IINet: Implicit Intra-inter Information Fusion for Real-Time Stereo Matching

Ximeng Li[1], **Chen Zhang**[1, 2], **Wanjuan Su**[1], **Wenbing Tao**[1,*]

[1]National Key Laboratory of Science and Technology on Multispectral Information Processing,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China
[2]Tuke Research
{ximengli, zhangchen_, suwanjuan, wenbingtao}@hust.edu.cn

Abstract

Recently, there has been a growing interest in 3D CNN-based stereo matching methods due to their remarkable accuracy. However, the high complexity of 3D convolution makes it challenging to strike a balance between accuracy and speed. Notably, explicit 3D volumes contain considerable redundancy. In this study, we delve into more compact 2D implicit network to eliminate redundancy and boost real-time performance. However, simply replacing explicit 3D networks with 2D implicit networks causes issues that can lead to performance degradation, including the loss of structural information, the quality decline of inter-image information, as well as the inaccurate regression caused by low-level features. To address these issues, we first integrate intra-image information to fuse with inter-image information, facilitating propagation guided by structural cues. Subsequently, we introduce the Fast Multi-scale Score Volume (FMSV) and Confidence Based Filtering (CBF) to efficiently acquire accurate multi-scale, noise-free inter-image information. Furthermore, combined with the Residual Context-aware Upsampler (RCU), our Intra-Inter Fusing network is meticulously designed to enhance information transmission on both feature-level and disparity-level, thereby enabling accurate and robust regression. Experimental results affirm the superiority of our network in terms of both speed and accuracy compared to all other fast methods.

Introduction

Stereo matching stands as a crucial technique for 3D perception. It leverages rectified binocular images to compute disparity through correspondence matching (Hirschmuller and Scharstein 2007). Deep stereo networks have exhibited significant potential with the aid of large training datasets (Mayer et al. 2016). While 3D CNN-based approaches (Chang and Chen 2018; Shen, Dai, and Rao 2021) and RNN-based methods (Lipson, Teed, and Deng 2021; Li et al. 2022) have demonstrate great accuracy, they come at the expense of substantial computational complexity. Consequently, various methods have been proposed to enhance real-time performance by expediting 3D cost aggregation. One class of methods (Xu et al. 2021) conducts aggregation at lower resolutions and super-resolves the outcomes. Another class (Gu et al. 2020; Yao et al. 2021) constructs cascading volumes and centers on reducing disparity samples at higher resolutions. Despite the promising temporal efficiency demonstrated by these techniques, striking a balance between accuracy and speed remains a challenge.

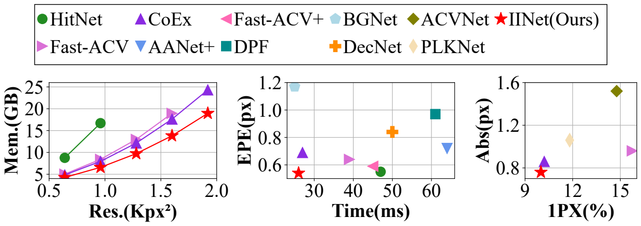


Figure 1: Left: Comparison of GPU memory cost as input resolution increases. While some fast methods fail with OOM, our method can process image at extremely high resolution. Mid: Comparison of EPE metric between fast methods on the SceneFlow. Our method achieve both best results and best efficiency. Right: D1 and Abs metric results on the Spring benchmark. Our method surpasses all other fast methods.

The challenge primarily arises from the retention of explicit 3D volumes and 3D convolution. While denser sampling and 3D aggregation at higher resolutions enhance accuracy, they introduce substantial time and memory burdens due to the $O(n^3)$ complexity of 3D convolution. Notably, explicit 3D volume contains huge amount of redundancy due to the sparsity of ground truth occupancy grids. To tackle redundancy and alleviate computational overhead, an intuitive approach is replacing 3D convolution with implicit 2D encoder-decoder network.

Nevertheless, the replacement of explicit 3D network with implicit 2D network causes potential issues. First, the 3D network serves the purpose of acquiring measurements (Guo et al. 2019) for local 3D structural patterns. Removal of the 3D network obstructs the acquisition of crucial structural information, leading to performance decline especially at occlusion or low-texture areas. Second, the 3D aggregation network plays a pivotal role in noise reduction and accuracy enhancement. Cost volumes without 3D aggregation tend to exhibit pronounced noise, making it a challenge to ef-

[*]Corresponding author

ficiently extract high-quality inter-image information. Last, the capacity of an implicit network diminishes as resolution increases, primarily due to low-level features and limited reception fields. Deep disparity features struggle to be transmitted effectively to higher resolutions, thereby hindering the achievement of robust and accurate regression. In this study, we employ various strategies to address these issues and enhance the performance. Consequently, we introduce a novel 2D implicit network as a substitute for the computationally intensive 3D network. Our approach not only significantly reduces time and memory costs but also demonstrates superior accuracy when compared to other fast methods.

To address the issue of performance decline resulting from the loss of structural information, we delve into intra-image context information. Numerous monocular studies (Ming et al. 2021) have established that intra information can regress relative depth by modeling structure like surfaces or edges. While intra information might lack reliability in terms of absolute spatial scale (Su and Ji 2022), it can help the propagation and rectification of inter-image information. Meanwhile, the restoration of the absolute scale can be achieved through inter information. Consequently, by fusing intra and inter information, we can enhance performance of 2D implicit network, particularly in low-texture and occlusion areas.

Simultaneously, the quality of inter information plays a critical role in providing accurate and noise-free scale information. To address this, we propose the Fast Multi-scale Score Volume (FMSV) and the Confidence Based Filtering (CBF). Based on the fact that well-textured regions often yield robust matches, our core idea revolves around boosting accuracy in well-matched regions and filtering out noise under the premise of high efficiency. The extraction of inter information is achieved through the construction of score volumes, serving as indicators of occupancy probability. By integrating the lightweight MLP with the top-k strategy, we efficiently generate multi-scale score volumes that contribute to heightened accuracy. Moreover, inspired by NP-CVP-MVSNet (Yang, Alvarez, and Liu 2022), we inject high-resolution disparity distributions into low-resolution volumes using a pretrain loss. Subsequently, the CBF is designed to filter noisy initial score volumes and prior disparities. In contrast to previous methods (Su, Xu, and Tao 2022) that employ networks to estimate aleatoric uncertainty, we directly employ average cross-entropy to estimate the uncertainty, guided by the insight that all cells corresponding to well-estimated pixels typically exhibit low entropy. Then confidences are generated with a linear transformation. By combining the two modules, we obtain accurate and noise-free inter information.

As to the challenge of implicit networks to regress robust and accurate disparities at high resolution, we tackle this by introducing the Intra-Inter Fusing (IIF) network which enhances information transmission at both feature-level and disparity-level. The IIF incorporates an U-Net (Ronneberger, Fischer, and Brox 2015) to efficiently fuse multi-scale intra-inter features. Additionally, drawing inspiration from RAFT (Teed and Deng 2020), we propose the Residual Context-aware Upsampler (RCU) to facilitate information transmission at disparity-level. The RCU leverages fused features and prior disparities to regress disparity residuals and up-sampling weights. Through cascaded RCUs, disparities from lower resolution are progressively refined with details and serve as priors for higher resolution. Benefiting from the compact network architecture, our method is capable of producing full-resolution disparities in real-time. As illustrated in Figure 1, our approach not only achieves SOTA accuracy but also excels in terms of efficiency among fast methods. Compared to Fast-ACVNet (Xu et al. 2023b), our network achieves a 17% improvement in accuracy on the SceneFlow, while reducing time costs by 33% and memory costs by 26%. Our contributions can be summarized as follows:

- We introduce a novel real-time stereo matching network that implicitly integrates intra and inter information, achieving remarkable performance in both accuracy and speed.
- We propose the Fast Multi-scale Score Volume (FMSV) and the Confidence Based Filtering (CBF) techniques, enabling efficient extraction of high-quality inter-image information.
- We present a compact Intra-Inter Fusing (IIF) network with the Residual Context-aware Upsampler (RCU) that produce robust and accurate disparities by enhancing information transmission at both feature and disparity level.

Related Works

Stereo Matching Stereo matching use rectified left and right images to regress disparities. Traditional methods define global energy functions, and algorithms like graph-cut (Hong and Chen 2004), dynamic programming (Ohta and Kanade 1985; Hirschmuller 2007), or random iteration (Bleyer, Rhemann, and Rother 2011) are employed to optimize the energy functions. Among deep learning approaches, PSMNet (Chang and Chen 2018) constructs concatenated volumes using siamese features, followed by the application of 3D convolutions to aggregate costs. GWC-Net (Guo et al. 2019) uses group-wise correlation to enhance features, CFNet (Shen, Dai, and Rao 2021) performs multi-scale fusion of cost volumes, GaNet (Zhang et al. 2019) adds explicit propagation step in aggregation, ACVNet (Xu et al. 2022a) utilizes pretrained models to filter out noise in the cost volumes. More recently, methods based on RNNs have emerged (Lipson, Teed, and Deng 2021). CRES (Li et al. 2022) improves pyramid construction, while IGEV (Xu et al. 2023a) introduces geometry volume for initialization. Another method STTR (Li et al. 2021) explores Transformer (Vaswani et al. 2017) architecture. Although these methods achieve great performance, they come with significant time overhead.

Implicit 3D Network Recently, with the popularity of techniques like NeRF (Mildenhall et al. 2021) and Transformer, implicit 3D methods have been widely studied. NeuS (Wang et al. 2021) leverages positional encoding and MLPs to reconstruct surfaces. SMDNet (Tosi et al.

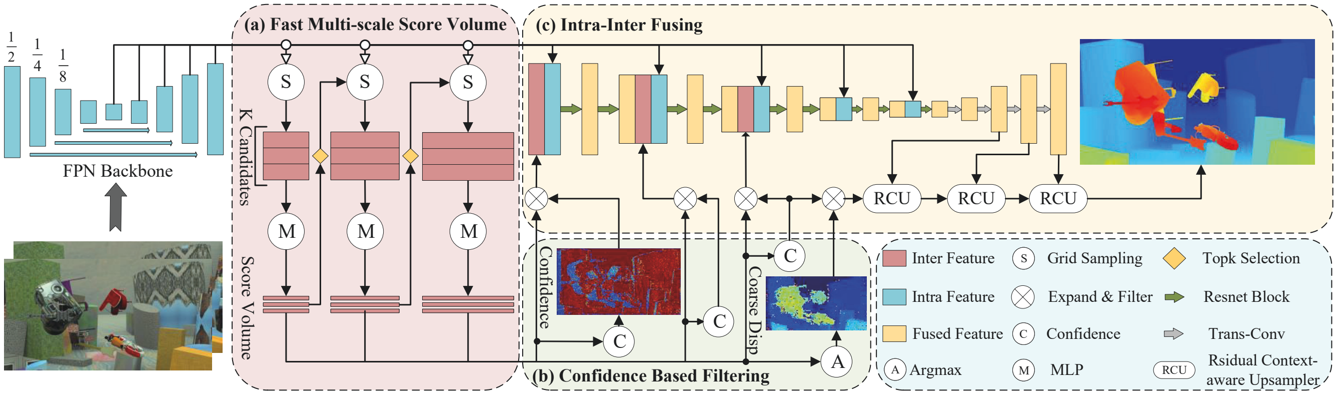


Figure 2: Overall architecture of IINet. (a) Our Fast Multi-scale Score Volume which efficiently extract inter-image information from stereo pairs. (b) The procedure of CBF. We first use score volumes to estimate confidence maps, then use them to filter noisy score volumes and coarse disparities. (c) The Intra-Inter Fusing module which use CNN encoder to fuse features and Residual Context-aware Upsamplers to refine output disparities. Rectangles of different sizes are used to represent features at various resolutions. While we use a bus to illustrate data flow for simplicity, the connections are separated across resolutions.

2021) directly regresses disparities from cost volume with MLPs. MVS2D (Yang et al. 2022) employs attention mechanisms to aggregate multi-view information. IIB (Yifan et al. 2022) employs the general Transformer architecture PerceiverIO (Jaegle et al. 2022) to regress depth. SimpleRecon (Sayed et al. 2022) uses pretrained image-encoder to integrate cost volume information. Implicit network helps eliminate the redundancy and improve feature with non-local information.

Context Information Context information has been extensively explored in monocular depth estimation (Miangoleh et al. 2021). It can regress relative depth by capturing the relationships between objects (Ming et al. 2021). Various techniques, including Transformer (Ranftl, Bochkovskiy, and Koltun 2021), MLP (Yuan et al. 2022), and diffusion (Duan, Guo, and Zhu 2023), have been employed to explore contextual information. In the field of correspondence matching, RAFT (Teed and Deng 2020) introduces an additional context encoder to assist the RNN network, while GMFlow (Xu et al. 2022b) employs cross-attention to extract contextual information. EP-MVSNet (Su and Tao 2023) uses context to guide edge-preserving depth map up-sampling. Recently, stereo matching methods have also leverage context information. COEX (Bangunharcana et al. 2021) utilizes contextual features to guide aggregation, while CSTTR (Guo et al. 2022) proposes a context branch to assist the attention module.

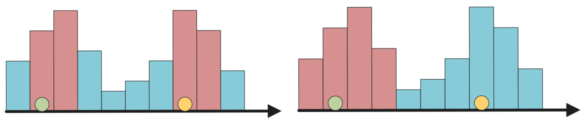
Method

The comprehensive architecture is illustrated in Figure 2. Given a pair of image, our process begins by extracting multi-scale context features. Subsequently, the Fast Multi-scale Score Volume (FMSV) depicted in Figure 2 (a) efficiently captures multi-scale inter-image information. A pretrain loss is employed to enhance the details. The Confidence Based Filtering module in Figure 2 (b) is devised to eliminate noise in both the score volumes and the coarse disparities. Furthermore, the Intra-Inter Fusing module illustrated in Figure 2 (c) applies an encoder-decoder architecture to fuse information. The features and priors traverse through cascaded Residual Context-aware Upsampler modules, regressing accurate, high-resolution disparities.

Fast Multi-scale Score Volume

Context Feature Extraction We employ MobileNetV3-Large (Howard et al. 2019) with pretrained weights and an FPN architecture as our backbone. The inputs consist of the original RGB image pairs, designated as I_L and I_R, while the outputs constitute a collection of multi-scale features F_L^i, F_R^i, where $i = 1, \ldots, n$, with 1 denoting the highest scale (1/2 resolution in our approach).

Multi-scale Score Volume Utilizing the siamese features F, we create a multi-scale score volume. The features of the right image are warped to the left image using the target disparity D_d. For a maximum disparity of 192, the number of target disparities is 24, 6, and 4 for 1/8, 1/4 and 1/2 resolutions, respectively. Then, the dot productions and patch features are concatenated to establish the cost volume. Instead of relying on 3D convolutions, we employ a lightweight MLP network for feature dimension reduction.



(a) Top-k candidates (b) Argmax candidates

Figure 3: Comparison between candidate selection strategies, e.g., four candidates in this figure, when probability distributions are ambiguous at low-res discontinue regions. Assuming GT disparities of adjacent pixels position in sections marked with circle, (a) top-k strategy selects candidates for both pixels but (b) argmax strategy only covers a single pixel, which leads to accumulative errors

The MLP compresses the dimensions of the cost volumes from $D^i \times C \times H^i \times W^i$ to $D^i \times H^i \times W^i$, generating the initial score volumes $C^i, i = 1, \ldots, L$. The weights of MLPs are shared across scales to enhance consistency. The process is formulated by:

$$C_d^i = \text{MLP}\left(\text{Cat}\left(F_L^i(x), F_R^i(x - D_d^i), F_L^i \odot F_R^i\right)\right), \quad (1)$$

where $\odot$ represents for dot product, and Cat indicates concatenating. Multi-scale volumes are efficiently constructed with top-k strategy to capture accurate inter information. Comparison between top-k strategy and argmax upsampling is shown in Figure 3. Disparity distributions tend to be ambiguous at edges due to the down-sampling process. The top-k strategy retains candidates with long distance, thus mitigating the issue of accumulative errors arising from unimodal distribution assumptions. The top-k score values are chosen, and the corresponding disparities serve as seeds. Each seed is expanded into two target disparities at the higher resolution, yielding D^{i-1}. This process begins at 1/8 resolution and is repeated until reaching 1/2 resolution.

Pretrain Loss Beyond the architecture, we incorporate a pretrain loss to guide the score volume's learning process. Drawing inspiration from NP-CVP-MVSNet (Yang, Alvarez, and Liu 2022), we employ the soft histogram distribution of ground truth (GT) disparities as supervision instead of which obtained through down-sampling. The distinction between these two operations is illustrated in Figure 4. For a patch at full resolution in 4(a), soft down-sampling injects hi-res patch information into low resolution volume through histogram. In contrast, conventional nearest down-sampling in Figure 4(c) loses high-frequency information. Specifically, the histogram volume is generated by folding the patch Φ_p at full resolution which correspond to a low-res pixel p. The distances between the GT values $GT_{p'}$ and each sampled disparity in D^i are computed for each pixel. These distances are truncated and summed within patches to form $dist_d^i(p)$ for each sampled disparity. This procedure can be seen as splatting hi-res GT values into low-res volumes. Finally, the histogram volume are obtained by normalizing volumes across the d dimension.

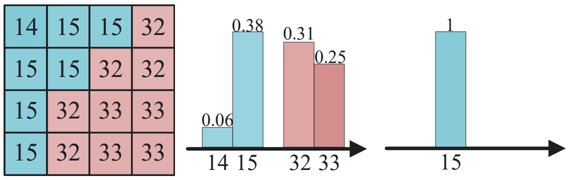
$$dist_d^i(p) = \sum_{p' \in \Phi_p} \begin{cases} 1 - \delta(p'), & \delta(p') < 1 \\ 0, & \delta(p') \geq 1 \end{cases},$$

$$\delta(p') = |GT_{p'} - D_d^i|, \quad (2)$$

$$Hist_d^i = \frac{dist_d^i}{\sum_{d=1}^{d=n(D^i)} dist_d^i}, \quad (3)$$

where $n(\cdot)$ denotes the count of a set. After obtaining the histogram volume $Hist_d^i$, a mask M^i is generated to mark the pixels whose volume cover GT disparities. We use the Kullback-Leibler (KL) divergence loss $\mathcal{L}_{KL}$ to enforce consistency between the estimated score volume and the histogram volume.

$$\mathcal{L}_{\text{KL}}^i = \text{KLloss}\left(\text{Softmax}\left(C^i\right), Hist^i\right) \cdot M^i. \quad (4)$$



(a) Full-res Disp. (b) Soft Down. (c) Nearest Down.

Figure 4: Difference between soft and nearest down-sampling. (a) Disparity distribution of a 4 × 4 patch at full resolution (b) Soft down-sampling utilizes the redundancy in low-res volume to store the histogram distribution for a hi-res patch, while (c) Nearest down-sampling directly use the center pixel's value as low-res ground truth. As a result, soft down-sampling contains more high frequency details and avoids misaligned supervision at lower resolution.

The purpose of the score volume is to represent the probability that a grid cell covers GT. However, the KL loss imposes constraints solely on the histogram distribution, which can lead to diminished response within cells. Thus, we propose an additional binary cross-entropy (BCE) loss, denoted as $\mathcal{L}_{BCE}$. Firstly, we employ the sigmoid function to map the score volumes to probability volumes, yielding P_d^i. Subsequently, we designate all grid cells covering the GT as positive samples, and the remaining cells as negative samples. Given that a majority of positive samples tend to cluster around a few disparity values, we use the Focal loss to balance the quantity differences. Additionally, we utilize $Hist^i$ as weights for positive samples to model visibility. The BCE loss is formulated as follows:

$$\mathcal{L}_{\text{BCE}}^i = -\sum_d Hist_d^i \alpha_t \left(1 - P_{d,t}^i\right)^\gamma \log\left(P_{d,t}^i\right) \cdot M^i, \quad (5)$$

$$P_d^i = \text{sigmoid}(C_d^i). \quad (6)$$

Our final pretrain loss is weighted sum of the two items:

$$\mathcal{L}_{\text{pre}} = \sum_{i=1}^{L} \lambda_i \left(\mathcal{L}_{\text{KL}}^i + \mathcal{L}_{\text{BCE}}^i\right). \quad (7)$$

We use the pretrain loss to pretrain our FMSV module to provide priors, then drop the loss and use the total loss to train the whole network end to end.

Confidence Based Filtering

We tackle noise by assessing confidence. The confidence is achieved by uncertainty estimation. Cross-entropy is utilized to represent the uncertainty of each grid cell. Then, to measure the uncertainty of a pixel, we calculate the average cross-entropy across the d dimension. The process can be formulated as:

$$U^i = \frac{1}{n(D^i)} \sum_{d=1}^{n(D^i)} \left(-P_d^i \log\left(p_d^i\right) - \overline{P_d^i} \log\left(\overline{P_d^i}\right)\right), \quad (8)$$

where $\overline{P}$ represents $1 - P$. U^i represents for uncertainty map at each resolution. By computing the average noise, this uncertainty quantifies the noise level of a pixel. We further utilize a linear transformation to convert the uncertainty into confidence $Conf^i$ for each resolution:

$$Conf^i = a\left(U^i - b\right) \cdot \max_d(P_d^i), \quad (9)$$

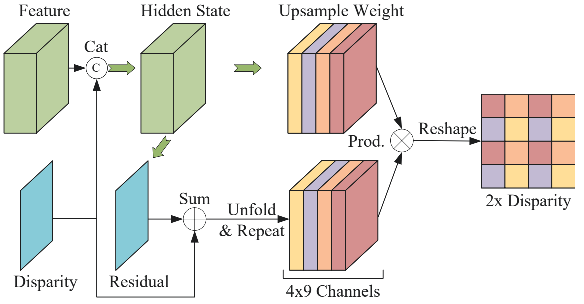


Figure 5: Illustration of a single residual context-aware upsampler module. We use same color to mark channels that correspond to the 2× disparity map after reshaping. The module is repeated for three times to output full resolution disparity.

where $b = -0.5\log(0.5)$ represents the maximum uncertainty, and a is a coefficient that makes confidence reach 1 when $P \geq P_{th}$. To prevent excessive filtering that could lead to sparsity, we set P_{th} to 0.9 and limit the maximum value to 1. The confidence estimation module generates low confidence values for occluded and low-texture regions. For the former, these regions exhibit low uncertainty alongside low probability. For the latter, these regions feature high probability paired with high uncertainty. By multiplying the estimated confidences with the score volumes, the filtered score volumes C_F^i are derived. The prior disparity map $disp_P^3$ is obtained by excluding regions with low confidence. This procedure is formulated as:

$$C_F^i = C^i \cdot Conf^i, \tag{10}$$

$$disp_P^3 = \text{argmax}_d(P^3) \cdot \left(Conf^3 > th\right), \tag{11}$$

where $\cdot$ represents for the Hadamard product and th represents for a threshold.

Intra-Inter Fusing Module

The Intra-Inter Fusing module employs an U-Net (Ronneberger, Fischer, and Brox 2015) for information fusion. In the decoder part, features and priors are passed through cascaded RCUs to regress full-resolution disparities.

Intra-Inter Encoder The inter features after filtering have holes in invalid areas. To address this, we introduce intra information to facilitate the propagation of the score volumes with local context. For efficiency, we employ a simple 2D convolution network to fuse the intra feature F_L^i with the filtered inter feature C_F^i. The multi-scale fusion network is established in a top-down manner to yield the fused features F_{fuse}^i.

$$F_{\text{fuse}}^i = \text{Conv}\left(\text{Cat}\left(C_F^i, F_L^i, \text{DsConv}\left(F_{\text{fuse}}^{i-1}\right)\right)\right). \tag{12}$$

Residual Context-aware Upsampler Subsequently, we utilize the fused features to implicitly predict multi-scale disparities. We employ latent variables to predict weights and disparity residuals. Instead of iterative optimization at the same resolution, we combine up-sampling with residual optimization, introducing high-frequency features while reducing computational overloads. We select 1/8 resolution as the lowest resolution for regression. The fused features F_{fuse}^i undergo an initial update through a U-Net (Ronneberger, Fischer, and Brox 2015). Subsequently, as depicted in Figure 5, these fused features are concatenated with the prior disparities $disp_P^i$ and encoded into latent variables H^i. Leveraging these latent variables H^i, we predict disparity residuals $\Delta disp^i$ and up-sampling weights W^i. These weights are $4{\times}9$ vectors that up-sample the disparity map by a factor of 2 using 9 neighboring pixels. The disparities are optimized using the residuals, resulting in $disp^i$, which are then up-sampled using W^i to generate priors $disp_P^{i-1}$ for higher resolutions. The process can be formulated as:

$$H^i = \text{Conv}\left(\text{Cat}\left(F_{\text{fuse}}^i, disp_P^i\right)\right). \tag{13}$$

$$W^i = \text{Softmax}\left(\text{Conv}\left(H^i\right)\right). \tag{14}$$

$$\Delta disp^i = \text{Conv}\left(H^i\right). \tag{15}$$

$$disp^i = disp_P^i + \Delta disp^i. \tag{16}$$

$$disp_P^{i-1} = \text{Unfold}(W^i) \cdot disp^i. \tag{17}$$

Finally, the lightweight decoding network outputs full-resolution disparity through cascaded Residual Context-aware Upsamplers.

Loss Function

L1 Loss We employ a loss function similar to SimpleRecon (Sayed et al. 2022). Our multi-scale disparity loss is constructed using L1 loss due to its stability. The multi-scale ground truth disparity G^i is derived through nearest interpolation. The L1 loss can be expressed as:

$$\mathcal{L}_{\text{depth}} = \sum_{i=0}^{3} \lambda_{i,disp} \left| disp^i - G^i \right|. \tag{18}$$

Grad and Normal Loss To ensure the smoothness of the disparity surface, we establish a multi-scale gradient loss and a normal loss at the full resolution. The gradient and normal losses can be defined as follows:

$$\mathcal{L}_{\text{grad}} = \sum_{i=0}^{3} \lambda_{i,grad} \left| \nabla disp^i - \nabla G^i \right|. \tag{19}$$

$$\mathcal{L}_{\text{normal}} = \lambda_{normal} \frac{1}{HW} \sum_{i,j} 1 - \hat{\mathbf{N}}_{i,j} \cdot \mathbf{N}_{i,j}. \tag{20}$$

Total Loss Our final end to end training loss is weighted sum of the three losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{grad}} + \mathcal{L}_{\text{normal}}. \tag{21}$$

Experiments

Datasets We use four datasets in our experiments. The SceneFlow (Mayer et al. 2016) is a large synthetic dataset consisting of 35, 454 training pairs and 4, 370 testing pairs. The Spring (Mehl et al. 2023), a novel large-scale dataset rendered from a Blender movie, includes 5,000 training pairs and 1,000 testing pairs, characterized by high resolution and high details. The KITTI (Geiger, Lenz, and Urtasun 2012), a real-world benchmark of driving scenarios,

Method	EPE (px)	D1 (%)	3px (%)	Time (ms)
DPF (Duggal et al. 2019)	0.97	-	-	61
AANet (Xu and Zhang 2020)	0.89	-	-	62
BGNet+ (Xu et al. 2021)	1.17	-	-	32
CoEx (Bangunharcana et al. 2021)	0.69	-	4.00	27
FastACV (Xu et al. 2023b)	0.64	2.49	2.82	39
HitNet (Tankovich et al. 2021)	0.55	2.26	**2.69**	47
IINet (ours)	**0.54**	**2.18**	2.73	**26**

Table 1: Results on the SceneFlow. We compare our results with recent fast or real-time methods. Time metric refers to the GPU latency.

is divided into KITTI12 and KITTI15 subsets, comprising 394 training pairs and 395 testing pairs. Lastly, the Middlebury2014 (Scharstein et al. 2014) is created from real-world images, we use its training set which contains 15 image pairs to test our generalization ability.

Metrics EPE (End-point Error) or Absolute Error refers to the absolute difference between the estimated and ground truth disparities. It serves as a measure of the overall numerical accuracy of the estimation. The N-px Error is a percentage error metric that calculates the proportion of pixels for which the EPE exceeds N pixels. The D1 error relaxes the 3px Error restriction by excluding pixels for which the Relative Absolute Error is less than 5 %. These two percentage error metrics evaluate the structural accuracy of the estimation.

Implementation Our model is developed using PyTorch and executed on an RTX 3090 GPU. For pretraining on the SceneFlow, we incorporate image augmentation with a probability of 0.2. The augmentation includes color adjustments, brightness enhancements, and contrast modifications to simulate varying exposure conditions. Furthermore, random vertical and rotation shifts are applied to the right image, followed by random cropping to resolution of 512×384. We employ the Adam optimizer and train for 105 epochs. The first 13 epochs use the pretrain loss to train the FMSV. Subsequently, the full network is trained using the total loss for the remaining epochs. The initial learning rate is set to 0.001, then reduced to 0.0001 at epoch 10, and halved at epochs 50, 70, 85, 95, and 100. For the Spring, we fine-tune the pretrained model using the training set for 65 epochs. The initial learning rate is set to 0.00005 and reduced by half at epochs 40 and 55.

Evaluation on SceneFlow

We first evaluate our method on the SceneFlow, the results are shown in Table 1. Among all the fast methods, ours achieve the best accuracy. Although our method's time cost is comparable to BGNet (Xu et al. 2021) and CoEx (Bangunharcana et al. 2021), it significantly outperforms them in terms of EPE and pixel-based metrics. Our accuracy performance surpasses that of Fast-ACVNet (Xu et al. 2023b) and closely aligns with HitNet (Tankovich et al. 2021), all while reducing computational time by 30% and 44%, respectively. The experimental results on SceneFlow demonstrate that our method can well balance efficiency and performance. Leveraging purely 2D convolution networks, our approach achieves multi-scale fusion, enabling the learning of both high-frequency and low-frequency components. By transmitting information at both feature-level and disparity-level with skip connection, our method progressively yields robust, fine-detail disparity maps.

Method	Fast	1px total(%)	D1 (%)	Abs (px)
FastACV (Xu et al. 2023b)	✓	15.75	4.26	0.96
ACVNet (Xu et al. 2022a)		14.77	5.35	1.52
CoEx (Bangunharcana et al. 2021)	✓	10.21	3.96	0.86
IINet (ours)	✓	10	3.78	0.76
CroCo (Weinzaepfel et al. 2023)		**7.13**	**2.71**	**0.47**

Table 2: Comparative results on the Spring Benchmark. We mark fast methods using ✓. The Abs metric equals to the EPE metric.

Evaluation on Spring

We proceed to evaluate our method on the Spring dataset, and the benchmark results are presented in Table 2. Our method has now rank the second on the leaderboard. Notably, the Spring benchmark is relatively new, leading to the inclusion of only recent methods in the comparison. To compare with other fast methods, we fine-tune CoEx (Bangunharcana et al. 2021) and Fast-ACVNet (Xu et al. 2023b) on the train set, utilizing their pretrained models from the SceneFlow. Subsequently, we submit the results to the Spring benchmark. These outcomes underscore the effectiveness and efficiency of our pipeline. Even without the use of Transformer architecture and extensive pretraining like CroCo-Stereo (Weinzaepfel et al. 2023), our approach still achieves commendable accuracy on high-detail datasets. As demonstrated in Figure 6, our implicit structural information integrated with contextual details contributes to stability in smooth regions. Meanwhile, our high-resolution decoder

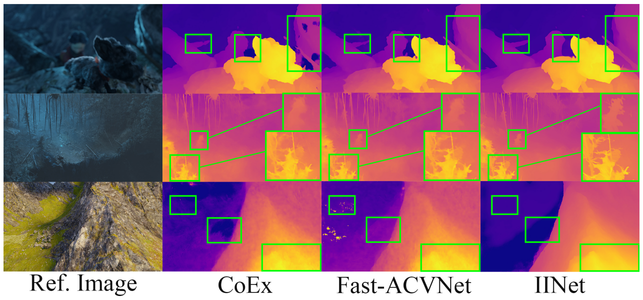


Figure 6: Visual results of hero frames from the Spring benchmark. Our network outperform other fast methods not only in high frequency details but also in low frequency smooth regions.

Method	KITTI12 D1(%)	KITTI15 D1(%)	Middle 2px(%)
DPF (Duggal et al. 2019)	16.8	15.9	30.83
BGNet (Xu et al. 2021)	24.8	20.1	37
CoEx (Bangunharcana et al. 2021)	13.5	11.6	25.51
FastACV (Xu et al. 2023b)	12.4	10.6	20.13
IINet (ours)	**11.6**	**8.5**	**19.57**

Table 3: Generalization performance on multiple datasets. The model is only trained on the SceneFlow.

enhances accuracy in intricate areas such as grass or hair, contributing to our advantages in 1px metric over other fast methods.

Generalization Ability

We proceed to assess the generalization capability of our model on the KITTI and Middlebury datasets. The corresponding results are presented in Table 3. The results affirm the good generalization capability of our model to real-world datasets. Our approach integrates a robust module based on confidence and leverages both inter and intra-image information for prediction. Consequently, our model demonstrates superior generalization ability. The results underscores the potential of our model for real-world applications.

Ablation Study

We conduct ablation studies on our network by systematically removing these components from the full model. These experiments are carried out using the SceneFlow with the same training parameters. The experimental results are presented in Table 4.

Intra Information We initiate our ablation study by investigating the impact of fusing inter and intra information. In this process, we remove the intra feature from the encoder. Notably, the pixel error metric experiences a huge degradation. We speculate that the inclusion of intra information aids in reconstructing low-confidence regions. Conversely, the network continues to exhibit good performance in terms of the EPE metric. This outcome highlights the efficacy of the implicit architecture for subpixel regression.

Residual Context-aware Upsampler For the ablation study on the RCU, we make two modifications to the components. Firstly, we substitute the context-aware upsampler with nearest interpolation, and this yields a considerable drop in performance. We speculate the high-resolution decoding network plays a role in regressing small residuals. It struggles to rectify large errors in this setting. Employing nearest interpolation directly would lead to the accumulative errors. Secondly, we replace the residual disparity optimization with independent regression, which further exacerbates the performance degradation. Our speculation is rooted in the fact that low-level features have limited receptive fields. Directly regressing disparities is more susceptible to texture influences, resulting in the emergence of artifacts.

Module	EPE (px)	D1 (%)	3px (%)	Time (ms)
Full	0.54	2.18	2.73	26
Ablation on Intra Information				
w\o Intra	0.62	2.54	3.13	24
Ablation on RCU and Filter				
w\o CU	0.69	2.72	3.27	25
w\o CU, F	0.73	2.82	3.42	25
w\o RCU	0.74	2.83	3.45	25
Ablation on FMSV				
MSV3	0.74	2.83	3.45	25
MSV2	0.78	2.9	3.53	21
MSV1	0.84	3.03	3.7	19

Table 4: Ablation study results. CU represents for context-aware upsampler. F represents for the confidence based filtering. MSV represents for multi-scale score volume, the number indicates the scales the volumes are constructed on.

Confidence Based Filtering We proceed the ablation study on the confidence based filtering module. Removing this module causes a decline in performance. Upon observing the uncertainty maps, we find that low-confidence regions are predominantly concentrated in the central areas of occlusions and low-texture surfaces. Our speculation is that noisy score volumes introduce randomness during the encoding phase. Subsequently, these noisy coarse disparities directly impede the efficacy of the residual decoding module. Thus, the application of a filtering operation becomes important in rectifying these areas.

Multi-scale Score Volume Lastly, we perform ablation on the FMSV by progressively diminishing the scale of score volumes, starting from 1/2, then to 1/4, and finally down only volume at 1/8 resolution is constructed. As the levels reduced, the results consistently degrade, with the EPE metric showing a more pronounced decline. This trend underscores the essential role of inter information for accurate disparity prediction. It is impractical to discard the score volumes and rely solely on the context features for disparity prediction, as demonstrated by empirical evidence.

Conclusion

We introduce a novel real-time stereo network based on efficient implicit 2D network instead of intricate 3D CNN. To counter degradation, we adopt a multi-pronged approach. Initially, we incorporate intra-information to facilitate information propagation. Subsequently, the Fast Multi-scale Score Volume and Confidence Based Filtering modules are developed to enhance the quality of inter-information. Furthermore, we elevate the accuracy and resilience of the 2D network by enhancing fusion at both feature and disparity levels. As a result, our network strikes a well balance between accuracy and efficiency.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 62176096.

References

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

Bleyer, M.; Rhemann, C.; and Rother, C. 2011. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, 1–11.

Chang, J.-R.; and Chen, Y.-S. 2018. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5410–5418.

Duan, Y.; Guo, X.; and Zhu, Z. 2023. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. arXiv:2303.05021.

Duggal, S.; Wang, S.; Ma, W.-C.; Hu, R.; and Urtasun, R. 2019. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4384–4393.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; and Tan, P. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2495–2504.

Guo, W.; Li, Z.; Yang, Y.; Wang, Z.; Taylor, R. H.; Unberath, M.; Yuille, A.; and Li, Y. 2022. Context-enhanced stereo transformer. In *European Conference on Computer Vision*, 263–279. Springer.

Guo, X.; Yang, K.; Yang, W.; Wang, X.; and Li, H. 2019. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3273–3282.

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

Hirschmuller, H.; and Scharstein, D. 2007. Evaluation of cost functions for stereo matching. In *2007 IEEE conference on computer vision and pattern recognition*, 1–8. IEEE.

Hong, L.; and Chen, G. 2004. Segment-based stereo matching using graph cuts. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, I–I. IEEE.

Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1314–1324.

Jaegle, A.; Borgeaud, S.; Alayrac, J.; Doersch, C.; Ionescu, C.; Ding, D.; Koppula, S.; Zoran, D.; Brock, A.; Shelhamer, E.; Hénaff, O. J.; Botvinick, M. M.; Zisserman, A.; Vinyals, O.; and Carreira, J. 2022. Perceiver IO: A General Architecture for Structured Inputs & Outputs. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Li, J.; Wang, P.; Xiong, P.; Cai, T.; Yan, Z.; Yang, L.; Liu, J.; Fan, H.; and Liu, S. 2022. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16263–16272.

Li, Z.; Liu, X.; Drenkow, N.; Ding, A.; Creighton, F. X.; Taylor, R. H.; and Unberath, M. 2021. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6197–6206.

Lipson, L.; Teed, Z.; and Deng, J. 2021. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, 218–227. IEEE.

Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; and Brox, T. 2016. 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*, 4040–4048.

Mehl, L.; Schmalfuss, J.; Jahedi, A.; Nalivayko, Y.; and Bruhn, A. 2023. Spring: A High-Resolution High-Detail Dataset and Benchmark for Scene Flow, Optical Flow and Stereo. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Miangoleh, S. M. H.; Dille, S.; Mai, L.; Paris, S.; and Aksoy, Y. 2021. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9685–9694.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Ming, Y.; Meng, X.; Fan, C.; and Yu, H. 2021. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438: 14–33.

Ohta, Y.; and Kanade, T. 1985. Stereo by intra-and interscanline search using dynamic programming. *IEEE Transactions on pattern analysis and machine intelligence*, (2): 139–154.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12179–12188.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Sayed, M.; Gibson, J.; Watson, J.; Prisacariu, V.; Firman, M.; and Godard, C. 2022. SimpleRecon: 3D reconstruction without 3D convolutions. In *European Conference on Computer Vision*, 1–19. Springer.

Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; and Westling, P. 2014. 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*, 31–42. Springer.

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

Su, Q.; and Ji, S. 2022. Chitransformer: Towards reliable stereo from cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1939–1949.

Su, W.; and Tao, W. 2023. Efficient Edge-Preserving Multi-View Stereo Network for Depth Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2348–2356.

Su, W.; Xu, Q.; and Tao, W. 2022. Uncertainty guided multi-view stereo network for depth estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7796–7808.

Tankovich, V.; Hane, C.; Zhang, Y.; Kowdle, A.; Fanello, S.; and Bouaziz, S. 2021. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14362–14372.

Teed, Z.; and Deng, J. 2020. 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*, 402–419. Springer.

Tosi, F.; Liao, Y.; Schmitt, C.; and Geiger, A. 2021. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8942–8952.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 27171–27183.

Weinzaepfel, P.; Lucas, T.; Leroy, V.; Cabon, Y.; Arora, V.; Brégier, R.; Csurka, G.; Antsfeld, L.; Chidlovskii, B.; and Revaud, J. 2023. 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*, 17969–17980.

Xu, B.; Xu, Y.; Yang, X.; Jia, W.; and Guo, Y. 2021. Bilateral grid learning for stereo matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12497–12506.

Xu, G.; Cheng, J.; Guo, P.; and Yang, X. 2022a. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12981–12990.

Xu, G.; Wang, X.; Ding, X.; and Yang, X. 2023a. Iterative Geometry Encoding Volume for Stereo Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21919–21928.

Xu, G.; Wang, Y.; Cheng, J.; Tang, J.; and Yang, X. 2023b. Accurate and Efficient Stereo Matching via Attention Concatenation Volume. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–13.

Xu, H.; and Zhang, J. 2020. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1959–1968.

Xu, H.; Zhang, J.; Cai, J.; Rezatofighi, H.; and Tao, D. 2022b. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8121–8130.

Yang, J.; Alvarez, J. M.; and Liu, M. 2022. Non-parametric depth distribution modelling based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8626–8634.

Yang, Z.; Ren, Z.; Shan, Q.; and Huang, Q. 2022. Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8574–8584.

Yao, C.; Jia, Y.; Di, H.; Li, P.; and Wu, Y. 2021. A decomposition model for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6091–6100.

Yifan, W.; Doersch, C.; Arandjelović, R.; Carreira, J.; and Zisserman, A. 2022. Input-level inductive biases for 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6176–6186.

Yuan, W.; Gu, X.; Dai, Z.; Zhu, S.; and Tan, P. 2022. Neural Window Fully-connected CRFs for Monocular Depth Estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3906–3915.

Zhang, F.; Prisacariu, V.; Yang, R.; and Torr, P. H. 2019. Ganet: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 185–194.