# Selective-Stereo: Adaptive Frequency Information Selection for Stereo Matching

Xianqi Wang*, Gangwei Xu*, Hao Jia, Xin Yang†
Huazhong University of Science and Technology
{xianqiw, gwxu, haojia, xinyang2014}@hust.edu.cn

## Abstract

*Stereo matching methods based on iterative optimization, like RAFT-Stereo and IGEV-Stereo, have evolved into a cornerstone in the field of stereo matching. However, these methods struggle to simultaneously capture high-frequency information in edges and low-frequency information in smooth regions due to the fixed receptive field. As a result, they tend to lose details, blur edges, and produce false matches in textureless areas. In this paper, we propose Selective Recurrent Unit (SRU), a novel iterative update operator for stereo matching. The SRU module can adaptively fuse hidden disparity information at multiple frequencies for edge and smooth regions. To perform adaptive fusion, we introduce a new Contextual Spatial Attention (CSA) module to generate attention maps as fusion weights. The SRU empowers the network to aggregate hidden disparity information across multiple frequencies, mitigating the risk of vital hidden disparity information loss during iterative processes. To verify SRU's universality, we apply it to representative iterative stereo matching methods, collectively referred to as Selective-Stereo. Our Selective-Stereo ranks $1^{st}$ on KITTI 2012, KITTI 2015, ETH3D, and Middlebury leaderboards among all published methods. Code is available at https://github.com/Windsrain/Selective-Stereo.*

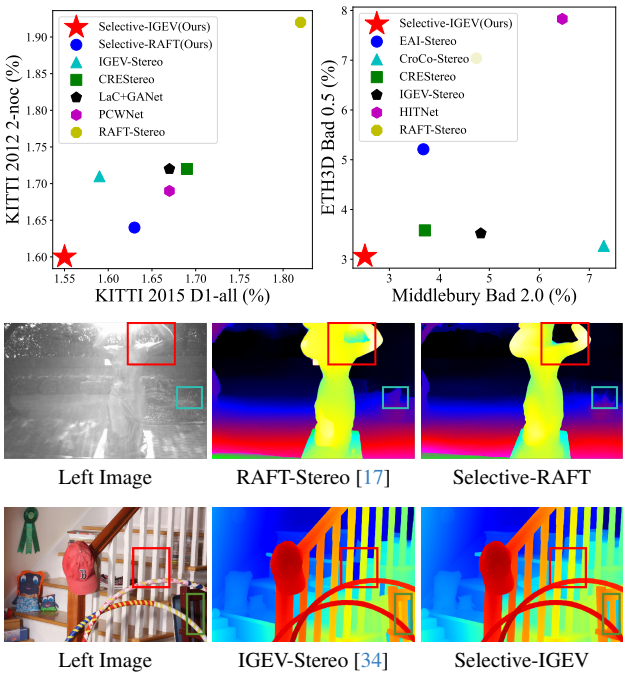



Figure 1. **Row 1:** Comparisons with state-of-the-art stereo methods on KITTI 2012 [13] and KITTI 2015 [21], ETH3D [24] and Middlebury [23] leaderboards. **Row 2:** Visual comparison with RAFT-Stereo on ETH3D. **Row 3:** Visual comparison with IGEV-Stereo on Middlebury. Our method distinguishes subtle details and sharp edges and performs well in weak texture regions.

## 1. Introduction

Stereo matching is a fundamental area of research in computer vision. It explores the calculation of displacement, referred to as disparity, between matching points in a pair of rectified images. This technique plays a significant role in various applications, including 3D reconstruction and autonomous driving.

With the advancement of deep learning, learning-based stereo matching [4, 14, 15, 17, 34, 35] have progressively displaced traditional methods and significantly enhancing the accuracy of disparity estimation. Initially, aggregation-based methods [7, 32, 41] led the development of stereo matching algorithms. These methods begin by defining a maximum range of disparity, constructing a 4D cost volume using feature maps, and subsequently employing 3D CNN to filter the volume and derive the final disparity map. Such methods focus on filtering the initially coarse cost volume, thus effectively aggregating geometry information. However, cost aggregation requires a large number of convolutions, resulting in high computational costs, making it difficult to be applied to high-resolution images.

Recently, a novel class of methods based on iterative optimization [11, 16, 17, 44] has been gaining prominence and

*Equal contribution.
†Corresponding author.

achieving state-of-the-art performance on several leaderboards. These methods begin by constructing an all-pairs cost volume, indexing a local cost volume from the original cost volume, and subsequently employing recurrent units [10] to calculate disparity residuals and update the disparity prediction. One major advantage of these methods is their ability to capture all candidate matching points without predefining the range of disparities. Additionally, these methods don't need to aggregate the cost volume using a large number of redundant convolutions. Instead, a continuous update of the disparity prediction is achieved through lightweight recurrent units during iterations. Therefore, these methods are capable of processing high-resolution images.

However, iterative methods encounter several challenges. Firstly, the all-pairs cost volume includes considerable noisy information [34], potentially causing the loss of crucial information when iterating the hidden information. Besides, as the network iterates, the hidden information increasingly incorporates global low-frequency information while losing local high-frequency information like edges and thin objects [44]. Secondly, the existing recurrent units possess a fixed receptive field, leading the network to solely concentrate on information at the current frequency and ignore other frequencies, such as detailed, edge, and textureless information.

In this paper, we propose Selective Recurrent Unit (SRU) to address the limitations of traditional recurrent units. As Chen *et al*. [5] mentions features contain information at different frequencies, high-frequency information describes rapidly changing fine details, while low-frequency information describes smoothly changing structures. Unlike traditional recurrent units that treat information at different frequencies equally, our SRU incorporates multiple branches of GRU, each with a distinct kernel size representing different receptive fields. The hidden information obtained from each GRU branch is fused and then fed into the next iteration. This fusion enables the capture of information from different receptive fields at different frequencies, while also performing secondary filtering to reduce noise information from local cost volume. To further enhance the fusion process, we propose a Contextual Spatial Attention (CSA) module to utilize the context information. Instead of simply summarizing information from different branches, CSA introduces attention maps extracted from the context information. After doing so, information captured by small kernels has large weights in regions like edge, while information captured by large kernels has large weights in regions like low-texture. These attention maps determine the weight of fusion, allowing the network to adaptively select suitable information based on different image regions. Besides, we prove the effectiveness and universality of our module by transferring it to different iterative networks. All networks are collectively referred to as Selective-Stereo. By doing so, we consistently improve the performance of these networks without introducing a significant increase in parameters and time.

We demonstrate the effectiveness of our method on several stereo benchmarks. On Scene Flow [20], our Selective-RAFT reaches the state-of-the-art EPE of 0.47, and our Selective-IGEV even achieves a new state-of-the-art EPE of 0.44. And as shown in Fig. 1, our Selective-RAFT surpasses RAFT-Stereo by a large margin and achieves competitive performance compared with the state-of-the-art methods on KITTI [13, 21] leaderboards. Our Selective-IGEV ranks $1^{st}$ on KITTI, ETH3D [24], and Middlebury [23] leaderboards among all published methods.

Our main contributions can be summarized as follows:

- We propose a novel iterative update operator SRU for iterative stereo matching methods.
- We introduce a new Contextual Spatial Attention module that generates attention maps for adaptively fusing hidden disparity information at multiple frequencies.
- We verify the universality of our SRU on several iterative stereo matching methods.
- Our method outperforms existing published methods on public leaderboards such as KITTI, ETH3D, and Middlebury.

## 2. Related Work

**Aggregation-based methods in stereo matching.** Several aggregation-based methods [4, 7, 8, 14, 15, 25, 35–37, 41] have shown significant progress in the domain of stereo matching in recent years. DispNet [20] establishes the groundwork for subsequent network architecture. GC-Net [15] proposes a 4D concatenate cost volume, which is subsequently regularized using 3D CNNs. Additionally, it also introduces the soft argmin function for disparity regression, resulting in a significant influence on subsequent methods. PSMNet [4] proposes a stacked hourglass 3D CNN, which improves the cost aggregation stage to enhance the network's ability to capture context information. Gwc-Net [14] proposes Group-wise Correlation Volume, which combines the advantages of correlation and concatenation volume. GA-Net [41] designs a semi-global guided aggregation layer and a local guided aggregation layer, inspired by the traditional semi-global matching algorithm, to further assist in aggregating global and geometry information in the network. Building upon Group-wise Correlation Volume, ACVNet [32] proposes Attention Concat Volume, which uses attention weights to suppress redundant information and maintain sufficient information for matching.

**Iterative-based methods in stereo matching.** In recent years, many iterative methods [17, 27, 33, 34], spearheaded by RAFT [17], have gradually become the mainstream of research. RAFT-Stereo [17] builds upon the optical flow

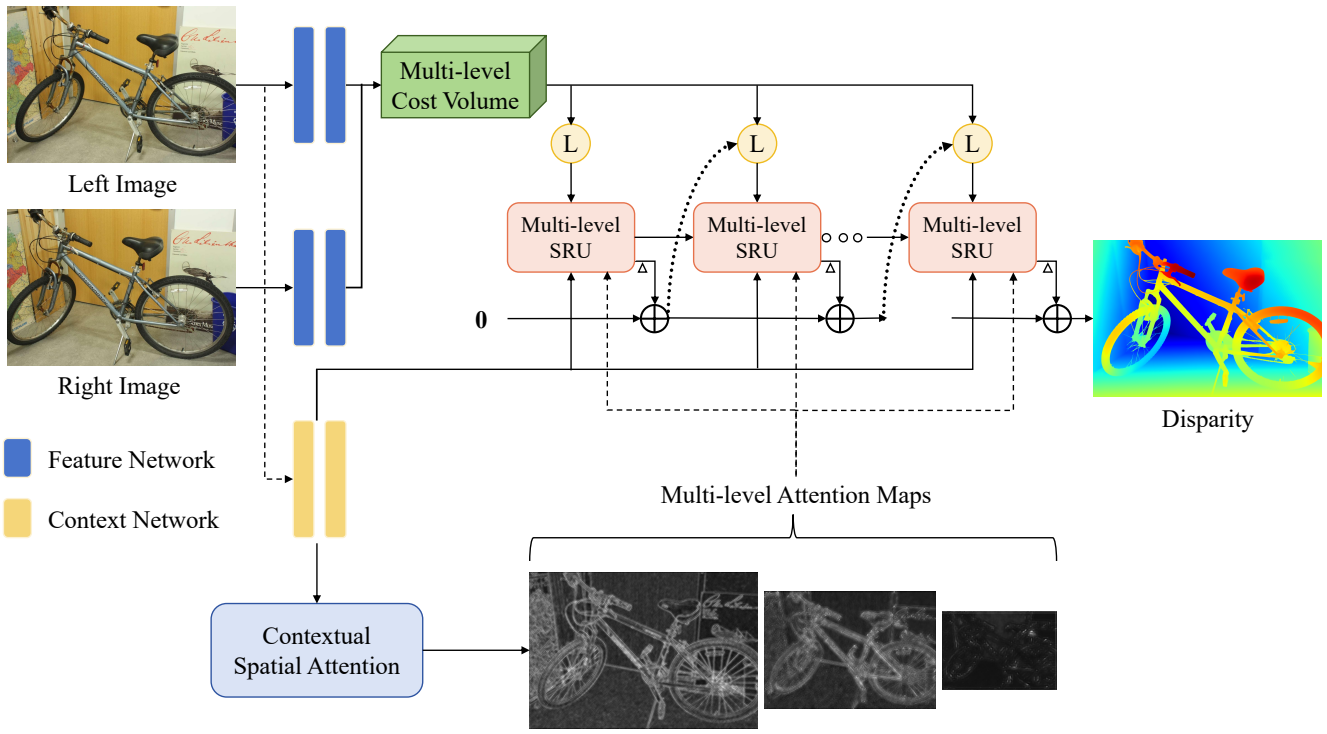



Figure 2. Overview of our proposed Selective-Stereo (Selective-RAFT version). The Contextual Spatial Attention (CSA) module extracts attention maps from context information as a guide for Selective Recurrent Units (SRUs). Then the network iteratively updates the disparity using local cost volumes retrieved from the correlation pyramid and attention maps given by CSA through SRUs.

method RAFT [27] by introducing an all-pairs cost volume pyramid that maintains high resolution. It extracts local correlation features from this pyramid, performs iterative disparity updates using GRU-based update operators, and incorporates a multi-level GRU to expand the receptive field. On this basis, IGEV-Stereo [34] asserts that the initial cost volume is excessively coarse. To alleviate the need for iterations and reduce time overhead, it proposes to use a lightweight cost aggregation network before iterations. CREStereo [16] designs a hierarchical network in a coarse-to-fine manner, as well as a stacked cascaded architecture for inference in place of the original single-resolution iterative structure. DLNR [44] proposes the use of LSTM as a replacement for GRU, providing the advantage of decoupling the update of hidden states from disparity prediction.

**Frequency information application in vision**. There are several works[5, 6, 39] that focus on using frequency information in computer vision. Chen *et al*. [5] propose the octave convolution to factorize the mixed feature maps by their frequencies. Xu *et al*. [39] propose a method of learning in the frequency domain and suggest that CNN models are more sensitive to low-frequency channels than high-frequency. DSGAN [12] introduces the frequency separation into super-resolution. LITv2 [22] proposes to disentangle the high/low-frequency patterns in an attention layer.

## 3. Method

In this section, we present the overall architecture of Selective-Stereo. Because our method can be plugged into different networks, we take Selective-RAFT (Fig. 2) as an example and focus on illustrating its key components.

### 3.1. Feature Extraction

To ensure fair comparisons, Selective-RAFT maintains consistency with RAFT-Stereo [17] by employing its feature extraction network. Feature extraction comprises two main components: feature network and context network.

**Feature Network.** Given the left and the right images $\mathbf{I}_{l(r)} \in \mathbb{R}^{3\times H\times W}$, we first downsample them to $1/2$ resolution using a $7 \times 7$ convolutional layer. Then, a series of residual blocks is employed to extract features and we apply another downsampling layer to get features at $1/4$ resolution in the middle. Lastly, a $1 \times 1$ convolutional layer is applied to get the final left and right features $\mathbf{f}, \mathbf{g} \in \mathbb{R}^{C\times \frac{H}{4}\times \frac{W}{4}}$ with suitable dimensions.

**Context Network.** Its architecture remains consistent with the feature network, and it adds a series of residual blocks and two additional downsampling layers, obtaining multi-level context features $\mathbf{f}_i^c$ ($i$ = 1, 2, 3) at 1/4, 1/8, 1/16 resolutions. Then we can get the initial hidden and the con-

text information:

$$\begin{aligned}\mathbf{h}_i &= \tanh(\mathbf{f}_i^c)\\ \mathbf{c}_i &= \mathrm{ReLU}(\mathbf{f}_i^c)\end{aligned} \tag{1}$$

### 3.2. Cost Volume Construction

Given the left and the right features $\mathbf{f}, \mathbf{g}$, we first construct an all-pairs correlate cost volume:

$$\mathbf{C}_{ijk} = \sum_h \mathbf{f}_{hij} \cdot \mathbf{g}_{hik}, \mathbf{C} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times \frac{W}{4}} \tag{2}$$

Then we construct a 4-level correlation pyramid $\{\mathbf{C}_i\}$ ($i$ = 1, 2, 3, 4) by using 1D average pooling with a kernel size of 2 and a stride of 2 at the last dimension.

### 3.3. Contextual Spatial Attention Module

To help information from different receptive fields and frequencies fuse, the Contextual Spatial Attention (CSA) module extracts multi-level attention maps from context information as guidence. As illustrated in Fig. 3, CSA can be divided into two submodules: Channel Attention Enhancement (CAE) and Spatial Attention Extractor (SAE). These submodules are derived from CBAM [31] and we simplify them to better adapt to stereo matching.

**Channel Attention Enhancement.** Given a context information map $\mathbf{c} \in \mathbb{R}^{C \times H \times W}$, we first use an average-pooling and a max-pooling operation on the spatial dimension to get two maps $\mathbf{f}_{avg}, \mathbf{f}_{max} \in \mathbb{R}^{C \times 1 \times 1}$. Then we use two convolutional layers to perform feature transformation on these maps separably. After that, we add these two maps together and use the sigmoid function to convert them into weights $\mathbf{M}_c \in \mathbb{R}^{C \times 1 \times 1}$ between 0 and 1. Lastly, using an element-wise product, the initial map can capture which channel map has high feature values to be enhanced, and which channel map has low feature values to be suppressed.

**Spatial Attention Extractor.** After the CAE module, we continue to use the same pooling operations, but now we pool on the channel dimension. Then we concatenate these pooling maps to form a map in $\mathbb{R}^{2 \times H \times W}$ and use one convolutional layer with a sigmoid function to generate the final attention map. Reviewing previous operations, this attention map has high weights in regions needing high-frequency information because this information possesses high feature values in the context information. Similarly, it has low weights in regions needing low-frequency information. In general, the attention map can explicitly distinguish regions that need information at different frequencies.

### 3.4. Selective Recurrent Unit

To capture information at different frequencies, Selective Recurrent Unit (SRU) uses attention maps extracted by CSA to fuse hidden information derived from GRUs with different kernel sizes.

**Multi-level update structure.** As illustrated in Fig. 4, SRUs at $1/8$, $1/16$ resolutions take the attention map, context information, hidden information at the same resolution, and the hidden information at adjacent resolutions as inputs. At $1/4$ resolution, SRUs take disparity, and local cost volume as additional inputs, and then their outputs will go through two convolutional layers to generate disparity residuals. The local cost volume is derived from the all-pairs correlation pyramid in the same way as RAFT-Stereo [17]. At last, disparities at $1/4$ resolution will be upsampled into full resolution using the convex combination.

**SRU's architecture**. A single GRU can be defined as follows:

$$\begin{aligned}z_k &= \sigma(\mathrm{Conv}([h_{k-1}, x_k], W_z)),\\ r_k &= \sigma(\mathrm{Conv}([h_{k-1}, x_k], W_r)),\\ \tilde{h}_k &= \tanh(\mathrm{Conv}([r_k \odot h_{k-1}, x_k], W_h)),\\ h_k &= (1 - z_k) \odot h_{k-1} + z_k \odot \tilde{h}_k\end{aligned} \tag{3}$$

where $x_k$ is the concatenation of disparity, correlation, hidden information, and context information previously defined. Unlike RAFT-Stereo [17] that divide the context information into $c_z$, $c_r$, $c_h$, we add it into $x_k$ because using convolutions with different kernel sizes can fully utilize context information.

As illustrated in Fig. 3, a single SRU can be defined as follows:

$$h_k = \mathbf{A} \odot h_k^s + (1 - \mathbf{A}) \odot h_k^l \tag{4}$$

where $\mathbf{A}$ denotes the attention map derived from CSA at the same resolution, $h_k^s$ denotes the GRU with smaller kernel sizes and $h_k^l$ denotes the larger one.

As Sec. 3.3 mentioned, the attention map has high weights in regions needing high-frequency information. Therefore, the GRU with smaller kernel sizes that can capture high-frequency information like edge, and thin objects should do element-wise products with the attention map directly, and the GRU with larger kernel sizes should do element-wise products with the contrary attention map.

**Receptive fields analysis.** The repective fields computing formula [1] can be defined as follows:

$$r_0 = \sum_{l=1}^{L}((k_l - 1)\prod_{i=1}^{l-1} s_i) + 1 \tag{5}$$

where $k_l$ denotes the kernel size, $s_i$ denotes the stride size, and $r_0$ denotes the whole network.

Given a multi-level structure like Fig. 4, if we take the $1/4$ resolution as the basis, and the downsampling operations can be regarded as a convolution with kernel size 3, stride size 2, this structure's receptive fields are $k$, $2k + 3$, $3k + 6$. That means it only has 3 fixed receptive fields in total.

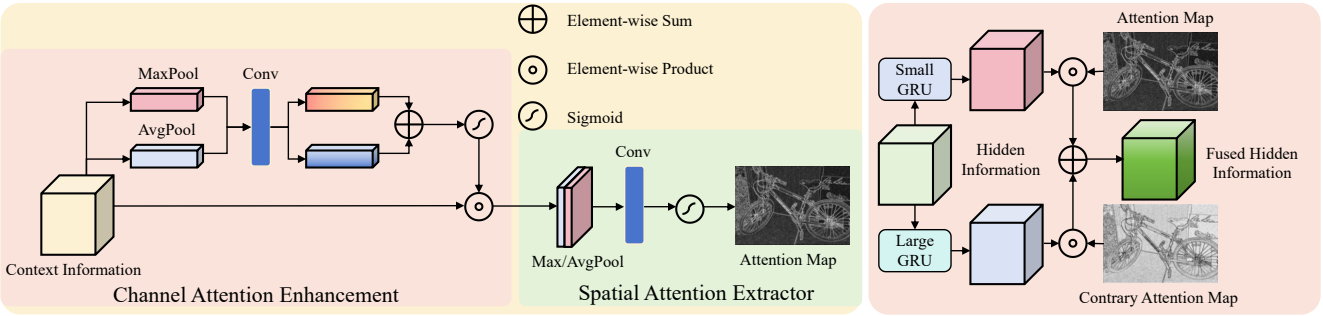



Figure 3. The architecture of proposed modules. Left: Contextual Spatial Attention (CSA) module. Right: Selective Recurrent Unit (SRU).

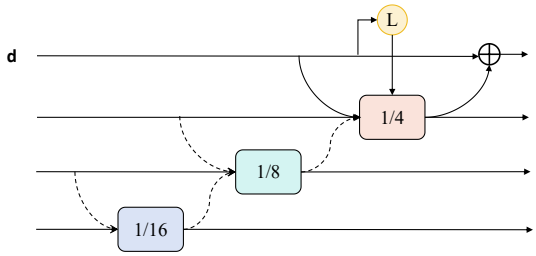



Figure 4. Multi-level SRU. Information is passed between SRUs at adjacent resolutions. Dashed arrows represent upsampling and downsampling operations. At $1/4$ resolution, disparity and local cost volume will be additional information put into SRUs.

If we replace GRUs with our SRUs with a small kernel size $s$ and a large kernel size $l$, the multi-level structure will have 6 receptive fields initially. Besides, pixels in hidden information are affected by different receptive fields during fusion, and the fusion is influenced by attention maps adaptively. In general, the multi-level SRU holds dynamic receptive fields, and it enables itself to capture information at different frequencies.

### 3.5. Loss Function

We supervise our network on the L1 distance between all predicted disparities $\{\mathbf{d}_i\}_{i=1}^{N}$ and the ground truth disparity $\mathbf{d}_{gt}$ with increasing weights. The total loss is defined as:

$$\mathcal{L} = \sum_{i=1}^{N} \gamma^{N-i} ||\mathbf{d}_i - \mathbf{d}_{gt}||_1 \tag{6}$$

where $\gamma = 0.9$, and $N$ is the number of iterations.

## 4. Experiments

**Scene Flow** [20] is a synthetic dataset including 35,454 training pairs and 4,370 testing pairs with dense disparity maps. For training and testing, we use the finalpass version, because it contains more realistic and difficult effects than the cleanpass version. **KITTI 2012** [13] and **KITTI 2015** [21] are datasets for real-world driving scenes. KITTI 2012 contains 194 training pairs and 195 testing pairs, and KITTI 2015 contains 200 training pairs and 200 testing pairs. **ETH3D** [24] is a collection of gray-scale stereo pairs containing 27 training pairs and 20 testing pairs for indoor and outdoor scenes. **Middlebury** [23] is a high-resolution dataset containing 15 training pairs and 15 testing pairs for indoor scenes.

### 4.1. Implementation Details

We implement our Selective-Stereo with PyTorch and the model is trained on NVIDIA RTX 3090 GPUs. For all experiments, we use the AdamW [19] optimizer and clip gradients to the range [-1, 1]. We use the one-cycle learning rate schedule with a learning rate of 2e-4. We first train our model on Scene Flow with a batch size of 8 for 200k steps as the pretrained model. The crop size is $320 \times 720$, and we use 22 update iterations during training.

### 4.2. Ablation Study

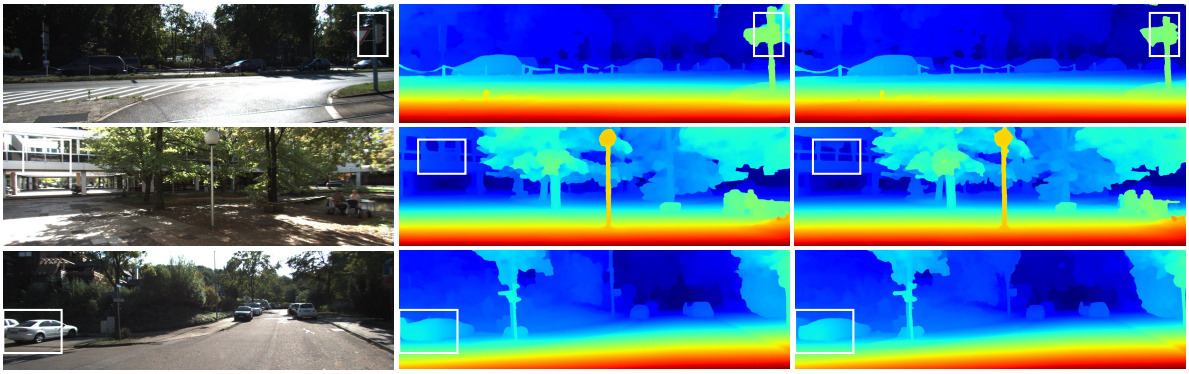
In this section, we evaluate our model in different settings to verify our proposed modules in several aspects. All results use 32 update iterations.

**Effectiveness of proposed modules**. To verify the effectiveness of our proposed modules, we take RAFT-Stereo [17] as the baseline and replace its GRUs with our SRUs. As shown in Tab. 1, the proposed SRU can improve the accuracy even without CSA. It means that if we just sum up the information from different branches, the growth of receptive fields can be beneficial for inference. If we add our CSA but invert the weights of the attention maps, the effect even decreases. That validates that our CSA's attention maps do indeed reflect the weights of information at different frequencies in regions. Therefore, if we add CSA normally, the full model (Selective-RAFT) can achieve the best performance with only a 4% increase in parameters.

**Universality of proposed modules.** To verify the universality of our proposed modules, we take three typical iterative stereo matching methods as the baseline and replace their GRUs with SRUs. Especially, in DLNR [44], the recurrent units are LSTMs but not GRUs, so we just replace GRUs inside SRUs with LSTMs to make a fair comparison.

| Model | GRU | SRU | CSA (Contrary) | CSA | EPE (px) | >1px (%) | Param (M) |
|---|---|---|---|---|---|---|---|
| Baseline (RAFT-Stereo) | ✓ | | | | 0.53 | 6.08 | 11.12 |
| SRU | | ✓ | | | 0.50 | 5.38 | 11.65 |
| SRU+CSA (Contrary) | | ✓ | ✓ | | 0.50 | 5.58 | 11.65 |
| Full model (Selective-RAFT) | | ✓ | | ✓ | **0.47** | **5.32** | 11.65 |

Table 1. Ablation study of the effectiveness of proposed modules on the Scene Flow test set. SRU denotes Selective Recurrent Unit, and CSA denotes Contextual Spatial Attention. Contrary means we invert the weights of the attention maps. The baseline is RAFT-Stereo.



Left Image IGEV-Stereo Selective-IGEV

Figure 5. Qualitative results on the test set of KITTI. Our Selective-IGEV outperforms IGEV in detailed and weak texture regions.

| Model | EPE (px) | >1px (%) | Param (M) |
|---|---|---|---|
| RAFT-Stereo [17] | 0.53 | 6.08 | 11.12 |
| Selective-RAFT | **0.47** | **5.32** | 11.65 |
| IGEV-Stereo [34] | 0.47 | 5.21 | 12.60 |
| Selective-IGEV | **0.44** | **4.98** | 13.14 |
| DLNR [44] | 0.49 | 5.06 | 57.37 |
| Selective-DLNR | **0.46** | **4.73** | 58.09 |

Table 2. Ablation study of the universality of proposed modules.

Left Image Selective-RAFT Selective-IGEV

Figure 6. Visualization of the attention map of different networks.

| Model | Number of Iterations | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 8 | 32 |
| RAFT-Stereo [17] | 2.08 | 1.13 | 0.87 | 0.75 | 0.58 | 0.53 |
| Selective-RAFT | **1.95** | **1.06** | **0.81** | **0.69** | **0.53** | **0.47** |
| IGEV-Stereo [34] | 0.66 | 0.62 | 0.58 | 0.55 | 0.50 | 0.47 |
| Selective-IGEV | **0.65** | **0.60** | **0.56** | **0.53** | **0.48** | **0.44** |

Table 3. Ablation study of the number of iterations.

| Kernel Sizes | EPE (px) | >1px (%) |
|---|---|---|
| $1 \times 1 + 1 \times 5$ | 0.48 | 5.41 |
| $3 \times 3 + 1 \times 5$ | 0.48 | **5.30** |
| $1 \times 1 + 3 \times 3$ | **0.47** | 5.32 |

Table 4. Ablation study of the size of convolutional kernels.

As shown in Tab. 2, all methods have a significant improvement in the EPE metrics on Scene Flow, and the insertion of modules only results in a slight increase in parameters. Besides, as shown in Fig. 6, the CSA module generates different attention maps in different networks. In Selective-RAFT, because the cost volume contains a large amount of noisy information, the network needs more large kernels to filter the local cost volume. On the contrary, the cost volume has already been aggregated in Selective-IGEV, so the network tends to maintain high-frequency information using small kernels. Moreover, the cost volume in Selective-IGEV faces an over-smooth problem [34], and that's why the attention map tends to increase the weights of large kernels to recover edge regions. In general, the CSA module shows different tendencies in different networks, which is a reflection of its adaptive ability.

**Number of iterations.** Our Selective-Stereo can achieve better performance with a smaller number of iterations. As shown in Tab. 3, our Selective-RAFT get the same performance with only 8 iterations compared to RAFT-Stereo [17], and for IGEV-Stereo [34], our Selective-IGEV also get a slight improvement with a few iterations. It shows that our modules can make secondary filtering to reduce noisy information from the initial cost volume.

**Size of convolution kernels.** We verify different kernel sizes on Scene Flow as shown in Tab. 4. At last, we choose the combination of $1 \times 1$ and $3 \times 3$ as our default config-

| Method | CSPN | LEAStereo | LaC + GANet | ACVNet | IGEV-Stereo | Selective-RAFT (Ours) | Selective-IGEV (Ours) |
|---|---|---|---|---|---|---|---|
| EPE (px) | 0.78 | 0.78 | 0.72 | 0.48 | 0.47 | 0.47 | **0.44** |

Table 5. Quantitative evaluation on Scene Flow test set.

| Method | KITTI 2012 | | | | | | KITTI 2015 | | | Run-time |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2-noc | 2-all | 3-noc | 3-all | EPE-noc | EPE-all | D1-bg | D1-fg | D1-all | (s) |
| AcfNet [42] | 1.83 | 2.35 | 1.17 | 1.54 | 0.5 | 0.5 | 1.51 | 3.80 | 1.89 | 0.48 |
| LEAStereo [9] | 1.90 | 2.39 | 1.13 | 1.45 | 0.5 | 0.5 | 1.40 | 2.91 | 1.65 | 0.30 |
| ACVNet [32] | 1.83 | 2.35 | 1.13 | 1.47 | 0.4 | 0.5 | 1.37 | 3.07 | 1.65 | 0.20 |
| RAFT-Stereo [17] | 1.92 | 2.42 | 1.30 | 1.66 | 0.4 | 0.5 | 1.58 | 3.05 | 1.82 | 0.38 |
| PCWNet [25] | 1.69 | 2.18 | **1.04** | **1.37** | 0.4 | 0.5 | 1.58 | 3.05 | 1.82 | 0.38 |
| LaC + GANet [18] | 1.72 | 2.26 | 1.05 | 1.42 | 0.4 | 0.5 | 1.44 | 2.83 | 1.67 | 1.80 |
| CREStereo [16] | 1.72 | 2.18 | 1.14 | 1.46 | 0.4 | 0.5 | 1.45 | 2.86 | 1.69 | 0.41 |
| IGEV-Stereo [34] | 1.71 | 2.17 | 1.12 | 1.44 | 0.4 | 0.4 | 1.38 | 2.67 | 1.59 | 0.18 |
| Selective-RAFT (Ours) | 1.64 | 2.09 | 1.10 | 1.43 | 0.4 | 0.5 | 1.41 | 2.71 | 1.63 | 0.45 |
| Selective-IGEV (Ours) | **1.59** | **2.05** | 1.07 | 1.38 | 0.4 | 0.4 | **1.33** | **2.61** | **1.55** | 0.24 |

Table 6. Quantitative evaluation on KITTI 2012 and KITTI 2015.

| Method | Edges | | Non-Edges | |
|---|---|---|---|---|
| | EPE | >1px | EPE | >1px |
| RAFT-Stereo [17] | 3.21 | 29.16 | 0.53 | 6.53 |
| Selective-RAFT | **2.40** | **21.63** | **0.40** | **4.65** |
| IGEV-Stereo [34] | 2.23 | 20.42 | 0.41 | 4.58 |
| Selective-IGEV | **2.18** | **20.01** | **0.38** | **4.35** |

Table 7. Quantitative evaluation on Scene Flow test set in different regions.

uration, because it achieves a competitive performance and reduces computational costs.

### 4.3. Comparisons with State-of-the-art

All fine-tuned models use the model pretrained on Scene Flow. Different target datasets use different finetune strategies. We validate two models called Selective-RAFT and Selective-IGEV using RAFT-Stereo [17] and IGEV-Stereo [34] as the baseline respectively.

**Scene Flow.** As shown in Tab. 5, we achieve a new state-of-the-art EPE of 0.44 on Scene Flow with Selective-IGEV, which surpasses LaC + GANet [18] by 38.89%. Besides, our Selective-RAFT also achieves a competitive EPE of 0.47 compared to IGEV-Stereo [34] with smaller parameters. To validate the ability to fuse information by regions of our modules, we then split Scene Flow test set into two regions: edge regions and non-edge regions using the Canny operator. As shown in Tab. 7, our Selective-RAFT outperforms RAFT-Stereo [17] by 25.23% and 24.53% in edge regions and non-edge regions. Due to IGEV-Stereo's aggregated cost volume [34], there's only a slight improvement in edge regions, but our Selective-IGEV still outperforms it by 7.32% in non-edge regions.

**KITTI.** We finetune our model on the mixed dataset of KITTI 2012 and KITTI 2015 with a batch size of 8 for 50k steps. Then we evaluate our Selective-Stereo on the test set of KITTI 2012 and KITTI 2015. As shown in Tab. 6, we achieve the best performance among all published methods for almost all metrics. On KITTI 2012, our Selective-RAFT outperforms RAFT-Stereo [17] by 14.58% and 13.64% on 2-noc and 2-all metrics, and our Selective-IGEV ranks $1^{st}$ on these metrics. On KITTI 2015, our Selective-RAFT outperforms RAFT-Stereo [17] by 10.44% on the D1-all metric, and our Selective-IGEV ranks $1^{st}$ on all metrics with only 16 iterations same as IGEV-Stereo [34]. As shown in Fig. 5, our Selective-IGEV outperforms IGEV-Stereo [34] in detailed and textureless regions.

**ETH3D.** Following CREStereo [16] and GMStereo [38], we use a collection of several public stereo datasets for training. The crop size is $384 \times 512$ and we first finetune the Scene Flow pretrained model on the mixed Tartan Air [29], CREStereo Dataset [16], Scene Flow [20], Sintel Stereo [3], InStereo2k [2] and ETH3D [24] datasets for 300k steps. Then we finetune it on the mixed CREStereo Dataset [16], InStereo2k [2] and ETH3D [24] datasets with for another 90k steps. As shown in Tab. 8, our Selective-RAFT outperforms RAFT-Stereo [17] by 17.90% on Bad 0.5 metric, and our Selective-IGEV achieves the best performance among all published methods for almost all metrics.

**Middlebury.** Also following CREStereo [16] and GMStereo [38], we first finetune the Scene Flow pretrained model on the mixed Tartan Air [29], CREStereo Dataset [16], Scene Flow [20], Falling Things [28], InStereo2k [2], CARLA HR-VS [40] and Middlebury [23] datasets using a crop size of $384 \times 512$ for 200k steps. Then we finetune it on the mixed CREStereo Dataset [16], Falling Things [28], InStereo2k [2], CARLA HR-VS [40] and Middlebury [23] datasets using a crop size of $384 \times 768$ with a batch size of 8 for another 100k steps. As shown in Tab. 8, our Selective-IGEV achieves the best performance among

| Method | ETH3D | | | | Middlebury | | | |
|---|---|---|---|---|---|---|---|---|
| | Bad 1.0 | Bad 0.5 | Bad 4.0 | AvgErr | Bad 2.0 | Bad 1.0 | Bad 4.0 | AvgErr |
| CroCo-Stereo [30] | 0.99 | 3.27 | 0.13 | 0.14 | 7.29 | 16.9 | 4.18 | 1.76 |
| GMStereo [38] | 1.83 | 5.94 | 0.08 | 0.19 | 7.14 | 23.6 | 2.96 | 1.31 |
| HITNet [26] | 2.79 | 7.83 | 0.19 | 0.20 | 6.46 | 13.3 | 3.81 | 1.71 |
| IGEV-Stereo [34] | 1.12 | 3.52 | 0.11 | 0.14 | 4.83 | 9.41 | 3.33 | 2.89 |
| RAFT-Stereo [17] | 2.44 | 7.04 | 0.15 | 0.18 | 4.74 | 9.37 | 2.75 | 1.27 |
| CREStereo [16] | **0.98** | 3.58 | 0.10 | 0.13 | 3.71 | 8.25 | 2.04 | 1.15 |
| EAI-Stereo [43] | 2.31 | 5.21 | 0.70 | 0.21 | 3.68 | 7.81 | 2.14 | 1.09 |
| DLNR [44] | - | - | - | - | 3.20 | 6.82 | 1.89 | 1.06 |
| Selective-RAFT (Ours) | 1.69 | 5.78 | 0.13 | 0.17 | - | - | - | - |
| Selective-IGEV (Ours) | 1.23 | **3.06** | **0.05** | **0.12** | **2.51** | **6.53** | **1.36** | **0.91** |

Table 8. Quantitative evaluation on ETH3D and Middlebury benchmarks. Note: Middlebury only allows one publish per paper, so we only publish our Selective-IGEV.

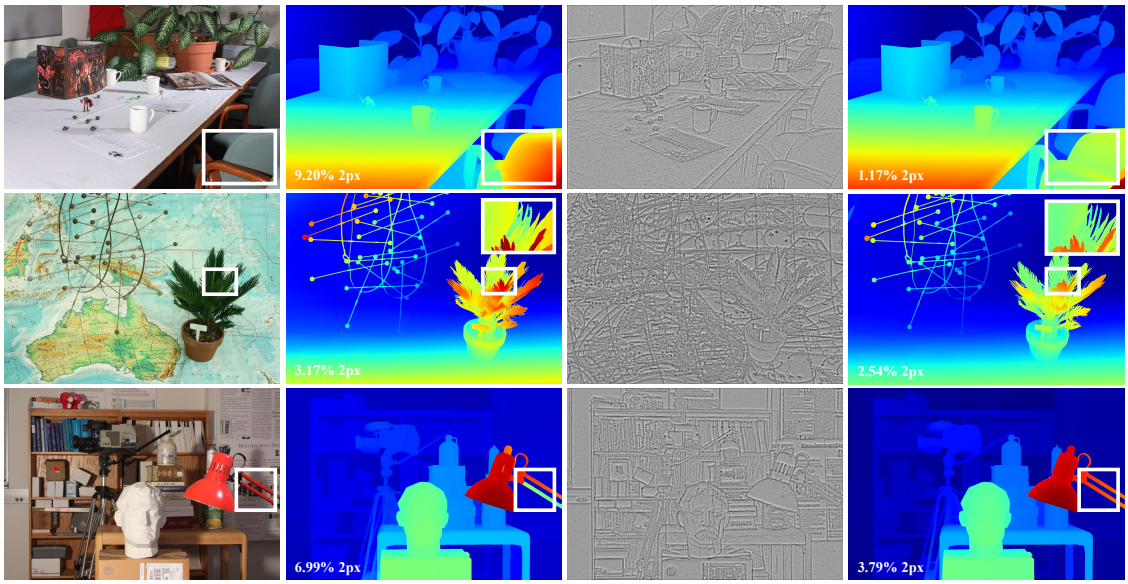



Figure 7. Qualitative results on the test set of Middlebury. The third column is the visualization of attention maps generated by the CSA module. Our Selective-IGEV outperforms IGEV in large textureless and thin object regions.

all published methods. As shown in Fig 7, compared to IGEV-Stereo [34], our Selective-IGEV performs better in textureless, and detailed regions. The third column in Fig 7 is the visualization of attention maps generated by the CSA module. It shows that attention maps can surely split regions that require information at different frequencies.

## 5. Conclusion

We propose Selective-Stereo, a novel iterative stereo matching method. The proposed Contextual Spatial Attention module and Selective Recurrent Unit help the network capture information at different frequencies for edge and smooth regions. Our Selective-Stereo ranks $1^{st}$ on KITTI, ETH3D, and Middlebury in almost all metrics among all published methods. It shows an ability to fuse information at different frequencies adaptively for edge and smooth regions with the help of attention maps extracted by CSA.

However, our method still faces some challenges. Firstly, although our method can fuse information adaptively using attention maps, the SRU's receptive field is still limited by predefined values. Secondly, adding branches or increasing the sizes of convolutional kernels leads to high memory and time costs, so we will explore the combination of lightweight convolutions and our method to reduce memory costs. Lastly, it's also a good direction to do research on the combination of convolutions and self-attention due to their different advantages and receptive fields.

**Acknowledgement.** This research is supported by National Natural Science Foundation of China (62122029, 62061160490, U20B200007).

## References


[1] André Araujo, Wade Norris, and Jack Sim. Computing receptive fields of convolutional neural networks. *Distill*, 4 (11):e21, 2019. 4

[2] Wei Bao, Wei 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*, 63:1–11, 2020. 7

[3] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012. 7

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 1, 2

[5] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3435–3444, 2019. 2, 3

[6] Yajie Chen, Xin Yang, and Xiang Bai. Confidence-weighted mutual supervision on dual networks for unsupervised cross-modality image segmentation. *Science China Information Sciences*, 66(11):210104, 2023. 3

[7] Junda Cheng, Xin Yang, Yuechuan Pu, and Peng Guo. Region separable stereo matching. *IEEE Transactions on Multimedia*, 2022. 1, 2

[8] Junda Cheng, Gangwei Xu, Peng Guo, and Xin Yang. Coatrsnet: Fully exploiting convolution and attention for stereo matching by region separation. *International Journal of Computer Vision*, pages 1–18, 2023. 2

[9] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, 33:22158–22169, 2020. 7

[10] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 2

[11] Miaojie Feng, Junda Cheng, Hao Jia, Longliang Liu, Gangwei Xu, and Xin Yang. Mc-stereo: Multi-peak lookup and cascade search range for stereo matching. *arXiv preprint arXiv:2311.02340*, 2023. 1

[12] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3599–3608. IEEE, 2019. 3

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1, 2, 5

[14] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3273–3282, 2019. 1, 2

[15] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. 1, 2

[16] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16263–16272, 2022. 1, 3, 7, 8

[17] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[18] Biyang Liu, Huimin Yu, and Yangqi Long. Local similarity pattern and cost self-reassembling for deep stereo matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1647–1655, 2022. 7

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[20] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 2, 5, 7

[21] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 1, 2, 5

[22] Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Fast vision transformers with hilo attention. *Advances in Neural Information Processing Systems*, 35:14541–14554, 2022. 3

[23] 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 *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 31–42. Springer, 2014. 1, 2, 5, 7

[24] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 1, 2, 5, 7

[25] Zhelun Shen, Yuchao Dai, Xibin Song, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. Pcw-net: Pyramid combination and warping cost volume for stereo matching. In *European Conference on Computer Vision*, pages 280–297. Springer, 2022. 2, 7

[26] 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. 8

[27] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2, 3

[28] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2038–2041, 2018. 7

[29] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020. 7

[30] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023. 8

[31] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 4

[32] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12981–12990, 2022. 1, 2, 7

[33] Gangwei Xu, Shujun Chen, Hao Jia, Miaojie Feng, and Xin Yang. Memory-efficient optical flow via radius-distribution orthogonal cost volume. *arXiv preprint arXiv:2312.03790*, 2023. 2

[34] 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. 1, 2, 3, 6, 7, 8

[35] 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*, 2023. 1, 2

[36] Gangwei Xu, Huan Zhou, and Xin Yang. Cgi-stereo: Accurate and real-time stereo matching via context and geometry interaction. *arXiv preprint arXiv:2301.02789*, 2023.

[37] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020. 2

[38] 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. 7, 8

[39] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1740–1749, 2020. 3

[40] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2019. 7

[41] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. 1, 2

[42] Youmin Zhang, Yimin Chen, Xiao Bai, Suihanjin Yu, Kun Yu, Zhiwei Li, and Kuiyuan Yang. Adaptive unimodal cost volume filtering for deep stereo matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12926–12934, 2020. 7

[43] 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. 8

[44] 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. 1, 2, 3, 5, 6, 8