{\text{r}}^{\text{mtc}}(j-d) \right\rangle \in \mathbb{R}^{h \times w \times w}, \tag{1}$$

where d is disparity, $j \in \{1, 2, \ldots, w\}$ is the horizontal index of a matching feature, and $\langle \cdot, \cdot \rangle$ denotes the inner product. The correlation volume slice $C^{\text{mtc}}(\hat{D}_t) \in \mathbb{R}^{(2r+1) \times h \times w}$ is obtained by bilinear sampling along the scanline dimension (last dimension) of C^{mtc} with a sampling window $\hat{D}_t = \left\{ d_t \mid d_t \in \left[\hat{d}_t - r, \hat{d}_t + r \right] \right\}$ centered on the estimated disparity $\hat{d}_t$, where $t \in \{1, 2, \ldots, T\}$ is the index of current iteration, T is the total number of refinement iterations, and $r \in \mathbb{Z}_+$ is the sampling window radius.

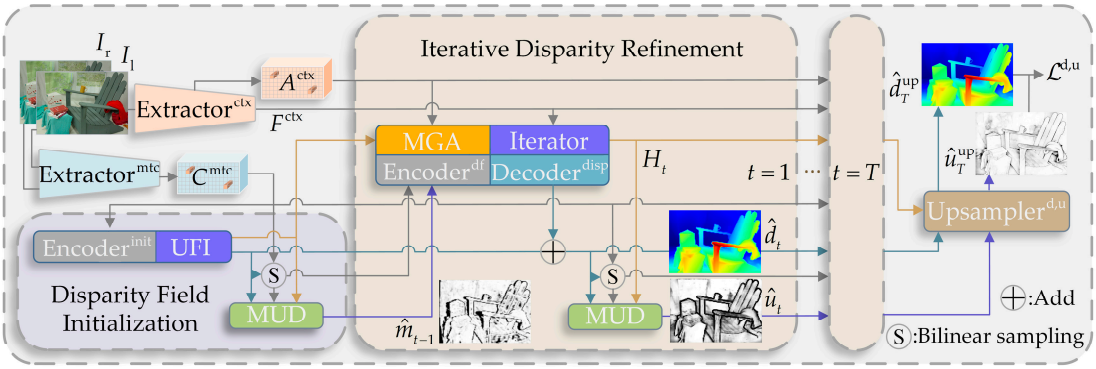


Figure 1. The architecture of the proposed matchability and uncertainty-aware iterative disparity refinement (MUIR) neural network, including feature extraction, disparity field initialization, iterative disparity refinement and upsampling.

During initialization, the initial disparity field's hidden state H_1 and disparity $\hat{d}_1$ are obtained through the simplified disparity field feature encoder Encoder$^{\text{init}}$ and the Initializer in the UFI module (see Section 3.2.1). In each iteration, the matchability mask $\hat{m}_{t-1} \in [0, 1]^{h \times w}$ is decoded by the MUD module from the hidden state (see Section 3.1.1). Next, the $\hat{m}_{t-1}$ is fed to the disparity field feature Encoder$^{\text{df}}$ and MGA modules to filter out the matching ambiguity noise and the unreliable hidden state (see Section 3.2.2), respectively. The hidden state H_t updated by the Iterator is fed to the residual Decoder$^{\text{disp}}$ to refine the estimated disparity: $\hat{d}_t = \hat{d}_{t-1} + \Delta\hat{d}_t$. The H_t and $\hat{d}_t$ are then utilized by the MUD to decode the next mask $\hat{m}_t$ and uncertainty $\hat{u}_t$. At the end of each iteration, the H_t, $\hat{d}_t$ and $\hat{u}_t$ can optionally be fed to the joint upsampling block Upsampler$^{\text{d,u}}$ to obtain the upsampled disparity $\hat{d}_t^{\text{up}}$ and uncertainty $\hat{u}_t^{\text{up}}$. The proposed iterative disparity refinement is completed after T iterations of the above process.

3.1. The Matchability and Uncertainty Decoder

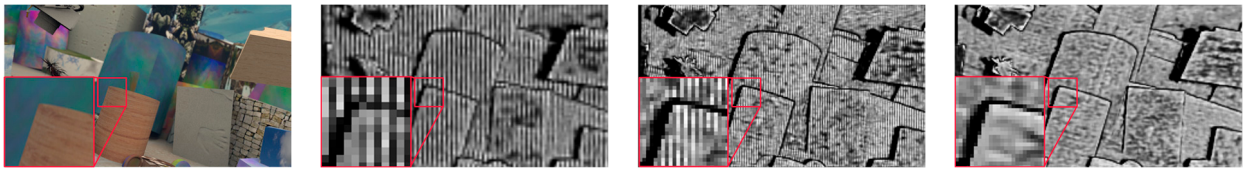
To reduce the susceptibility of the model to mismatched pixels, we propose to integrate a new MUD module into the iterative disparity refinement framework, which is trained to decode the matchability mask $\hat{m}_t$ and the disparity uncertainty $\hat{u}_t$ from the disparity field hidden state, thus enabling the model to be aware of the reliability of feature matching and estimated disparity. Notably, our method requires neither additional mask supervision nor additional dedicated training for the uncertainty prediction module.

3.1.1. Prediction of Matchability Mask

We observed that in RAFT-based models, even when apparent matching ambiguity noise is carried in the correlation volume slice sampled from ill-posed regions, the accuracy of estimated disparity is only slightly affected. This implies that the model may have an implicit matchability awareness. To verify this insight, the disparity field feature encoding in the baseline model [12] is modified as:

$$\begin{aligned} \hat{m}_t &= \sigma(f^{\mathrm{m}}_{\theta^{\mathrm{m}}}(H_t, C^{\mathrm{mtc}}(\hat{D}_t), \hat{d}_t)) \\ F^{\mathrm{corr}}_t &= f^{\mathrm{corr}}_{\theta^{\mathrm{corr}}}(C^{\mathrm{mtc}}(\hat{D}_t), \hat{m}_t) = \mathrm{Conv}(\llbracket C^{\mathrm{mtc}}(\hat{D}_t), C^{\mathrm{mtc}}(\hat{D}_t) \odot \hat{m}_t, \hat{m}_t \rrbracket; \theta^{\mathrm{corr}})' \end{aligned} \tag{2}$$

where $f^{\mathrm{m}}_{\theta^{\mathrm{m}}}$ and $f^{\mathrm{corr}}_{\theta^{\mathrm{corr}}}$ denote convolutional networks with parameters θ^{m} and θ^{corr}, respectively, $\llbracket \cdot \rrbracket$ denotes the concatenation operation, $\odot$ denotes element-wise multiplication and $\sigma(\cdot)$ denotes the sigmoid function. The volume slice $C^{\mathrm{mtc}}(\hat{D}_t)$ is multiplied by matchability mask $\hat{m}_t$ to filter the matching ambiguity noise. From preliminary pre-training experiments we find that the modified model is enabled to decode the matchability mask from the disparity field's hidden state in each iteration, as shown in Figure 2b,c.

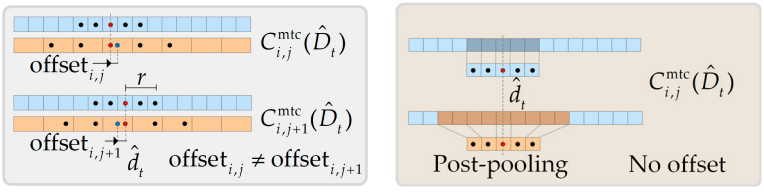


(**a**) (**b**) (**c**) (**d**)

Figure 2. The matchability masks decoded from disparity field's hidden state. (**a**) The input left image; (**b**,**c**) the matchability mask from the models run at 1/8 and 1/4 resolution, respectively; (**d**) the matchability mask from the model with our modified pooling method runs at 1/4 resolution. The details are visible in the zoomed-in red box.

As depicted in Figure 2b,c, the obtained matchability mask identifies ill-posed regions. However, it exhibits distinct strip-like textures with a periodicity that correlates with the resolution of the image.

Specifically, as illustrated in Figure 3a, the offset varies between the disparity $\hat{d}_t$ and the center of the multi-scale pooling (illustrated as blue dots) within the correlation volume C^{mtc} along the scanline direction. For two adjacent points $p_{i,j}$ and $p_{i,j+1}$ with the same disparity $\hat{d}_t$, their offsets are denoted as $\mathrm{offset}_{i,j} < 0$ and $\mathrm{offset}_{i,j+1} > 0$, respectively. This offset is related to the horizontal pixel coordinate j and the disparity $\hat{d}_t$. Consistent with our experimental observations, this varying offset introduces a periodic modulation of the disparity field F^{df}_t, causing the model to misinterpret certain horizontal features as unmatchable. This results in strip-like artifacts appearing in the matchability mask. Although these artifacts are gradually attenuated during training, the risk of decreasing accuracy is not negligible.



(**a**) (**b**)

Figure 3. Illustrations for correlation volume pooling methods. (**a**) The pooling method used in the baseline [12]. (**b**) The proposed post-pooling method. The pooling center is kept at $\hat{d}_t$.

It is challenging to verify this hypothesis directly through network parameters, our proposed post-pooling method, which centers on $\hat{d}_t$ (as shown in Figure 3b), effectively

eliminates the offset. As a result, the artifacts in the matchability mask are removed, as evidenced in Figure 2d. This experimental outcome provides indirect confirmation of our hypothesis regarding the cause of the strip-like artifacts.

The above preliminary experiments demonstrate the latent implicit matchability awareness in the baseline can be extracted in the form of matchability mask. Notably, the mask is a byproduct obtained from the training of the disparity refinement task, without any additional explicit supervision. We randomly binarize the mask like a drop block [45] to avoid the $C^{\mathrm{mtc}}(\hat{D}_t)$ being unnecessary numerically scaled by the mask $\hat{m}_t$:

$$\hat{m}_t^{\mathrm{bnr}} = \begin{cases} \hat{m}_t & \text{if } \tau^{\mathrm{bnr}} > 0.5 \\ [\hat{m}_t] & \text{otherwise} \end{cases}, \tag{3}$$

where $\tau^{\mathrm{bnr}} \sim B(1, 0.5)$ and $[\cdot]$ denotes binary rounding operation.

3.1.2. Prediction of Disparity Uncertainty

The disparity estimation in ill-posed regions relies heavily on geometric priors rather than feature matching, limiting the accuracy. The more stable and practical way for downstream tasks is to filter out less reliable predictions. However, uncertainty or confidence estimation that rely on cost volume [19,43] is not directly applicable to RAFT-based methods. Therefore, we follow [17,39] in capturing aleatoric uncertainty to indicate disparity error. Unlike methods that use a dedicated uncertainty subnetwork, our method integrates the uncertainty decoder into the iterative disparity refinement framework for joint training.

In stereo matching, the disparity error is assumed to be Laplacian-distributed [41]. The probability density and the joint-training negative maximum likelihood loss function are as follows:

$$\begin{aligned} \mathrm{p}(\hat{d}|I_{\mathrm{l}}, I_{\mathrm{r}}; \theta) &= \tfrac{1}{2\beta} e^{-\frac{\|\hat{d}-d_{\mathrm{gt}}\|_1}{\beta}} \\ \mathcal{L}_1^{\mathrm{d,u}} = -\log \mathrm{p}(\hat{d}|I_{\mathrm{l}}, I_{\mathrm{r}}; \theta) &= \log 2 + \log \beta + \tfrac{\|\hat{d}-d_{\mathrm{gt}}\|_1}{\beta} \end{aligned}, \tag{4}$$

where d_{gt} is disparity ground truth, β is the scale parameter of Laplacian distribution, θ denotes the model parameters, and $\|\cdot\|_1$ denotes the L1 norm. To ensure numerical stability during training [17], the uncertainty is defined as $\hat{u} = \log \beta^2$, and the constant term is omitted:

$$\mathcal{L}_1^{\mathrm{d,u}} = \frac{\hat{u}}{2} + e^{-\frac{\hat{u}}{2}} \left\| \hat{d} - d_{\mathrm{gt}} \right\|_1. \tag{5}$$

Equation (5) can be simply applied to the loss in the baseline [12]:

$$\mathcal{L}_2^{\mathrm{d,u}} = \sum_{t=1}^{T} \gamma^{T-t} \left(\frac{\hat{u}_t}{2} + e^{-\frac{\hat{u}_t}{2}} \left\| \hat{d}_t - d_{\mathrm{gt}} \right\|_1 \right), \tag{6}$$

where γ is a weighting parameter (usually set to 0.9). Preliminary experiments are performed by adding the uncertainty prediction network to the baseline as the loss function replaced by Equation (6). The endpoint error (EPE) curves in Figure 4 show the decrease in disparity accuracy due to the simple design of the joint training.

To avoid an unnecessary decrease in disparity accuracy, the uncertainty is constrained as $\hat{u}_t \in [u_{\min}, u_{\max}]$, resulting in a modified joint-training loss function:

$$\begin{aligned} \mathcal{L}^{\mathrm{u}} &= \tfrac{\hat{u}}{2} + e^{-\frac{\hat{u}}{2}} \min(\max(\left\| \hat{d}_t - d_{\mathrm{gt}} \right\|_1, \varepsilon_{\min}), \varepsilon_{\max})) \\ \mathcal{L}^{\mathrm{d,u}} &= \sum_{t=1}^{T} \gamma^{T-t} (\left\| \hat{d}_t - d_{\mathrm{gt}} \right\|_1 + \mathcal{L}^{\mathrm{u}}) \end{aligned}, \tag{7}$$

where $[\varepsilon_{\min}, \varepsilon_{\max}]$ denotes the range of disparity errors perceived by the model, corresponding to the effective uncertainty range $[u_{\min} = \log \varepsilon_{\min}^2, u_{\max} = \log \varepsilon_{\max}^2]$. The range

constraint is used to avoid model degradation caused by errors that are too small or too large [17]. The constrained uncertainty is computed as:

$$\begin{aligned} \hat{v}_t &= f^{\mathrm{u}}_{\theta^{\mathrm{u}}}(H_t, C^{\mathrm{mtc}}(\hat{D}_t), \hat{d}_t) \\ \hat{u}_t &= \log(\varepsilon^2_{\min} + (\varepsilon^2_{\max} - \varepsilon^2_{\min})\sigma(\hat{v}_t)) \end{aligned}, \tag{8}$$

where $f^{\mathrm{u}}_{\theta^{\mathrm{u}}}$ denotes a simple convolutional network with parameter θ^{u}. In addition, to keep the pixel-wise correspondence between disparity and uncertainty after upsampling, the $\hat{d}^{\mathrm{up}}_t$ and $\hat{u}^{\mathrm{up}}_t$ are computed together via the joint disparity–uncertainty upsampler. It is shown in Figure 4 that the proposed method greatly alleviates the decrease in disparity accuracy introduced by the joint training.

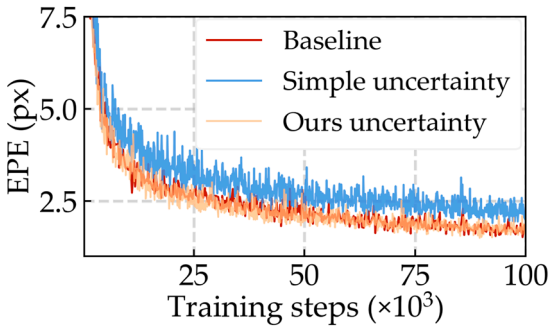


Figure 4. Disparity error curve during training. "EPE" is short for endpoint error.

The above matchability mask decoding network $f^{\mathrm{m}}_{\theta^{\mathrm{m}}}$ and uncertainty decoding network $f^{\mathrm{u}}_{\theta^{\mathrm{u}}}$ are merged, shown in Figure 5, to construct the proposed MUD module:

$$(\hat{m}_t, \hat{u}_t) = \mathrm{MUD}(H_t, C^{\mathrm{mtc}}(\hat{D}_t), \hat{d}_t; \theta^{\mathrm{m}}, \theta^{\mathrm{u}}) \tag{9}$$

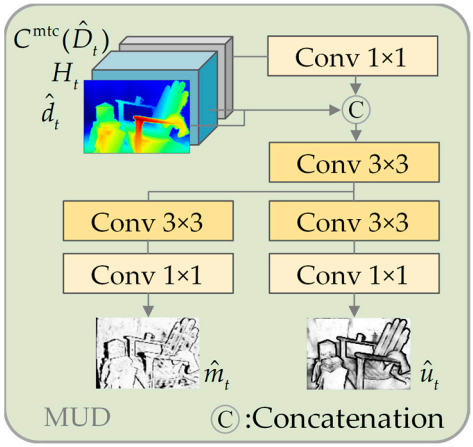


Figure 5. The structure of the matchability and uncertainty decoder (MUD). The disparity field's hidden state is decoded via several convolution layers to obtain matchability mask $\hat{m}_t$ and disparity uncertainty $\hat{u}_t$.

3.2. The Disparity Field Initialization and Aggregation Build upon the MUD

The baseline is restricted by the range-limited disparity field initialization and the slowly increasing range of aggregation, which particularly hinders disparity refinement in regions with large parallaxes (requiring too many iterations). To alleviate the above problem, based on the proposed MUD module, we further propose the uncertainty-preferred disparity field initialization (UFI) and the masked hidden state global aggregation (MGA) modules.

3.2.1. The UFI Module

In the RAFT-based methods, [16,32] use additional processes to obtain the initial disparity, while [12–14,31] use the first iteration to obtain the initial disparity. The initial disparity of these methods is limited by the radius of the disparity window or the number of disparities in the cost volume, and they are unable to model regions with disparities exceeding the limit. In contrast, the proposed UFI module uses multi-disparity proposals $d_k^{\text{init}} \in [0, d_1^{\text{init}}, d_2^{\text{init}}, d_3^{\text{init}}]$ to initialize the disparity field, where d_k^{init} denotes the top three candidates with highest correlation score. They are parallel encoded by $\text{Encoder}^{\text{init}}$ to obtain the disparity field features F_k^{df}, which are then fed to the Initializer to initialize the hidden state:

$$\begin{aligned} q_k &= \tau(f_{\theta^{\text{q}}}^{\text{q}}(F_k^{\text{df}}, F^{\text{ctx}})) \\ z_k &= \sigma(f_{\theta^{\text{z}}}^{\text{z}}(H_{k-1}^{\text{init}}, q_k)) \\ H_k^{\text{init}} &= H_{k-1}^{\text{init}} + (q_k - H_{k-1}^{\text{init}}) \odot z_k \end{aligned}, \tag{10}$$

where $\tau(\cdot)$ denotes the tanh function, $f_{\theta^{\text{q}}}^{\text{q}}$ and $f_{\theta^{\text{z}}}^{\text{z}}$ are the reset gate and update gate networks with parameters θ^{q} and θ^{z}, respectively. The initial disparity candidates $\hat{d}_k^{\text{init}}$ and uncertainty $\hat{u}_k^{\text{init}}$ are decoded from the initial hidden state $H_1 = H_3^{\text{init}}$. The initial disparity is selected as $\hat{d}_1 = \underset{\hat{d}_k^{\text{init}}}{\text{argmin}}(\hat{u}_k^{\text{init}})$. Notably, the entire UFI module, with the exception of the reset gate $f_{\theta^{\text{z}}}^{\text{z}}$, is computed in parallel between sampling windows. The preliminary experiment shown in Figure 6 indicates an apparent improvement in initial disparity accuracy with the UFI module. It validates the effectiveness of multi-disparity window scan-and-select initialization.

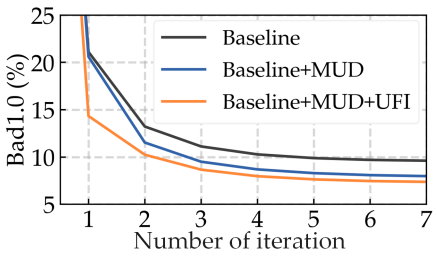


Figure 6. The proportion of disparity outliers after each iteration. "Bad1.0" is the metric that indicates the percentage of pixels with an error greater than 1 pixel.

3.2.2. The MGA Module

Attention weights in vision transformers [46,47] are used to aggregate features by assigning higher importance to those with relatively high dot-product scores. This process involves computing the dot product between different features. The dot-product scores are then normalized using the SoftMax function to produce attention weights. In GMA [48], attention weights $A^{\text{ctx}} \in [0,1]^{h^2 \times w^2}$ guided by contextual feature are reused during iteration for the global aggregation of flow field feature, thus improving the occlusion handling in optical flow task. Inspired by GMA, we propose a new MGA module, where the disparity field's hidden state is globally aggregated to introduce long-range dependencies of the hidden state. The attention weights are computed using the contextual feature encoded with 2D rotary positional embedding (RoPE) [49] to inject relative positional information:

$$A^{\text{ctx}} = \zeta(\text{rope}(f_{\theta^{\text{qry}}}^{\text{qry}}(F^{\text{ctx}})) \otimes \text{rope}(f_{\theta^{\text{key}}}^{\text{key}}(F^{\text{ctx}}))), \tag{11}$$

where $f^{\text{qry}}_{\theta^{\text{qry}}}$, $f^{key}_{\theta^{key}}$ denote convolutional networks with parameters θ^{qry} and θ^{key}, and $\zeta(\cdot)$ denotes the SoftMax function. The iterative update of disparity field with the MGA module is denoted as:

$$\begin{aligned} \hat{m}^{\text{hm}}_t &= f^{\text{hm}}_{\theta^{\text{hm}}}(\text{H}_{t-1}, \text{C}^{\text{mtc}}(\hat{D}_1), \hat{d}_{t-1}, \hat{m}_{t-1}) \\ H^{\text{glb}}_t &= \text{MGA}(A^{\text{ctx}}, H_{t-1}, \hat{m}^{\text{hm}}_t) = A^{\text{ctx}} \otimes (f^{\text{vel}}_{\theta^{\text{vel}}}(H_{t-1}) \odot \hat{m}^{\text{hm}}_t), \\ H_t &= \text{ConvGRU}(H_{t-1}, H^{\text{glb}}_t, F^{\text{df}}_t, F^{\text{ctx}}) \end{aligned} \tag{12}$$

where $f^{\text{hm}}_{\theta^{\text{hm}}}$, $f^{\text{vel}}_{\theta^{\text{vel}}}$ denote the convolutional network with parameters θ^{hm} and θ^{vel}. The adaptive mask $\hat{m}^{\text{hm}}_t$ is used to dynamically adjust the attention weights during iteration, thus suppressing the propagation of unreliable hidden states, thus ensuring the effectiveness of aggregation.

4. Experiments

In this section, the proposed MUIR is pre-trained on the synthetic Scene Flow [21] dataset and ablation experiments are performed on the proposed MUD, UFI and MGA modules to evaluate the contribution of each improvement (in Section 4.3). The generalization experiments on KITTI [23,24], Middlebury [25], ETH3D [26] and New Tsukuba [50] are then performed to evaluate the generalization performance of the pre-trained MUIR (in Section 4.4).

4.1. Datasets and Evaluation Metrics

- The Scene Flow dataset [21] is a large synthetic dataset commonly used for pre-training models. The finalpass version of the stereo image pairs, synthesized using Blender, contains motion and defocus blur effects that simulate the real scenes. It provides 35,454 training image pairs and 4370 test image pairs with disparity ground truth at a resolution of 960 × 540. It should be noted that other existing synthetic datasets, such as Fallingthings [51], CREStereo [13], TartanAir [52], etc., are not used to ensure a fair comparison with existing work.
- The KITTI [23,24] provides datasets of real outdoor scenes captured by on-board calibrated cameras and LIDAR on the road, where the KITTI2012 [23] and the KITTI2015 [24] contain 194 pairs and 200 pairs of training data with sparse disparity ground truth at a resolution of about 1220 × 370.
- The Middlebury [25] provides datasets of real indoor scenes obtained using dense structured light, containing stereo image pairs with multiple asymmetric exposure combinations of multiple scenes. The Middlebury 2014 and Middlebury 2021 contain 23 and 24 sets of semi-dense disparity ground truth with maximum resolutions of 3052 × 1968 and 1920 × 1080, respectively.
- The ETH3D [26] provides a dataset of real indoor and outdoor scenes captured by a high-precision laser scanner. It provides 27 training grayscale stereo image pairs with corresponding semi-dense disparity ground truth at a resolution about 950 × 500.
- The New Tsukuba [50] dataset provides a realistic computer-generated stereo dataset that includes four different illumination conditions. It provides 1800 stereo image pairs, each with 256 levels of disparity ground truth, at a resolution of 640 × 480.

For quantitative evaluation and comparison, we mainly use endpoint error EPE, the 50th percentile error A50, and the percentage of bad pixels Bad1.0 and D1 to evaluate the disparity accuracy. The EPE is also called "avgerr" in the Middlebury benchmark and "Average error" in ETH3D. The Bad1.0 is the percentage of pixels with an error greater than one pixel. The D1 is the percentage of pixels with an absolute error greater than three pixels or relative error greater than 5% of the ground truth.

4.2. Implementation Details

The MUIR proposed in this paper is implemented using PyTorch [53] and trained from scratch on NVIDIA RTX2080Ti and RTX TITAN GPUs. During pre-training, we use the AdamW [54] optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 1×10^{-5}) and the one-cycle scheduler [55] to adjust the learning rate with maximum value of 2×10^{-4} The training image pairs are cropped to a size of 320×704. The models are pre-trained for 2×10^5 steps on the Scene Flow training set with the batch size set to eight. The data augmentation strategy aligns with [12]. To ensure the practicality and portability, the proposed MUIR fast model is run at 1/8 resolution and with seven iterations. The performance of the MUIR fast model is primarily evaluated and compared with the baseline, which is the fast version of RAFT-Stereo [12] with the Slow-Fast bi-level GRUs and seven iterations. The performance of MUIR with 15 iterations at 1/4 resolution is also reported where necessary.

4.3. Ablation Study

Ablation experiments are performed on the proposed MUIR pre-trained model to discuss the effectiveness of the proposed MUD, UFI and MGA modules. The detailed quantitative evaluation results are reported in Table 1. It should be noted that due to the absence of ill-posed region labels, the model performance in occluded regions is evaluated as an approximation. The occlusion labels are obtained from the warping error between left and right view disparities. The warping error occurs when sampling the right view's disparity based on the left view's disparity, resulting in a warped disparity. A point is marked as occluded when this warping error exceeds a certain threshold. The EPE and Bad1.0 metrics are divided into all pixels (All), non-occluded pixels (Noc) and occluded pixels (Occ). The absolute prediction error (APE) and the area under the sparsification error curve (AUSE) are calculated to quantitatively evaluate the numerical and distributional accuracy of the estimated uncertainty.

Table 1. Ablation study on the proposed MUIR. The evaluation is conducted on the held-out Scene Flow [21] test set [1]. Rows (1–7) and rows (8–11) contain models performed at 1/8 and 1/4 resolution, respectively. The best results in each category are bolded.

	Models [1]	Param (M)	Time (s)	EPE (px)			A50 (px)	Bad1.0 (%)			APE (px)	AUSE (%)
				All	Noc	Occ		All	Noc	Occ		
(1)	Baseline fast [12]	9.86	0.052	0.866	0.544	2.523	0.180	9.573	6.142	28.890	-	-
(2)	+Simple uncertainty [17]	10.17	0.041	1.077	0.557	3.325	0.161	9.910	6.339	30.010	0.820	5.113
(3)	+Ours uncertainty	10.32	0.040	0.911	0.544	2.646	0.207	9.354	5.935	28.618	0.482	5.598
(4)	+MUD	10.30	0.050	0.797	0.494	2.378	0.127	8.390	5.287	25.728	0.433	5.042
(5)	+MUD+UFI	12.89	0.059	0.766	0.456	2.328	0.131	8.010	4.963	24.927	0.423	4.872
(6)	+MUD+UFI+GMA [48]	13.64	0.065	0.731	**0.436**	2.215	0.124	7.779	**4.794**	24.339	**0.405**	**4.768**
(7)	MUIR fast (+MUD+UFI+MGA)	14.00	0.066	**0.711**	0.461	**2.144**	**0.099**	**7.769**	4.853	**24.002**	**0.405**	4.977
(8)	Baseline [12]	11.11	0.215	0.690	0.406	2.075	0.155	7.255	4.304	23.366	-	-
(9)	+MUD	12.92	0.222	0.643	0.367	1.956	0.135	6.197	3.554	20.705	0.253	3.497
(10)	+MUD+UFI	15.95	0.268	0.617	0.351	1.917	0.120	5.957	3.389	20.063	0.248	**3.289**
(11)	MUIR (+MUD+UFI+MGA)	17.15	0.295	**0.570**	**0.325**	**1.767**	**0.102**	**5.819**	**3.302**	**19.618**	**0.236**	3.493

[1] These results were measured with maximum disparity set to 192 pixels following [12].

The proposed MUIR fast and MUIR pre-trained models are evaluated in rows 7 and 11 of Table 1. Compared to the baseline (rows 1, 8), the disparity EPE is reduced by about 17.9% and 17.4%, respectively, and the A50 is reduced by about 45% and 34.2%, respectively, indicating a lower and more concentrated distributed disparity error. In addition, the Bad1.0 in occluded regions is reduced by about 16.9% and 16%, indicating the improved robustness of MUIR for ill-posed regions.

4.3.1. Ablation of the MUD Module

The MUD module is proposed in Section 3.1 and is designed to reduce the susceptibility to mismatched pixels. The matchability mask and disparity uncertainty of each iteration can be decoded from the hidden state using the MUD. In row 3 of Table 1, the modified joint disparity–uncertainty estimation proposed in Section 3.1.2 is evaluated. Compared to the original method [17] (row 2), our method reduces the disparity EPE by about 15.4% and the uncertainty APE by 41.2%. It alleviates the unnecessary decrease in disparity accuracy during joint training. Following this, in rows 4 and 9 of Table 1, the MUD module is evaluated. Compared to the uncertainty-aware only method (row 3), the disparity EPE and uncertainty APE is reduced by about 12.5% and 11.3%, respectively. Compared to the baseline (row 1, 8), the EPE and Bad1.0 Occ are reduced by about 8% and 10.9%, respectively. This demonstrates the improved robustness in ill-posed regions.

Notably, the contribution of the MUD module is not only for the improvements in quantitative evaluation results, but also for enhancing the interpretability of models through the estimated matchability mask and disparity uncertainty shown in Figure 7d,f. Both the occluded and textureless regions are masked out in the matchability mask. The disparity error converted from the estimated uncertainty keeps highly consistent with the error map. With a specified error threshold (1px in Figure 7f), flying pixels at depth discontinuity boundaries can be easily filtered out, as shown in Figure 7g,h.

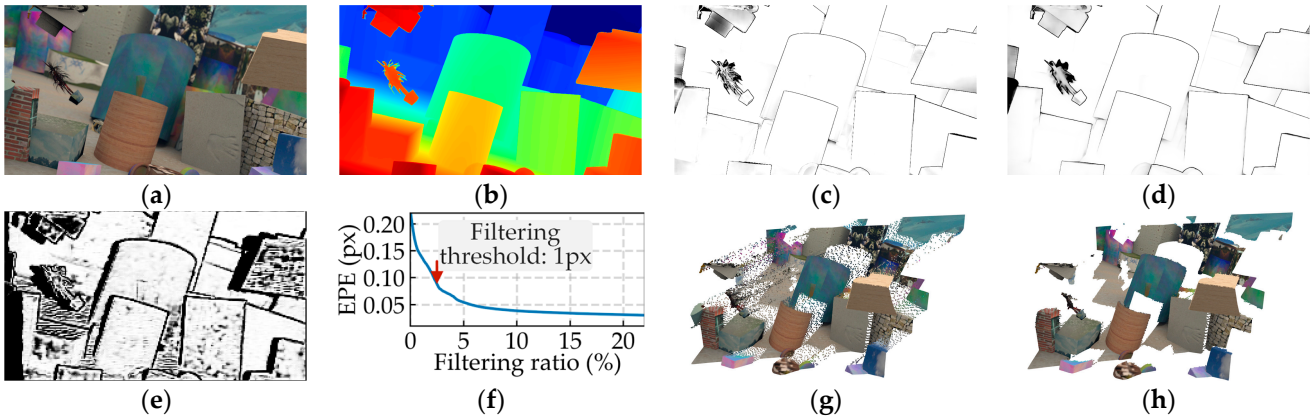


Figure 7. Illustration of the output of the MUIR and the alternative output filtering. (**a**) The input left image; (**b**) the estimated disparity map; (**c**) the disparity error map; (**d**) the error map converted from the estimated uncertainty; (**e**) the estimated matchability mask; (**f**) the error-filtering ratio curve (the original EPE is 0.223 px); (**g**,**h**) the point cloud converted from the estimated disparity map w/o filtering (**g**) and with filtering threshold set to 1 px (**h**). The results are best viewed when zoomed in.

4.3.2. Ablation of UFI Module

The UFI module, proposed in Section 3.2.1, is an initialization module that employs the multi-disparity window scan-and-select method to provide more appropriate initialized hidden state and more accurate initial disparity for the following refinement process. The UFI module is evaluated in rows 5 and 10 of Table 1. Compared to the method without UFI (rows 4 and 9), the disparity EPE is reduced by 3.9%. This is attributed to the fact that the UFI module decouples the initialization from the iterative refinement process, thus easing the parameter optimization burden on the iterator. The disparity accuracy of models at each iteration are shown in Table 2. The UFI module is evaluated in rows 3 and 6. Compared to the baseline (rows 1 and 5), the initial disparity EPE and Bad1.0 are reduced by about 27.5% and 31.8%, respectively, showing that the main contribution of UFI is in improving the initial disparity accuracy.

Table 2. Ablation study for models with different iterations on the Scene Flow test set. Rows (1–4) and rows (5–8) shows models' performance at 1/8 and 1/4 resolution, respectively. The best results in each category are bolded.

	Models	EPE (px)						Bad1.0 (%)					
		Init [1]	3rd itr [2]	5th itr	7th itr	10th itr	15th itr	Init [1]	3rd itr[2]	5th itr	7th itr	10th itr	15th itr
(1)	Baseline fast [12]	1.576	0.957	0.882	0.866	-	-	21.054	11.025	9.847	9.573	-	-
(2)	+MUD	1.558	0.891	0.817	0.797	-	-	20.920	9.890	8.678	8.390	-	-
(3)	+MUD+UFI	**1.142**	0.853	0.783	0.766	-	-	**14.362**	9.307	8.267	8.010	-	-
(4)	MUIR fast	1.156	**0.817**	**0.736**	**0.711**	-	-	14.541	**9.032**	**8.028**	**7.769**	-	-
(5)	Baseline [12]	2.159	0.924	0.786	0.743	0.706	0.690	25.844	10.641	8.347	7.668	7.344	7.225
(6)	+MUD	2.218	0.873	0.723	0.678	0.655	0.643	27.393	9.481	7.260	6.614	6.319	6.197
(7)	+MUD+UFI	**1.182**	0.801	0.699	0.658	0.631	0.617	**15.851**	**8.480**	6.811	6.294	6.060	5.957
(8)	MUIR	1.239	**0.758**	**0.636**	**0.593**	**0.573**	**0.570**	16.746	8.599	**6.710**	**6.149**	**5.893**	**5.819**

[1] "Init" is short for initialization; it is also treated as 1st iteration in other works [14,32]. [2] "itr" is short for iteration.

4.3.3. Ablation of MGA Module

The MGA module, proposed in Section 3.2.2, is a hidden state aggregation module that introduces long-range dependencies into the hidden state for enhancing the propagation of matching information. Unlike GMA [48], the MGA performs global aggregation directly on the adaptively masked hidden state. The MGA is evaluated in rows 7 and 11 of Table 1. Compared to the models without hidden state aggregation (rows 5 and 10), the disparity EPE is reduced by about 7.2%. In contrast, the EPE of the model with GMA (row 6) is reduced by only 4.6%. This indicates that the adaptive masked aggregation of MGA is more effective. In addition, according to rows 4 and 7 of Table 2, compared to the models without MGA (rows 3 and 6), the average EPE reduction ratio is improved by about 33.7%, which shows the improvement of disparity refinement efficiency.

4.4. Generalization Evaluations

To further evaluate the cross-domain generalization performance of the proposed MUIR, the zero-shot generalization evaluation is conducted on the KITTI2015, Middlebury Stereo Evaluation 3 and ETH3d training sets. The results are shown in Table 3.

Table 3. Quantitative zero-shot generalization evaluation on the KITTI 2015 [24], Middlebury Eval 3 [25] and ETH3D [26] training sets. Rows (1–6) and rows (7 and 8) show models' performance at 1/8 and 1/4 resolution, respectively. The best results in each category are bolded.

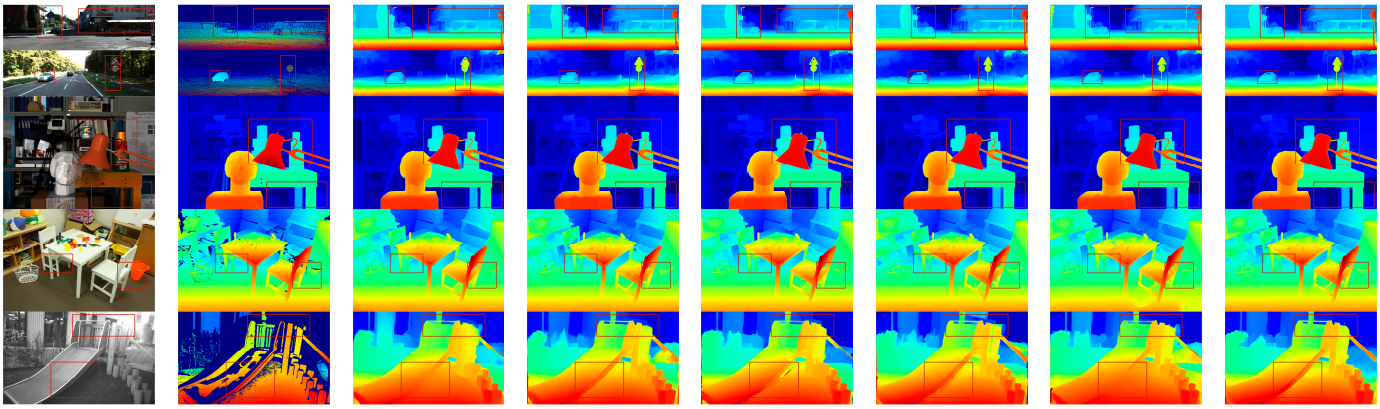
	Models	Param (M)	Scene Flow		KITTI 2015			Middlebury Eval 3 [1]			ETH3D		
			Bad1.0 (%)	EPE (px)	D1-all (%)	EPE (px)	APE (px)	Bad2.0 (%)	EPE (px)	APE (px)	Bad1.0 (%)	EPE (px)	APE (px)
(1)	Baseline fast [12]	9.86	9.573	0.866	5.887	**1.217**	-	15.153	1.943	-	6.270	0.523	-
(2)	IGEV fast [32] [2]	12.38	8.283	0.676	5.880	1.218	-	12.448	1.665	-	6.629	0.568	-
(3)	EAI fast [14]	14.13	8.595	0.808	**5.834**	1.246	-	14.979	2.198	-	6.022	0.416	-
(4)	MUIR fast	14.00	7.769	0.711	6.291	1.244	**0.812**	12.235	1.688	1.087	3.801	**0.312**	**0.174**
(5)	MUIR-IGEV fast [2]	13.39	**7.101**	**0.625**	6.439	1.292	0.833	11.496	1.611	0.966	5.980	0.587	0.281
(6)	MUIR-EAI fast	18.05	7.200	0.675	6.442	1.271	0.848	**11.155**	**1.394**	**0.813**	**3.694**	0.324	0.197
(7)	Baseline [12]	11.11	7.225	0.690	5.843	1.402	-	10.099	1.286	-	**2.922**	0.295	-
(8)	MUIR	17.15	**5.819**	**0.570**	**5.719**	**1.214**	0.760	**8.323**	**1.081**	0.578	3.889	**0.275**	0.174

[1] The evaluation in Middlebury Eval 3 is conducted on the half resolution. [2] The results of IGEV fast and MUIR-IGEV fast are measured with the maximum disparity set to 192 pixels, following [32].

In Table 3, rows 2 and 6 show that, compared to the baseline, the disparity EPE of MUIR fast is reduced by about 19.2% in Middlebury Eval 3 and 39.4% in ETH3D. To further study the applicability of MUIR to other RAFT-based variants, the baseline is replaced with IGEV-Stereo fast (row 5) and EAI-Stereo fast (row 6), fast versions of [14,32] running at 1/8 resolution with seven iterations. The modified models are denoted MUIR-IGEV fast (row 5) and MUIR-EAI fast (row 6). Compared to each baseline, the MUIR-IGEV fast achieves 7.6% and 9.8% outlier proportion reduction in Middlebury Eval 3 and ETH3D, respectively, and

the MUIR-EAI fast also achieves 25.5% and 38.7% outlier proportion reduction. It should be noted that, compared to the corresponding baselines, the D1-all of MUIR, MUIR-IGEV and MUIR-EAI are increased by 6.9%, 9.5% and 10.4%, respectively, due to the excessive domain gap in KITTI datasets.

The visualization of partial experimental results is shown in Figure 8. The generalization experimental results show that the proposed MUIR can be applied to multiple RAFT-based models and provide performance improvements especially in some ill-posed regions.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**)

Figure 8. Visual comparison between various models. The datasets are arranged from top to bottom in the following order: KITTI2012, KITTI2015, New Tsukuba, Middlebury and ETH3d. (**a**) The input left image; (**b**) the disparity ground truth. (**c**–**h**) are estimated disparity maps of: (**c**) RAFT-Stereo [12]; (**d**) our proposed MUIR-RAFT; (**e**) IGEV-Stereo [32]; (**f**) our proposed MUIR-IGEV; (**g**) EAI-Stereo [14]; (**h**) our proposed MUIR-EAI. The details are marked in the red box and best viewed when zoomed in.

5. Conclusions

The MUIR proposed in this paper utilizes the novel MUD module to improve the performance of RAFT-based models. The MUD module learns to decode the feature matchability mask and the disparity uncertainty concurrently from the disparity field's hidden state. This innovative approach reduces the model's susceptibility to mismatched pixels and alleviates the unnecessary reduction in disparity accuracy that occurs during joint training. Furthermore, building on the MUD module, we introduce the UFI and MGA modules to improve refinement efficiency. Comprehensive experimental evaluations on synthetic and real-world datasets show that our method reduces disparity error and occluded outlier proportion by up to 17.9% and 16.9%, respectively, compared to the baseline, demonstrating the improvement in disparity accuracy and robustness of MUIR. Future research may focus on incorporating global image features and scalable disparity encoding techniques to further optimize scale generalization.

Author Contributions: Conceptualization, J.W.; methodology, J.W.; formal analysis, J.W.; investigation, J.W. and Y.T.; resources, H.G.; writing—original draft preparation, J.W.; writing—review and editing, J.W., W.Z. and Y.T.; supervision, Y.T. and H.G.; funding acquisition, H.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science & Technology Commission of Shanghai Municipality (STCSM), grant number 21010502900.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: We thank the anonymous reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hirschmuller, H. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 328–341. [CrossRef] [PubMed]
2. Bleyer, M.; Rhemann, C.; Rother, C. PatchMatch Stereo-Stereo Matching with Slanted Support Windows. In Proceedings of the British Machine Vision Conference 2011, Dundee, Scotland, 29 August–2 November 2011; pp. 14.1–14.11.
3. Scharstein, D.; Szeliski, R. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42. [CrossRef]
4. Chang, J.-R.; Chen, Y.-S. Pyramid Stereo Matching Network. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5410–5418.
5. Guo, X.; Yang, K.; Yang, W.; Wang, X.; Li, H. Group-Wise Correlation Stereo Network. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3268–3277.
6. Zhang, F.; Prisacariu, V.; Yang, R.; Torr, P.H.S. GA-Net: Guided Aggregation Net for End-To-End Stereo Matching. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 185–194.
7. Khamis, S.; Fanello, S.; Rhemann, C.; Kowdle, A.; Valentin, J.; Izadi, S. StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11219, pp. 596–613, ISBN 978-3-030-01266-3.
8. Laga, H.; Jospin, L.V.; Boussaid, F.; Bennamoun, M. A Survey on Deep Learning Techniques for Stereo-Based Depth Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1738–1764. [CrossRef] [PubMed]
9. Yang, Q.; Wang, L.; Yang, R.; Stewenius, H.; Nister, D. Stereo Matching with Color-Weighted Correlation, Hierarchical Belief Propagation, and Occlusion Handling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 492–504. [CrossRef] [PubMed]
10. Huq, S.; Koschan, A.; Abidi, M. Occlusion Filling in Stereo: Theory and Experiments. *Comput. Vis. Image Underst.* **2013**, *117*, 688–704. [CrossRef]
11. Zhan, Y.; Gu, Y.; Huang, K.; Zhang, C.; Hu, K. Accurate Image-Guided Stereo Matching with Efficient Matching Cost and Disparity Refinement. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *26*, 1632–1645. [CrossRef]
12. Lipson, L.; Teed, Z.; Deng, J. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. In Proceedings of the 2021 International Conference on 3D Vision (3DV), London, UK, 1–3 December 2021; pp. 218–227.
13. Li, J.; Wang, P.; Xiong, P.; Cai, T.; Yan, Z.; Yang, L.; Liu, J.; Fan, H.; Liu, S. Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18–24 June 2022; pp. 16242–16251.
14. Zhao, H.; Zhou, H.; Zhang, Y.; Zhao, Y.; Yang, Y.; Ouyang, T. EAI-Stereo: Error Aware Iterative Network for Stereo Matching. In Proceedings of the 16th Asian Conference on Computer Vision (ACCV), Macao, China, 4–8 December 2022; Volume 13841, pp. 3–19.
15. Meister, S.; Hur, J.; Roth, S. UnFlow: Unsupervised Learning of Optical Flow with a Bidirectional Census Loss. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 7251–7259.
16. Xu, H.; Zhang, J.; Cai, J.; Rezatofighi, H.; Yu, F.; Tao, D.; Geiger, A. Unifying Flow, Stereo and Depth Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13941–13958. [CrossRef] [PubMed]
17. Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
18. Zbontar, J.; LeCun, Y. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *J. Mach. Learn. Res.* **2016**, *17*, 1–32.
19. Zhang, Y.; Chen, Y.; Bai, X.; Yu, S.; Yu, K.; Li, Z.; Yang, K. Adaptive Unimodal Cost Volume Filtering for Deep Stereo Matching. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12926–12934.
20. Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; Bry, A. End-to-End Learning of Geometry and Context for Deep Stereo Regression. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 66–75.
21. 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 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4040–4048.
22. 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 Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), ELECTR NETWORK, Prague, Czech Republic, 27 September–1 October 2021; pp. 3542–3548.
23. Geiger, A.; Lenz, P.; Urtasun, R. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.

24. Menze, M.; Heipke, C.; Geiger, A. Joint 3D Estimation of Vehicles and Scene Flow. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* **2015**, *II-3/W5*, 427–434. [CrossRef]
25. Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; Westling, P. High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. In *Pattern Recognition*; Jiang, X., Hornegger, J., Koch, R., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2014; Volume 8753, pp. 31–42, ISBN 978-3-319-11751-5.
26. Schöps, T.; Schönberger, J.L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; Geiger, A. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2538–2547.
27. Duggal, S.; Wang, S.; Ma, W.-C.; Hu, R.; Urtasun, R. DeepPruner: Learning Efficient Stereo Matching via Differentiable PatchMatch. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4383–4392.
28. Li, Z.; Liu, X.; Drenkow, N.; Ding, A.; Creighton, F.X.; Taylor, R.H.; Unberath, M. Revisiting Stereo Depth Estimation from a Sequence-to-Sequence Perspective with Transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), ELECTR NETWORK, Montreal, QC, Canada, 11–17 October 2021; pp. 6177–6186.
29. Teed, Z.; Deng, J. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. Computer Vision-ECCV 2020. In Proceedings of the 16th European Conference; Proceedings. Lecture Notes in Computer Science (LNCS 12347), Glasgow, UK, 23–28 August 2020; pp. 402–419.
30. Cho, K.; van Merrienboer, B.; Bahdanau, D.; Bengio, Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv* **2014**, arXiv:1409.1259.
31. Zhao, H.; Zhou, H.; Zhang, Y.; Chen, J.; Yang, Y.; Zhao, Y. High-Frequency Stereo Matching Network. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 1327–1336.
32. Xu, G.; Wang, X.; Ding, X.; Yang, X. Iterative Geometry Encoding Volume for Stereo Matching. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 21919–21928.
33. Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; Zhou, X. LoFTR: Detector-Free Local Feature Matching with Transformers. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, ELECTR NETWORK, Virtual, 19–25 June 2021; pp. 8918–8927.
34. Poggi, M.; Mattoccia, S. Learning from Scratch a Confidence Measure. In Proceedings of the British Machine Vision Conference 2016, York, UK, 19–22 September 2016; pp. 46.1–46.13.
35. Seki, A.; Pollefeys, M. Patch Based Confidence Prediction for Dense Disparity Map. In Proceedings of the British Machine Vision Conference 2016, York, UK, 19–22 September 2016; pp. 23.1–23.13.
36. Poggi, M.; Mattoccia, S. Learning to Predict Stereo Reliability Enforcing Local Consistency of Confidence Maps. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 12–26 July 2017; pp. 4541–4550.
37. Tosi, F.; Poggi, M.; Benincasa, A.; Mattoccia, S. Beyond Local Reasoning for Stereo Confidence Estimation with Deep Learning. In *Computer Vision–ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11210, pp. 323–338, ISBN 978-3-030-01230-4.
38. Kim, S.; Kim, S.; Min, D.; Sohn, K. LAF-Net: Locally Adaptive Fusion Networks for Stereo Confidence Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 205–214.
39. Truong, P.; Danelljan, M.; Timofte, R.; Van Gool, L. PDC-Net+: Enhanced Probabilistic Dense Correspondence Network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10247–10266. [CrossRef] [PubMed]
40. Eldesokey, A.; Felsberg, M.; Holmquist, K.; Persson, M. Uncertainty-Aware CNNs for Depth Completion: Uncertainty from Beginning to End. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 12011–12020.
41. Zhang, J.; Yao, Y.; Luo, Z.; Li, S.; Shen, T.; Fang, T.; Quan, L. Learning Stereo Matchability in Disparity Regression Networks. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), ELECTR NETWORK, Milan, Italy, 10–15 January 2021; pp. 1611–1618.
42. Mehltretter, M.; Heipke, C. Aleatoric Uncertainty Estimation for Dense Stereo Matching via CNN-Based Cost Volume Analysis. *ISPRS J. Photogramm. Remote Sens.* **2021**, *171*, 63–75. [CrossRef]
43. Mehltretter, M. Joint Estimation of Depth and Its Uncertainty from Stereo Images Using Bayesian Deep Learning. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* **2022**, *2*, 69–78. [CrossRef]
44. Chen, L.; Wang, W.; Mordohai, P. Learning the Distribution of Errors in Stereo Matching for Joint Disparity and Uncertainty Estimation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 17235–17244.
45. Ghiasi, G.; Lin, T.-Y.; Le, Q.V. DropBlock: A Regularization Method for Convolutional Networks. In Proceedings of the Advances in Neural Information Processing Systems 31 (NIPS), Montreal, Canada, 2–8 December 2018; Volume 31.

46. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, Austria, 3–7 May 2021.
47. Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; Li, Y. MaxViT: Multi-Axis Vision Transformer. In *Computer Vision – ECCV 2022*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Lecture Notes in Computer Science; Springer Nature Switzerland: Cham, Switzerland, 2022; Volume 13684, pp. 459–479, ISBN 978-3-031-20052-6.
48. Jiang, S.; Campbell, D.; Lu, Y.; Li, H.; Hartley, R. Learning to Estimate Hidden Motions with Global Motion Aggregation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), ELECTR NETWORK, Montreal, QC, Canada, 11–17 October 2021; pp. 9752–9761.
49. Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; Liu, Y. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing* **2024**, *568*, 127063. [CrossRef]
50. Martull, S.; Peris, M.; Fukui, K. Realistic CG Stereo Image Dataset with Ground Truth Disparity Maps. *Sci. Program.* **2012**, *111*, 117–118.
51. Tremblay, J.; To, T.; Birchfield, S. Falling Things: A Synthetic Dataset for 3D Object Detection and Pose Estimation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2119–21193.
52. Wang, W.; Zhu, D.; Wang, X.; Hu, Y.; Qiu, Y.; Wang, C.; Hu, Y.; Kapoor, A.; Scherer, S. TartanAir: A Dataset to Push the Limits of Visual SLAM. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), ELECTR NETWORK, Las Vegas, NV, USA, 24 October 2020–24 January 2021; pp. 4909–4916.
53. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
54. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
55. Smith, L.N.; Topin, N. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. In Proceedings of the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, Baltimore, MD, USA, 15–17 April 2019; Volume 11006.