UASNet: Uncertainty Adaptive Sampling Network for Deep Stereo Matching

Yamin Mao[1] Zhihua Liu[1] Weiming Li[1] Yuchao Dai[2]
Qiang Wang[1] Yun-Tae Kim[3] Hong-Seok Lee[3]
[1]Samsung Research Center of Beijing [2] Northwestern Polytechnical University
[3]Samsung Advanced Institute of Technology

Abstract

Recent studies have shown that cascade cost volume can play a vital role in deep stereo matching to achieve high resolution depth map with efficient hardware usage. However, how to construct good cascade volume as well as effective sampling for them are still under in-depth study. Previous cascade-based methods usually perform uniform sampling in a predicted disparity range based on variance, which easily misses the ground truth disparity and decreases disparity map accuracy. In this paper, we propose an uncertainty adaptive sampling network (UASNet) featuring two modules: an uncertainty distribution-guided range prediction (URP) model and an uncertainty-based disparity sampler (UDS) module. The URP explores the more discriminative uncertainty distribution to handle the complex matching ambiguities and to improve disparity range prediction. The UDS adaptively adjusts sampling interval to localize disparity with improved accuracy. With the proposed modules, our UASNet learns to construct cascade cost volume and predict full-resolution disparity map directly. Extensive experiments show that the proposed method achieves the highest ground truth covering ratio compared with other cascade cost volume based stereo matching methods. Our method also achieves top performance on both SceneFlow dataset and KITTI benchmark.

1. Introduction

Inferring disparity (or depth) from stereo images is a fundamental task in many applications, such as robotics [19], autonomous driving [20] and augmented reality [29]. State-of-the-art stereo matching algorithms can be divided into two categories according to their cost volume representation. One is to construct 4D cost volume throughout the entire disparity search range [2][27][4][5][30][8], and the other is to construct cascade cost volume with a narrowed disparity range [7][24][3][12][13][17]. Nowadays, cascade cost volume methods are popular since they can achieve high resolution depth map with efficient hardware usage.

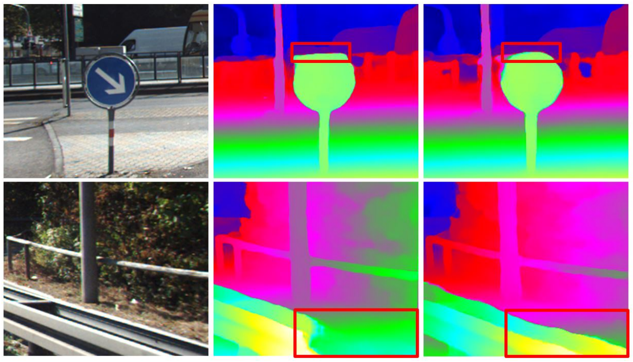
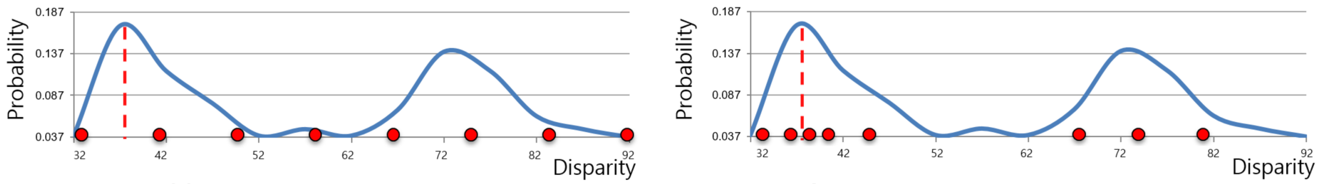


Figure 1. Comparison of disparity estimation results between a recent method [7] (middle column) and our proposed UASNet (right column). The baseline method [7] estimates wrong disparities in areas depicted by red boxes, where similar textures lead to stereo matching ambiguity. In comparison, our UASNet estimates correct disparities based on our novel design of URP to improve disparity range prediction and UDS to improve the disparity sampling (see text for details).

For cascade cost volume methods, an important component is the method to narrow the disparity range. Previous methods [7][24] narrow disparity range through simply adding a constant offset to the initial predicted disparity. This assigns the same offset for all pixels and may miss the ground truth disparity when the predicted error is large as in Figure 1. Recently, UCS-Net [3] predicts per-pixel disparity range based on variance. However, only using statistical variance and manually designed rules to predict the offset is not enough to handle the complex matching ambiguities. For example, a number of different matching distributions may produce the same variance, which is insufficient to predict disparity range. With the predicted disparity range, another key challenge is designing of disparity sampling. Constrained by cost volume construction implementation, the number of samples of per-pixel disparity range has to be the same. Previous methods [3][6] utilize a uniform sampling, which easily misses the ground truth disparity in a large disparity range. As shown in Figure 2, the uniform sampling causes the sampling points (red circles) far away



(a) Uniform sampling

(b) Uncertainty-based sampling

Figure 2. Comparison of two disparity sampling methods. The black blue curve describes matching probability distribution along the disparity dimension. Subfigure (a) shows uniform disparity sampling. The sampling points (red circles) may fall far away from the true disparity (red dashed line). Subfigure (b) shows our proposed UDS. The sampling points are dense in intervals of high matching certainty, which significantly increases opportunity to obtain the true disparity sample.

from the true disparity (red dashed line).

In order to address the above issues, this paper proposes an uncertainty adaptive sampling network (UASNet) to construct cascade volume with improved disparity range prediction as well as effective sampling. The key to our method is that we propose a novel uncertainty distribution-guided range prediction (URP) to precisely estimate per-pixel disparity range and an uncertainty-based disparity sampler (UDS) to adaptively adjust sampling interval to localize disparity with improved accuracy. Specifically, the URP explores the more discriminative uncertainty distribution to handle the complex matching ambiguities. It utilizes a deep learning module to learn per-pixel disparity range from the uncertainty distribution. A disparity range supervision is added to the network to learn discriminative features to improve disparity range prediction. The UDS discretizes per-pixel predicted candidate range according to matching uncertainty. In this way, samples are dense in high matching certainty range and it becomes easy to obtain the true disparity sample of the subsequent stage. With the proposed modules, cascade cost volume is constructed and the full resolution disparity map is obtained.

Figure 1 visualizes comparison between a recent cascade cost volume method [7] and our UASNet. It can be seen that some pixels on the top area of traffic sign and a part of road curb denoted by red boxes can not find the correct matches in [7]. In comparison, our UASNet in the right column can learn the correct disparities with improved range prediction and accurate uncertainty-based sampling.

To summarize, our network improves stereo matching accuracy by the proposed UASNet with a cascade cost volume representation. Our contributions are summarized as follows:

- We propose an URP module to explore the uncertainty distribution to handle the complex matching ambiguities and improve disparity ranges prediction. It achieves the highest ground truth covering ratio in comparison with other range prediction methods [7][3][6].
- We propose an UDS strategy to adaptively adjust per-pixel sampling interval according to the matching uncertainty which achieves dense sampling in high matching certainty range and thus is easy to obtain the true disparity sample for the subsequent stage.
- Our proposed method achieves top performance on both SceneFlow dataset [15] and KITTI benchmark [16].

2. Related Work

2.1. Deep Stereo Matching

The stereo matching problem has been intensively studied for a long time and has achieved significant progresses in recent years [27][1][25][32][22]. Most of the stereo matching processes can be summarized as four steps [2][18], e.g. feature extraction, cost construction, cost aggregation and disparity computation. Compared with traditional methods [9][11][10], deep stereo matching shows great potential in feature extraction and cost aggregation processes, which significantly boost the matching accuracy on stereo benchmarks [21][31].

Early work in deep stereo matching was started by Zbontar and LeCun [28] that proposed a deep network to match image patches, followed by traditional cost regularization. Later, GC-Net [14] incorporated all components of stereo matching into a single end-to-end learning model. Following GC-Net, PSM-Net [2] proposed a stacked 3D convolution hourglasses structure to aggregate cost volume. To further improve the matching accuracy, CSPN [4] proposed a convolutional spatial propagation network to aggregate non-local cost information and GANet [30] introduced a semi-global aggregation layer for cost aggregation. Later, GwcNet [8] proposed an enhanced cost volume presentation by introducing a group-wise correlation cost volume. It worths noticing that these methods constructed cost volume in the entire disparity search range. However, due to high computation and memory cost, they have to construct low resolution cost volume, which limits further improvement in matching accuracy.

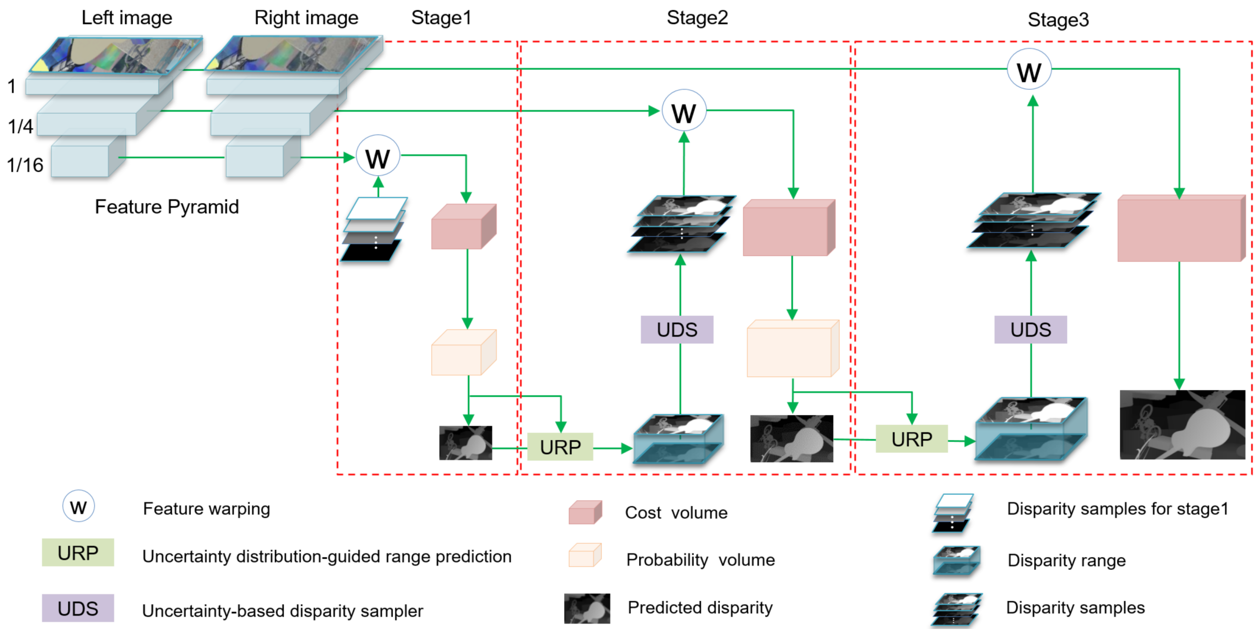


Figure 3. The pipeline of our proposed UASNet in a cascade cost volume representation. We firstly compute per-pixel disparity range by URP as well as uncertainty-based adaptive samples by UDS. Then, we apply feature warping based on the adaptive samples to build sparse cost volume. Finally, we regress disparity map from cost volume. In our experiments, we construct effective three-stage cascade cost volumes with spatial resolution changes from 1/16, 1/4 to 1 times of the original resolution.

2.2. Coarse-to-fine Stereo Matching

To deal with issues in computation and memory cost, coarse-to-fine stereo matching methods have been proposed to gradually construct high-resolution cost volumes with efficient hardware usage [24][23][12][13][7][3]. AnyNet [24] constructed a high-resolution cost volume of a narrowed disparity range by adding a small fixed offset on the initial disparity. An early work with cascade cost volume method [7] leveraged the predicted disparity in previous stage to progressively reduce the search range. Both methods assign the same disparity range for all pixels and may easily miss the right matches if prediction errors are large in the coarse stage. An inherent design issue that all cascade-based stereo methods must address is how to choose the disparity range and sampling to construct the cascade cost volume.

Recent works [6] [3] predicted per-pixel disparity ranges to address the range selection issue. DeepPruner [6] used a differentiable PatchMatch layer to prune per-pixel range. UCS-Net [3] improved accuracy of predicted range based on the variance of the probability distribution. However, it is still insufficient to estimate the disparity range using the simple one-dimensional variance, especially for complex scenes with matching ambiguities. Meanwhile, in large disparity range, uniform sampling performed in the above methods [6] [3] results in sparse sampling that easily misses the true disparity.

Different from previous works, here we explore the more informative uncertainty distribution to predict the disparity range and discretize per-pixel range by uncertainty-based disparity sampling. Notably, our proposed modules can be plugged into any exiting cascade cost volume networks to provide accurate disparity range prediction and sampling strategy. Here, we select cascade cost volume method [7] and iterative stereo depth estimation method [6] as the backbone.

3. Method

Figure 3 illustrates the pipeline of our network. It consists of three-stage cascade cost volumes, for which the spatial resolution increases from 1/16, 1/4 to 1 times of the original resolution. The network first extracts multi-scale features through a feature pyramid module. In the first stage, our method builds a low-resolution cost volume by warping right feature map across a full disparity search range and concatenating it with left feature map. Then, a probability volume is learned from the cost volume to predict the corresponding disparity map. In the second stage, an upsampled disparity map and the probability volume are passed through URP and UDS for generating fine adaptive disparity samples. Based on the samples, a high-resolution cost volume is established to predict the disparity map. In the final stage, a full resolution cost volume is constructed based on the estimated finer samples and regressed to predict the full resolution disparity map. The following sections will introduce our proposed URP, UDS and loss function.

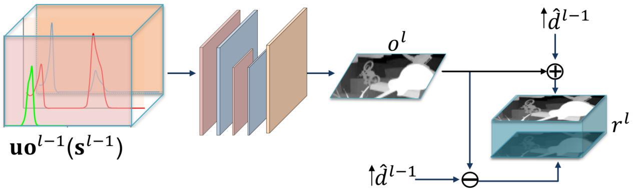


Figure 4. Uncertainty distribution-guided range prediction process. This module inputs uncertainty-based offset vector $\mathbf{uo}^{l-1}\left(\mathbf{s}^{l-1}\right)$ from l–1 stage, and learns per-pixel offset o^l through an encoder-decoder network. The output disparity range r^l is calculated by the offset o^l and the upsampled disparity map $\uparrow d^{l-1}$ of $l-1$ stage.

3.1. Uncertainty Distribution-guided Range Prediction

In stereo matching problem, disparity is generally regressed from probability volume **p** and disparity sample set **s**. It can be represented as a probability weighted average of disparity sample set with (Eq.(1)) [14].

$$\hat{d} = \sum_{d=s_1}^{s_n} d \times p\left(d\right) \tag{1}$$

here, $\hat{d}$ denotes the regressed disparity, $\mathbf{s} = \{s_1, ..., s_n\}$ is the set of disparity samples, and $p\left(d\right)$ is the probability that disparity equal to d. In practice, the probability distribution of disparity samples varies with each pixel. For a pixel in textured regions, the probability distribution is a unimodal distribution and the true disparity generally corresponds to the peak. For a pixel in texture-less or repeated texture areas, the probability distribution tends to have several local peaks due to ambiguous feature matching. For a pixel in occluded region, the probability distribution is flat because there is no correct matching.

To describe the uncertainty of probability distribution, UCS-Net [3] leverages the variance for uncertainty estimation. The UCS-Net considers the uncertainty of each pixel and uses statistical variance to predict per-pixel offset. Actually, different distributions can correspond to the same variance. In the supplementary material (Figure S2), we show six different distributions generating the same variance, in which four cases cannot estimate the disparity range correctly, and the other two cases predict a larger disparity range than the offset between the true and predicted disparities.

To deal with this issue, our proposed URP explores the more informative uncertainty distribution to predict per-pixel disparity ranges. Instead of using one-dimension variance information as in [3], we leverage multi-dimension uncertainty distribution of each pixel. The offset between disparity sample s_i and the regressed disparity $\hat{d}$ also affects the final offset estimation. Therefore, we calculate the uncertainty-based offset of each-pixel sample by multiplying the uncertainty with the offset, as shown in Eq.(2). Here, the uncertainty refers to the matching probability of the disparity sample.

$$uo^{l-1}\left(s_i^{l-1}\right) = p^{l-1}\left(s_i^{l-1}\right) * \left(s_i^{l-1} - \hat{d}^{l-1}\right)^2 \tag{2}$$

here, uo^{l-1} and p^{l-1} denote the uncertainty-based offset and the probability of sample s_i at stage $l-1$. $\hat{d}^{l-1}$ is the regressed disparity at stage $l-1$.

For each pixel, we obtain a multi-dimension uncertainty-based offset vector $\mathbf{uo}^{l-1}\left(\mathbf{s}^{l-1}\right) = \left\{uo^{l-1}\left(s_1^{l-1}\right), ..., uo^{l-1}\left(s_n^{l-1}\right)\right\}$ corresponding to sample set **s**. Then, instead of manually designed rules as in [3], our URP model uses a deep learning module to learn per-pixel disparity offset from $\mathbf{uo}^{l-1}\left(\mathbf{s}^{l-1}\right)$. A disparity range loss supervises the network to learn discriminant features to solve the complex matching ambiguities and predict a compact range to cover the true disparity.

Figure 4 visualizes the URP process. Given the per-pixel uncertainty-based offset vector $\mathbf{uo}^{l-1}\left(\mathbf{s}^{l-1}\right)$, an encoder-decoder structure is applied to learn informational feature and predict the per-pixel offset o^l at stage l. Finally, we compute the per-pixel lower bound by subtracting per-pixel offset from regressed disparity map $\hat{d}^{l-1}$ and upper bound by summing the offset and $\hat{d}^{l-1}$.

3.2. Uncertainty-Based Disparity Sampler

To successfully construct regular cost volume, the number of samples of each pixel has to be the same. However the uniform sampling easily misses the true disparity in a large search range. To handle this problem, our proposed UDS discretizes per-pixel predicted candidate range based on matching uncertainty, which enables the sampling point distribution consistent with the matching probability distribution. In this way, dense samples are located in high matching certainty range and it is easy to obtain the true disparity sample at later stage.

Figure 5 visualizes our proposed UDS process. Inputing an upsampled probability volume $\mathbf{p}^{l-1}$ and a predefined sampling number N, the estimated range r^l is discretized into $N-1$ parts. Firstly we normalize the probability within range r^l along the disparity dimension. Then, starting from the lower bound, if the area of accumulative probability histogram reaches $i/(N-1)$ accordingly, where $i = 0, 1, ..., N-1$, then each end point is considered as sample s_i. Finally, we obtain N samples $s_0, s_1, ..., s_{N-1}$. In this way, samples are dense in high probability area.

Specifically, the disparity sample s_i is computed as follows:

$$\begin{aligned} &if\ \ p\left(d \le d_{k-1}\right) < i/(N-1) \le p\left(d \le d_k\right): \\ &s_i = d_{k-1} + \frac{\frac{i}{N-1} - p\left(d \le d_{k-1}\right)}{p\left(d_k\right)},\ \ k = 1, 2, ..., K \end{aligned} \tag{3}$$

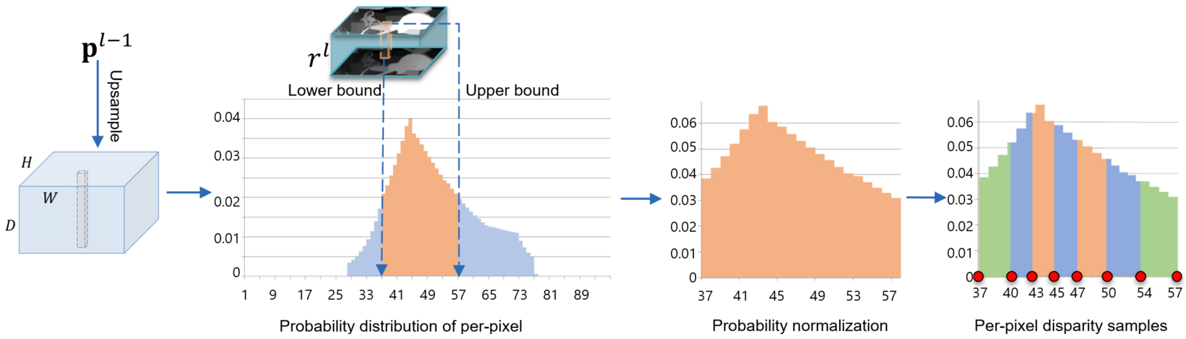


Figure 5. Example of uncertainty-based disparity sampler. We truncate each pixel probability distribution with the predicted range r^l and normalize it. Then, disparity samples are obtained by equalizing the accumulative probability.

here, $d_k = d_{min} + k * \frac{d_{max} - d_{min}}{K}$ refers to disparity value within range r^l, K is a fixed constant, and d_{min} and d_{max} are the lower bound and upper bound of range r^l, respectively. The lower bound d_{min} is sample s_0. $p\left(d \leq d_k\right) = \sum_{d=d_{min}}^{d_k} p^{l-1}\left(d\right)$ is the accumulative probability.

3.3. Loss Function

Disparity Loss. The end point error (EPE) with smooth L_1 is used for computing disparity loss.

$$EPE\left(d_{gt}, \hat{d}\right) = L_{smooth}\left(d_{gt} - \hat{d}\right) \tag{4}$$

here, d_{gt} refers to ground truth disparity and $\hat{d}$ is the predicted disparity. We supervise the disparity outputs of all stages and the total predicted disparity loss for our network as follows:

$$L_{disp} = \sum_{l=1}^{3} w^l \sum_{i=1}^{4} \lambda_i \cdot EPE\left(d_{gt}^l, \hat{d}_i^l\right) \tag{5}$$

where $EPE\left(d_{gt}^l, \hat{d}_i^l\right)$ refers to the loss for the ith disparity prediction at stage l, and w^l and λ_i refer to their corresponding loss weights. Each stage has four outputs from a pre-hourglass module and three stacked 3D hourglass modules. In our training process, each stage outputs the disparity maps and the loss is back propagated. For testing process, only the final stage outputs the disparity map.

Disparity Range Loss. In order to learn a compact search space for each pixel, we use a disparity range loss [6]. This loss includes two parts: a relaxation loss and a absolute loss. The relaxation loss aims to ensure the disparity range large enough to cover the ground truth disparity and the absolute loss constrains the disparity range as small as possible.

Eq.(6) introduces the relaxation loss. If the lower bound is larger than the ground truth disparity, a large penalty is implemented to encourage the lower bound to be smaller than the ground truth. In contrast, for the predicted upper bound, the relaxation loss encourages it to be bigger than the ground truth.

$$\begin{aligned} L_{relax\_min} &= \begin{cases} \gamma \cdot L_1\left(d_{gt}, d_{min}\right), & if\ d_{min} \leq d_{gt} \\ (1-\gamma) \cdot L_1\left(d_{gt}, d_{min}\right), & otherwise \end{cases} \\ L_{relax\_max} &= \begin{cases} \gamma \cdot L_1\left(d_{gt}, d_{max}\right), & if\ d_{gt} \leq d_{max} \\ (1-\gamma) \cdot L_1\left(d_{gt}, d_{max}\right), & otherwise \end{cases} \end{aligned} \tag{6}$$

where γ is a weight and smaller than 0.1 in the experiments. L_1 refers to L_1 loss. L_{relax_min} and L_{relax_max} constrain the lower and upper bounds, respectively, which allows the disparity range to be large enough to cover the ground truth. However, a too large range reduces the possibility of sampling at the ground truth disparity. To deal with this, the EPE absolute loss is used to enforce that the range does not become too large. Therefore our design combines both the L_{relax_min}, L_{relax_max}, EPE(d_{gt}, d_{min}) and EPE(d_{gt}, d_{max}) together to guarantee a reasonable predicted range.

$$\begin{aligned} L_{range\_loss} = \sum_{l=2}^{3} & \left(\alpha_l \cdot \left(L_{relax\_min}^l + L_{relax\_min}^l\right)\right. \\ & \left. + \beta_l \cdot \left(EPE\left(d_{gt}^l, d_{min}^l\right) + EPE\left(d_{gt}^l, d_{max}^l\right)\right)\right) \end{aligned} \tag{7}$$

where the α_l and β_l are two balancing weights at stage l, and a bigger β means a smaller coverage range. Here, we only need to predict the disparity range of stage2 and stage3. **Total Loss Function.** Total loss function is defined as: $L = L_{disp} + L_{range\_loss}$.

4. Experiments

In this section, we describe the details of experiments including the datasets, evaluation metrics, training settings and make ablation study to verify the proposed components of the network. Then, we compare our results with the SOTA methods on public datasets.

4.1. Datasets and Implementation Details

SceneFlow[15] is a large synthetic dataset that contains 35,454 pairs of training image and 4,370 pairs of test image. Finalpass version is used to train our model because it is

Method	Cascade	Cost1	Cost2	Cost3	Range	Sampling		SceneFlow		KITTI2015	
		1/16	1/4	1		Uniform	UDS	EPE ↓ $[px]$	CR ↑ $[\%]$	All ↓ $[\%]$	CR ↑ $[\%]$
Cas$_2$[7]	2	✓	✓		FIX	✓		0.649	99.04	2.0	99.35
Cas$_2$+variances[3]	2	✓	✓		VAN	✓		0.645	97.72	1.91	98.62
Cas$_2$+Ours(URP)	2	✓	✓		URP	✓		0.623	**99.39**	1.87	99.70
Cas$_2$+Ours(URP+UDS)	2	✓	✓		URP		✓	**0.619**	99.35	**1.86**	**99.73**
Cas$_3$[7]	3	✓	✓	✓	FIX	✓		0.581	97.36	1.75	98.66
Cas$_3$+Ours(URP)	3	✓	✓	✓	URP	✓		0.554	**98.71**	1.69	99.28
Cas$_3$+Ours(URP+UDS)	3	✓	✓	✓	URP		✓	**0.527**	98.70	**1.66**	**99.34**

Table 1. The ablation study on SceneFLow and KITTI2015 benchmark. Here, FIX, VAN and URP refer to range prediction methods with a fixed offset, variance-based and URP-based modules, respectively. CR refers to covering ratio.

Method	Range	Sampling		SceneFlow	
		Unif	UDS	EPE ↓$[px]$	CR ↑$[\%]$
DeepPruner*[6]	PRU	✓		0.996	98.34
DeepPruner*+Ours(URP)	URP	✓		0.934	**98.99**
DeepPruner*+Ours(URP+UDS)	URP		✓	**0. 901**	98.97

Table 2. The ablation study on SceneFLow dataset. Here, PRU and URP refer to prunered range and URP-based predicted range. Unif refers to uniform sampling. CR refers to covering ratio.

close to the real scene which has motion blur and defocus. The network is trained for 64 epochs with Adam optimizer. The initial learning rate is 0.001 and is down-scaled by 2 after epoch 10, 12, 14 and ends at $1.25e^{-4}$. In training process, we randomly crop input image to 512 × 256 patch. The coefficients w^l and λ_i in disparity loss are the same as [7]. For disparity range loss, the coefficients α_2, β_2 are set as 4.0, 0.7 and α_3, β_3 are set as 4.0, 2.8.

KITTI [16]. KITTI 2012 dataset consists of 194 training image pairs and 195 test image pairs. KITTI 2015 dataset contains 200 training images and 200 test images. In our implementation, we combine KITTI 2012 with 2015 dataset together and there are totally 394 stereo image pairs. We randomly select 347 images for training and the rest are used for validation. The pretrained model on SceneFlow dataset is finetuned on KITTI for a further 600 epochs. The learning rate is 0.001 for the first 200 epochs and $1e^{-4}$ for the rest epochs. Since the ground truth of KITTI is captured by laser scanner directly, it is much more sparser than SceneFlow dataset, so we augment data like HSM-Net [26]. Specifically, asymmetric chromatic augmentation is used to improve the robustness of the network to handle different lighting and exposure conditions. We also apply asymmetric occlusion augmentation by replacing the randomly selected rectangular area on the left image with the average value of the entire image, which helps disparity estimation in the occluded area.

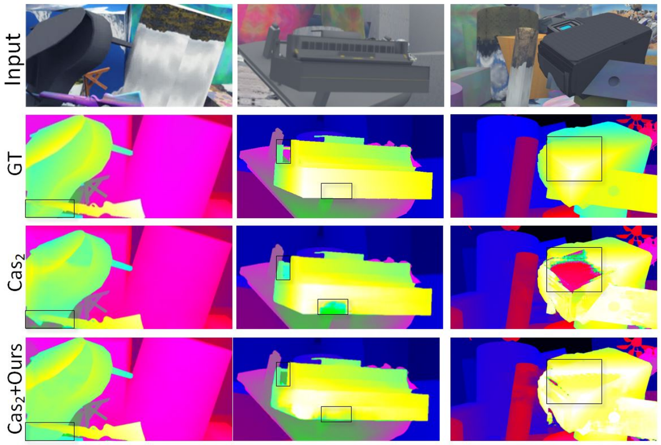


Figure 6. Qualitative comparison results on SceneFlow dataset. Our proposed methods work better on textureless regions, such as the area highlighted with black bounding box, while baseline method predicts wrong disparity due to fixed offset and uniform sampling method.

4.2. Ablation Study

In this section, we make ablation study to validate the improvement of our proposed URP and UDS. On SceneFlow dataset and KITTI 2015 benchmark, we improve Cas$_2$ and Cas$_3$ methods [7] with our proposed modules and denote them as Cas$_2$+Ours(URP), Cas$_2$+Ours(URP+UDS), Cas$_3$+Ours(URP) and Cas$_3$+Ours(URP+UDS). Cas$_2$ refers to 2-stage cascade cost volume model, the size of cost volume is $H/4 \times W/4 \times C \times 12$ and $H/2 \times W/2 \times C/2 \times 12$ respectively. For Cas$_3$, additional $H \times W \times C/4 \times 8$ cost volume is constructed to directly output full resolution disparity map.

Besides, we compare our model with current SOTA range prediction methods [3] [6]. Specifically, for UCS-Net[3], we replace range prediction module in Cas2 with variances-based method in UCS-Net and names it as Cas2+variances. For DeepPruner* [6], we replace the range predictor model of DeepPruner* with our

	SceneFlow	**KITTI2012**				**KITTI2015**			
Method	**EPE** $\downarrow$	**Bad2.0** $\downarrow$	**Bad3.0** $\downarrow$	**Ref2.0** $\downarrow$	**Ref3.0** $\downarrow$	**D1-bg** $\downarrow$	**D1-fg** $\downarrow$	**ALL** $\downarrow$	**Noc** $\downarrow$
	[px]	noc[%]	noc[%]	noc[%]	noc[%]	all[%]	all[%]	all[%]	noc[%]
PSMNet [CVPR18'][2]	1.09	2.44	1.49	13.77	8.36	1.86	4.62	2.32	2.14
DeepPruner* [ICCV19'][6]	0.86	-	-	-	-	1.87	3.56	2.15	1.95
GwcNet [CVPR19'][8]	0.77	2.16	1.32	12.49	7.80	1.74	3.93	2.11	1.92
Cas [CVPR20'][7]	0.62	-	-	-	-	1.59	4.03	2.00	1.78
GA-Net [CVPR19'][30]	0.84	1.89	1.19	10.75	6.22	1.48	3.46	1.81	1.63
AcfNet [AAAI20'][32]	0.87	1.83	1.17	11.17	6.93	1.51	3.80	1.89	1.72
CSPN [TPAMI19'][4]	0.78	**1.79**	1.19	-	6.92	1.51	2.88	1.74	1.61
LEAStereo [NIPS20'][5]	0.78	1.90	**1.13**	9.66	5.35	**1.40**	2.91	**1.65**	**1.51**
UASNet(Ours)	**0.53**	1.81	1.18	**8.02**	**4.55**	1.44	**2.79**	1.66	**1.51**
Rank	**1**	2	3	**1**	**1**	2	**1**	2	**1**

Table 3. Quantitatively comparison with the SOTA methods. UASNet(Ours) refers to Cas$_3$+Ours(URP+UDS).

URP and name it as DeepPruner*+Ours(URP). Here, DeepPruner* refers to DeepPruner-Best model. For a fair comparison, the model size of DeepPruner*+Ours(URP) is slightly smaller than the DeepPruner*. Furthermore, we add UDS to DeepPruner*+Ours(URP) and name it as DeepPruner*+Ours(URP+UDS).

Evaluation of Predicted Disparity Precision. As shown in Tabel 1, Cas$_2$+Ours(URP) decreases EPE error by 4% and ALL error (bad pixel ratio with 3 pixel threshold for all labeled pixels) by 6.5% in comparison with Cas$_2$. Cas$_3$+Ours(URP+UDS) decreases EPE error by 9.3% and ALL error by 5.1% when comparing with Cas$_3$. Furthermore, Cas$_2$+Ours(URP) has lower EPE and ALL errors than Cas$_2$+variances[3].

As shown in Table 2, DeepPruner*+Ours(URP) decreases EPE error by 6.2% and improves covering ratio from 98.34% to 98.99% when comparing with DeepPruner*. DeepPruner*+Ours(URP+UDS) decreases EPE error by 9.5% than DeepPruner*. Here, the EPE result of DeepPruner* is slightly higher than the result stated in [6], because we use the finalpass version of SceneFlow dataset which is harder than the cleanpass version.

Figure 6 visualizes comparison results of Cas$_2$[7] and Cas$_2$+Ours(URP+UDS) on SceneFlow dataset. Our method performs a better disparity estimation in some challenging areas, such as the area highlighted with black bounding box, where baseline method predicts wrong disparity due to fixed offset and uniform sampling. In the supplementary material (Figure S3), we visualize some comparison results of uniform sampling and our UDS method on KITTI 2015 benchmark. It demonstrates that UDS can adaptively adjust sampling interval to correctly estimate the disparity, especially for thin and long structures.

Evaluation of Predicted Disparity Range. We quantify the pixel proportion that the predicted disparity range covers ground truth disparity in an image, named covering ratio (CR). The result shows that the CR of Cas$_2$+Ours(URP) and Cas$_3$+Ours(URP) achieves 99.39%, 98.71% on SceneFlow and 99.70%, 99.28% on KITTI respectively, which validates our predicted range can cover almost all ground truth disparity. Here, the CR of the stage3 is lower than stage2 because we use a bigger β in disparity range loss to ensure a compact range. When evaluated on sceneflow dataset with 960*540 image resolution, our URP method can additionally cover the ground truth disparity in 1814 pixels and 6998 pixels for stage2 and stage3 respectively than the Cas method. Furthermore, our URP method achieves the highest ground truth covering ratio in comparison with the fixed offset method [7], variance-based range prediction method [3] and range pruning method[6].

Figure 7 visualizes our predicted disparity range of stage 2 and stage 3. It can be seen that the predicted ranges are small in most cases, which greatly reduce the computation and memory burdens for building cost volume. Meanwhile, almost all the ground truth disparities lie between the upper bound curve and lower bound curve.

4.3. Comparison with the SOTA Methods

Table 3 quantitatively compares our method with the SOTA methods containing PSM-Net [2], DeepPruner* [6], GwcNet [8], Cas[7], GA-Net [30], AcfNet[32], CSPN[4] and LEAStereo[5]. Among them, LEAStereo is a Neural Architecture Search (NAS) method. Our method achieves the best performance on SceneFlow testing dataset and achieves top performance on the KITTI Stereo 2012 and 2015 benchmarks.

Figure 8 visualizes the comparison of our method and the SOTA methods. Three images posted on KITTI 2015 leaderboard are presented and compared. Notice that our method has better performance on foreground object boundaries, such as road signs, which are highlighted by yellow squares in the images and the error map is zoomed in for better visualization.

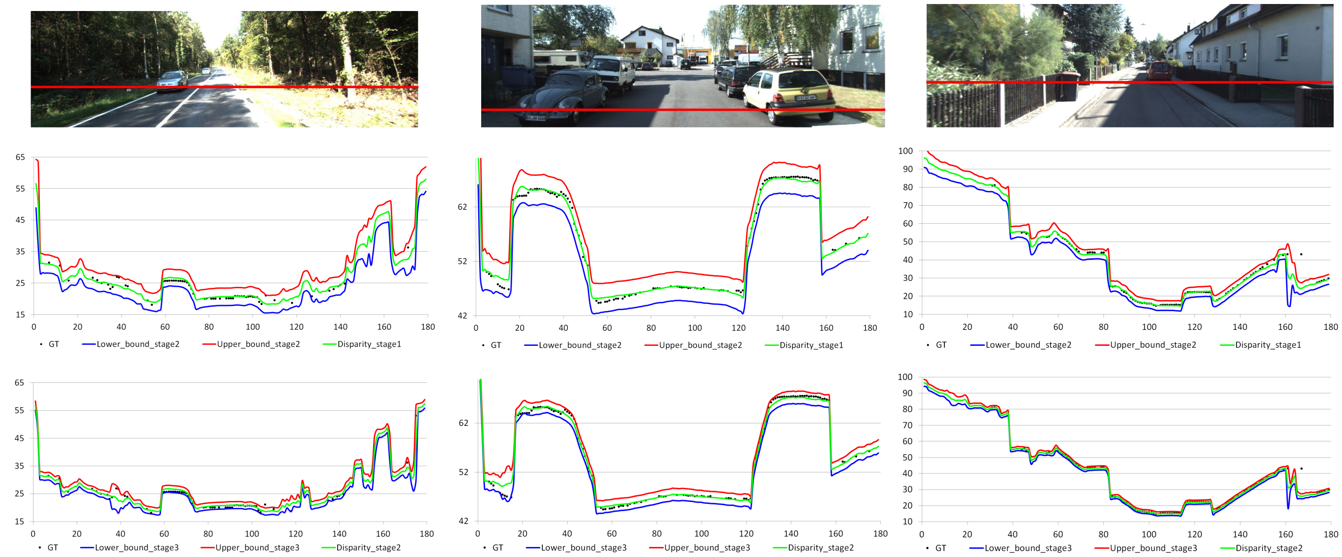


Figure 7. Predicted disparity range visualization of stage 2 and stage 3. The predicted disparity range (red line: upper bound; blue line: lower bound) is large enough to cover the ground truth disparity (black points), and small enough to reduce the computation and memory burden of cost volume building.

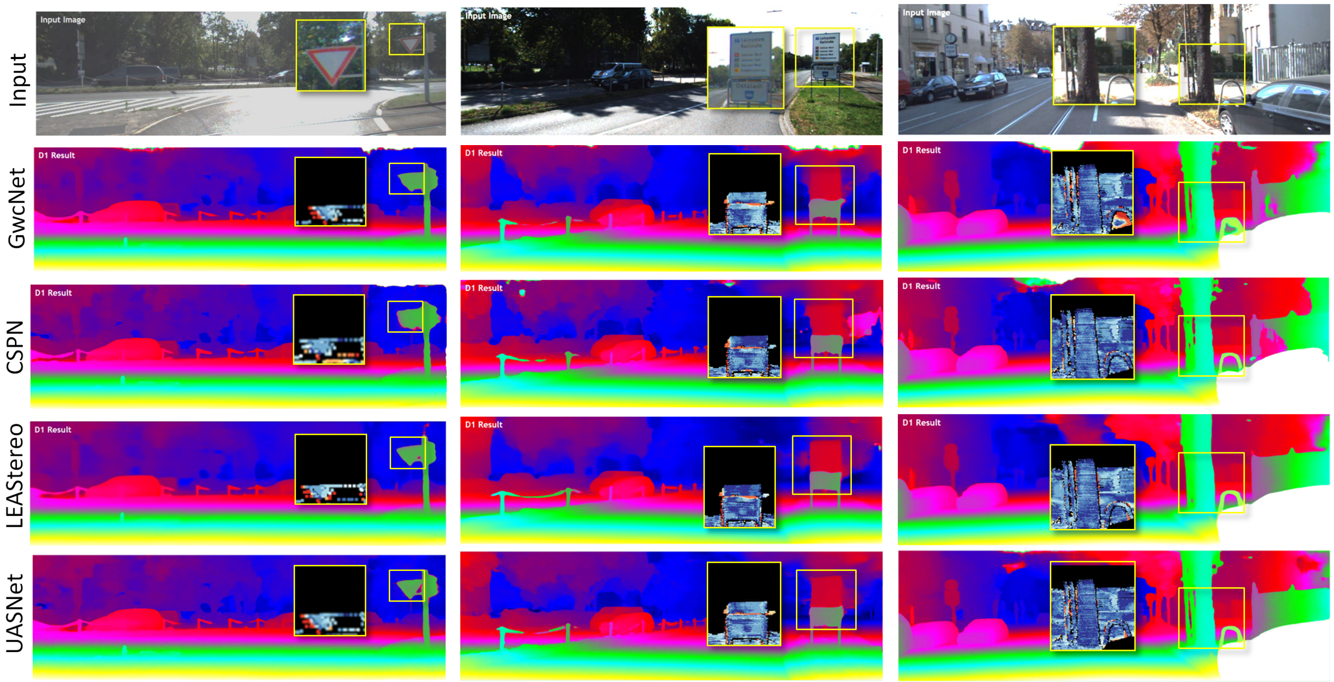


Figure 8. Visualized comparison of our results with the SOTA methods. The results are the colorized disparity maps generated from KITTI.

5. Conclusion

In this paper, we propose an uncertainty adaptive sampling network (UASNet) to construct cascade cost volume with improved disparity range prediction as well as effective sampling. Through experiments we validate the effectiveness of each design component, which work together to improve our UASNet performance over recent SOTA methods such as the cascade cost volume [7] and DeepPruner* [6]. Further experiments show that our method achieves top performance on three stereo matching benchmarks: SceneFlow[15] , KITTI2012 and KITTI2015 [16].

6. Acknowledgments

Yuchao Dai was supported in part by National Key Research and Development Program of China (2018AAA0102803) and National Natural Science Foundation of China (61871325).

References

[1] Abhishek Badki, Alejandro Troccoli, Kihwan Kim, Jan Kautz, Pradeep Sen, and Orazio Gallo. Bi3d: Stereo depth estimation via binary classifications. In *CVPR*, pages 1600–1608, 2020.

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

[3] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *CVPR*, pages 2524–2534, 2020.

[4] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *TPAMI*, 2019.

[5] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Tom Drummond, Hongdong Li, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. In *NeurIPS*, 2020.

[6] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *ICCV*, pages 4384–4393, 2019.

[7] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, pages 2495–2504, 2020.

[8] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *CVPR*, pages 3273–3282, 2019.

[9] Heiko Hirschmller. Stereo processing by semiglobal matching and mutual information. *TPAMI*, 30, 2007.

[10] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *CVPR*, volume 2, pages 807–814, 2005.

[11] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *TPAMI*, 35(2):504–511, 2013.

[12] Yinlin Hu, Rui Song, and Yunsong Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *CVPR*, pages 5704–5712, 2016.

[13] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *CVPR*, pages 8981–8989, 2018.

[14] 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 *ICCV*, pages 66–75, 2017.

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

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

[17] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, pages 4161–4170, 2017.

[18] Zhibo Rao, Mingyi He, Yuchao Dai, Zhidong Zhu, and Renjie He. Nlca-net: a non-local context attention network for stereo matching. *APSIPA Trans. Signal Inf. Process.*, 9, 2020.

[19] Korbinian Schmid, Teodor Tomic, Felix Ruess, Heiko Hirschmüller, and Michael Suppa. Stereo vision based indoor/outdoor navigation for flying robots. In *IROS*, pages 3955–3962, 2013.

[20] Sayanan Sivaraman and Mohan M Trivedi. A review of recent developments in vision-based vehicle detection. In *IEEE IV*, pages 310–315, 2013.

[21] Xiao Song, Xu Zhao, Liangji Fang, Hanwen Hu, and Yizhou Yu. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. *Int J Comput Vis*, 128(4):910–930, 2020.

[22] Xiao Song, Xu Zhao, Hanwen Hu, and Liangji Fang. Edgestereo: A context integrated residual pyramid network for stereo matching. In *ACCV*, pages 20–35, 2018.

[23] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *CVPR*, pages 195–204, 2019.

[24] Yan Wang, Zihang Lai, Gao Huang, Brian H Wang, Laurens Van Der Maaten, Mark Campbell, and Kilian Q Weinberger. Anytime stereo image depth estimation on mobile devices. In *ICRA*, pages 5893–5900, 2019.

[25] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *CVPR*, pages 1959–1968, 2020.

[26] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *CVPR*, pages 5515–5524, 2019.

[27] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In *CVPR*, pages 6044–6053, 2019.

[28] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *CVPR*, pages 1592–1599, 2015.

[29] Nadia Zenati and Noureddine Zerhouni. Dense stereo matching with application to augmented reality. In *ICSPC*, pages 1503–1506, 2007.

[30] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip H. S. Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *CVPR*, pages 185–194, 2020.

[31] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *ECCV*, pages 420–439, 2020.

[32] Youmin Zhang, Yimin Chen, Xiao Bai, Suihanjin Yu, and Kuiyuan Yang. Adaptive unimodal cost volume filtering for deep stereo matching. In *AAAI*, volume 34, pages 12926–12934, 2020.