

# Digging Into Normal Incorporated Stereo Matching

Zihua Liu\*

Tokyo Institute of Technology  
Tokyo, Japan  
zliu@ok.sc.e.titech.ac.jp

Zhicheng Wang

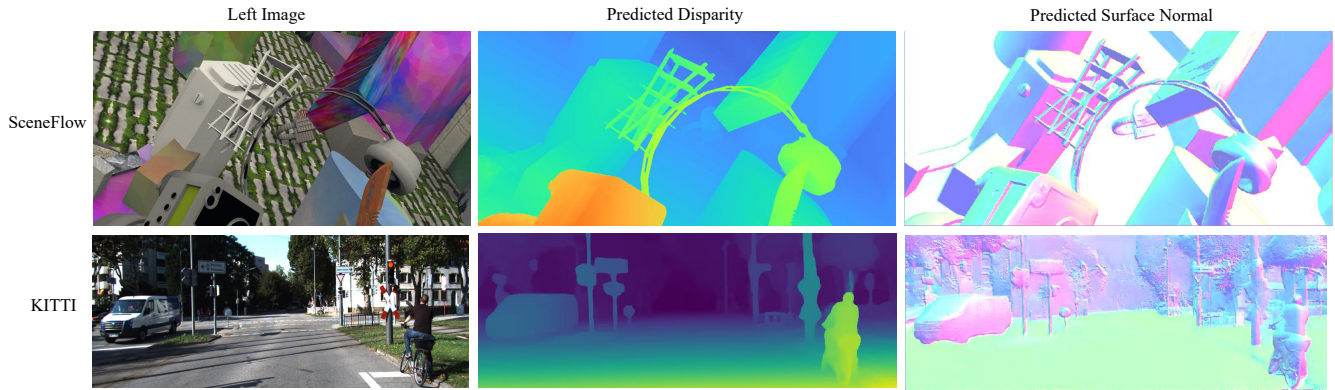
CAD Research Center, Tongji University.  
Shanghai, China  
zhichengwang@tongji.edu.cn

Songyan Zhang\*

CAD Research Center, Tongji University.  
Shanghai, China  
spyder@tongji.edu.cn

Masatoshi Okutomi<sup>†</sup>

Tokyo Institute of Technology  
Tokyo, Japan  
mxo@ctrl.titech.ac.jp



**Figure 1: Visualization of predicted disparity and surface normal on SceneFlow and KITTI 2015 datasets. Our proposed NINet successfully preserves detailed structure and consistency even in the challenging bordered, low-texture and occluded regions.**

## ABSTRACT

Despite the remarkable progress facilitated by learning-based stereo matching algorithms, disparity estimation in low-texture, occluded, and bordered regions still remain bottlenecks that limit the performance. To tackle these challenges, geometric guidance like plane information is necessary as it provides intuitive guidance about disparity consistency and affinity similarity. In this paper, we propose a normal incorporated joint learning framework consisting of two specific modules named non-local disparity propagation (NDP) and affinity-aware residual learning (ARL). The estimated normal map is first utilized for calculating a non-local affinity matrix as well as a non-local offset to perform spatial propagation at the disparity level. To enhance geometric consistency, especially in low-texture regions, the estimated normal map is then leveraged to

calculate a local affinity matrix which provides the residual learning with information about where the correction should refer and thus improve the residual learning efficiency. Extensive experiments on several public datasets including Scene Flow, KITTI 2015, and Middlebury 2014 validate the effectiveness of our proposed method. By the time we finished this work, our approach ranked 1st for stereo matching across foreground pixels on the KITTI 2015 dataset and 3rd on the Scene Flow dataset among all the published works.

## CCS CONCEPTS

• **Computing methodologies** → **Matching**;

## KEYWORDS

Surface Normal, Stereo Matching, Residual Learning, Spatial Propagation

## ACM Reference Format:

Zihua Liu, Songyan Zhang, Zhicheng Wang, and Masatoshi Okutomi. 2022. Digging Into Normal Incorporated Stereo Matching. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3503161.3548312>

## 1 INTRODUCTION

Stereo matching has been a fundamental task in computer vision for decades of years. Given a pair of rectified stereo images, the

\*Both authors contributed equally to this work.

<sup>†</sup>Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '22, October 10–14, 2022, Lisboa, Portugal.

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9203-7/22/10...\$15.00

<https://doi.org/10.1145/3503161.3548312>

core of stereo matching is to find the corresponding pixels along the epipolar line so that the disparity can be calculated, which has been closely related to various applications like robotic navigation, autonomous driving, and VR/AR.

Recently, the application of convolutional neural networks significantly facilitates the development of stereo matching [2, 7, 12, 28, 34]. Most of the learning-based methods can be divided into two categories according to the construction of cost volume which is an essential step for stereo matching as mentioned in [20]. The pioneer works [4, 12] employ the inner production for measuring the similarity between pixels and propose a lightweight correlation layer for building the cost volume which is widely adopted in the following works [10, 24, 28]. The correlation layer takes advantage of efficient 2D convolution networks but suffers from the loss of context information is inevitable as only a single channel of feature is preserved. GCNet proposes to build a 4D cost volume where the complete stereo features are directly concatenated along the disparity dimension for abundant context information. Most of the top-performance methods like [6, 27, 33] leverage this concatenation-based cost volume which follows 3D convolution networks to improve stereo matching. Even though that learning-based methods help the disparity estimation reach a new peak, ambiguous matching still exists especially in bordered, occluded, and low-texture regions, which remains a challenge. EDNet[34] proposes a novel method to learn an error-aware attention map which is utilized to guide the residual learning concentrate on the inaccurate regions. However, there still exists a limitation that the model only has intuitive information about where the error occurs but lacks guidance about where the correction should be learned from. Some previous works manage to improve the disparity estimation with stronger geometric constraints [9, 17, 31] by conducting a joint learning framework with the surface normal, but most of them just simply enforce the consistency between normal estimation and disparity estimation. The normal regression is usually independent of the stereo matching which means the plane information fails to be fully utilized by stereo matching.

To tackle the challenge of stereo matching in bordered, occluded, and low-texture regions as well as explicitly leverage normal information, in this paper, we propose a normal incorporated joint learning network named NINet. Our idea lies that the relatively easier task of normal estimation can provide abundant plane information as well as pattern affinity. And disparities within the same plane should be consistent. To realize this, we propose two specific modules which are the non-local disparity propagation(NDP) module and affinity-aware residual learning(ARL) module, respectively. The former NDP module leverages the estimated normal and computes a non-local affinity matrix and a corresponding location offset indicating the sampled points for spatial propagation at the disparity level. This operation can efficiently improve the disparity estimation performance, especially in bordered and occluded regions. Moreover, to mitigate the effect of low-texture regions, we provide another ARL module to further correct the disparity estimation. Following EDNet [34], we introduce error based attention method with the modification of utilizing the normal estimation to calculate a local affinity matrix. Using the reweighted feature map based on the local affinity matrix to regress the disparity residual. To sum up, our contributions can be summarized as follows:

- We propose a normal incorporated joint learning framework to explicitly leverage normal estimation and exploit plane information to provide intuitive geometric guidance for disparity estimation.
- We propose two specific modules: non-local spatial propagation (NDP) at the disparity level and affinity-aware residual learning (ARL) at the feature level. The former module efficiently improves the disparity estimation in both bordered and occluded regions while the latter module primarily improves the performance in low-texture regions.
- We propose a novel method to generate dense pseudo normal ground truth on sparsely annotated KITTI 2015 datasets which enables the application of our method in real-world scenes.
- We conduct extensive experiments on several public datasets including Scene Flow, KITTI 2015, and Middlebury 2014. Our proposed NINet outperforms most of the previous top-performance approaches and ranks 1st on KITTI 2015 leaderboard for foreground pixels. Our method demonstrates remarkable generalizability as well, which further validates the effectiveness of our method.

## 2 RELATED WORKS

In this section, we briefly review some pioneer learning-based works for disparity estimation as well as giving an introduction about surface normal incorporated stereo matching methods. Finally, we summarize the spatial propagation mechanism utilized in this field.

### 2.1 Learning-based Stereo Matching

The pioneer learning-based work DispNetC[12] follows the previous work [4] which introduces a correlation layer to calculate the inner product of the corresponding pixels for measuring the similarity. Many real-time approaches like [24, 28] takes this method for constructing the cost volume. To preserve abundant context information which may be lost in the correlation operation, GCNet [7] employs the concatenation of the left and right features along the disparity dimension to construct a 4D cost volume. The concatenation-based cost volume following 3D convolution networks for aggregation are widely adopted in many state-of-the-art works like [5, 21, 25, 33]. PSMNet[2] proposes a spatial pyramid pooling module for incorporating global context information into image features along with stacked hourglass 3D convolution networks for extending the context information. As 3D convolution networks are highly memory-consuming, StereoNet [8] decreases the layer number of 3D convolution networks and improves the performance by using RGB images for edge refinement. To incorporate both the advantages of correlation-based cost volume and concatenation-based volume, GwcNet[6] proposes the group-wise correlation volume to form a 4D correlation-based cost volume which is then concatenated with the concatenation cost volume. The recent remarkable work ACVNet[27] introduces an attention-based strategy to suppress redundant information in the cost volume which achieves top performance and real-time efficiency.

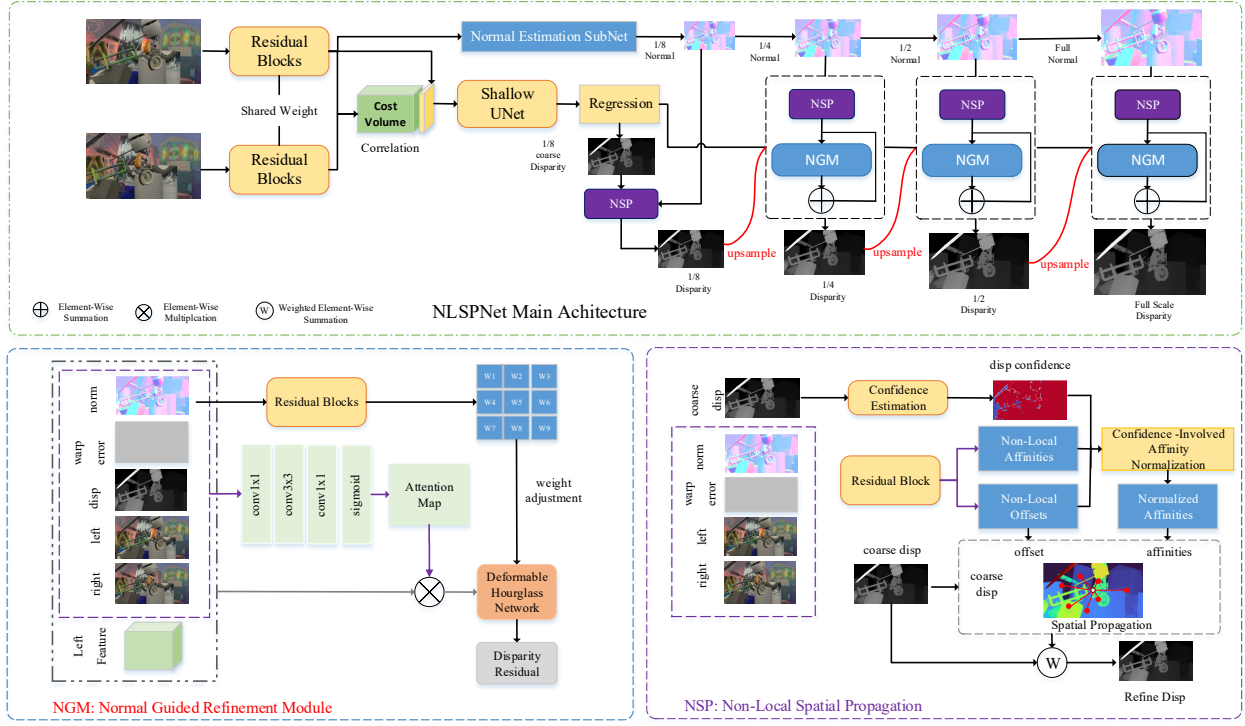


Figure 2: An overview of our proposed NINet. Our model is mainly composed of two modules named ARL and NDP. Note that some skip-connection operations are omitted here for simplifying the visualization.

## 2.2 Normal Assisted Depth Estimation

Many indoor depth estimation utilizes the normal as an assisted geometry constraints [9, 17, 18, 26, 35]. GeoNet [17] proposes a normal-to-depth branch and a depth-to-normal branch for jointly learning. VNL [31] designs a loss term named virtual normal directions to impose geometric constraints. The work [9] is also a joint learning-based model which proposes a consistency loss to refine the depth from depth/normal pairs. An energy optimization-based approach is introduced in [36] for multi-view stereo matching. IDNSolver [36] proposes an iterative energy optimization strategy which is a novel view for multi-view stereo matching. The ASNDepth [11] introduces an adaptive surface normal constraint to effectively correlate the depth estimation with geometric consistency. Most related works follow the pipeline of enforcing the consistency between the depth and the normal to impose more implicit geometric constraints.

## 2.3 Spatial Propagation for Disparity

Spatial propagation is a popular mechanism in depth completion whose input data is a sparse depth map generated by the sensor of LiDAR. Representative works are [15, 25, 29]. When it comes to the depth estimation task, PAPNet[35] conducts two types of propagation which are task-specific propagation and cross-task propagation, respectively. The cross-task propagation integrates affinity from different tasks including semantic segmentation, surface normal estimation, and depth estimation while the task-specific propagation performs an iterative diffusion in the feature space to widely spread the cross-task affinity patterns. The recent work PatchMatchNet

[23] introduces an iterative multi-scale patchmatch algorithm to learn adaptive propagation for multi-view stereo matching. HitNet [22] imitates its predecessor [14] and performs spatial propagation with the assistance of predicted slanted plane hypotheses.

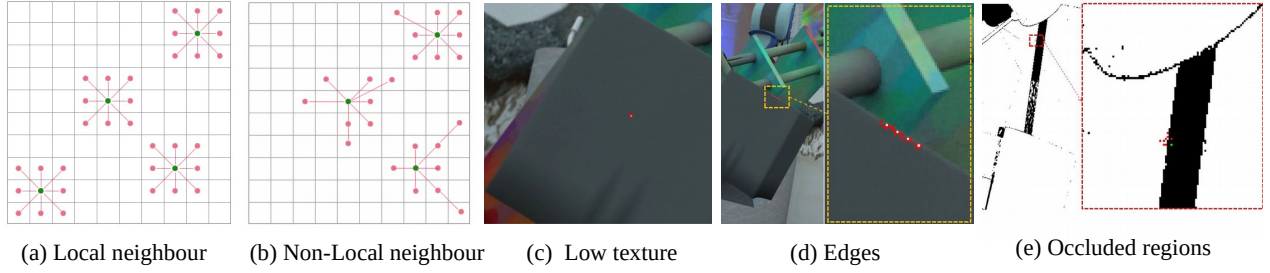
## 3 METHODOLOGY

In this section, we will provide a thorough introduction to our proposed normal incorporated stereo matching network(NINet) in the following order: the architecture of our model is demonstrated in Subsection 3.1 following the instruction of normal estimation subnet in Subsection 3.2. Details about our proposed two modules can be found in Subsection 3.3 and Subsection 3.4. We describe our multi-scale training mechanism in Subsection3.5.

### 3.1 Network Architecture

The overview architecture of our proposed NINet is shown in Figure 2. We proposed a normal incorporated joint learning framework for surface normal and disparity estimation. A weight-sharing 2D residual blocks are used for feature extraction from stereo images which are then fed into two parallel branches for normal estimation and disparity estimation, respectively. The normal estimation will be discussed in Subsection 3.2. As for the stereo matching process, we imitate [12, 24] and employ inner production for calculating the similarity between pixels. Given a pair of stereo feature maps, *i.e.*  $f_l$  and  $f_r \in \mathbb{R}^{H \times W \times C}$ , the correlation is computed as :

$$Corr(d, x, y) = \frac{1}{C} < f_l(x - d, y), f_r(x, y) >, \quad (1)$$



**Figure 3: Illustration of local spatial propagation (a), non-local spatial propagation (b). (c) shows sampled points in low-texture regions. (d) demonstrates sampled points at edges. (e) shows sampled points in occluded regions. Red points indicated the selected ones for propagating disparity to the targeted white/green point. It's obvious that our method successfully learns to dynamically sample points for propagation according to different patterns.**

where  $\langle A, B \rangle$  indicates the inner production between  $A$  and  $B$ . The aggregation part takes a shallow UNet-like network which follows our proposed NDP and ARL modules for disparity estimation at multi scales.

### 3.2 Normal Estimation Sub-Net

Since there is a certain degree of the geometric relationship between disparity and surface normal, we directly use the network structure of DispNetC[12] to estimate the surface normal and use the coarse-to-fine strategy to obtain the surface normal prediction results at different scales with a normalization at the feature dimension. For the reason that Sceneflow and KITTI datasets do not provide ground truth for surface normal, we use a Sobel operator to obtain plane information as the supervision information. See Section 4.2 for the different processing methods of different datasets.

### 3.3 Non-Local Disparity Propagation

Inspired by the spatial propagation mechanism utilized in the depth completion task[4], we introduce a normal incorporated non-local disparity propagation module in which we hub NDP to generate non-local affinities and offsets for spatial propagation at the disparity level. The motivation lies that the sampled pixels for edges and occluded regions are supposed to be selected. The propagation process aggregates disparities via plane affinity relations, which alleviates the phenomenon of parallax blurring at object edges due to frontal parallel windows. And the disparities in occluded areas are also be optimized at the same time by being propagated from non-occluded areas where the predicted disparities with high confidence. The architecture of NDP module can be seen in Figure 2.

In the NDP module, the coarse disparity and its corresponding warped error are firstly utilized to obtain the disparity confidence map through the confidence estimation branch. A sigmoid activation function is used to generate the disparity confidence map. This confidence map can reflect the reliability of the current disparity which benefits the spatial propagation process.

For the disparity spatial propagation, previous works such as [25] use a local fixed  $3 \times 3$  window to consider all the possible directions of neighbor pixels for disparity aggregation. Its process can be shown in Figure3.a. For a specific location  $p$ , we have its

8-closest neighbor  $N_p$ :

$$N_p = \{p + p_n | p_n \in \{(-1, -1), (-1, 0), (1, 0), (1, 1)\}\} \quad (2)$$

After we obtain the affinity map  $A \in R^{H \times W \times 8}$  and confidence map  $C \in R^{H \times W \times 1}$ . The local spatial propagation for a specific position  $p$  can be described as follows:

$$\tilde{d}_p = \sum_{n \in N_p} (A_p(n) \times c_n \times d_n) + (1 - \sum_{n \in N_p} (A_p(n) \times c_n) \times d_p) \quad (3)$$

where  $n$  enumerates the locations in  $N_p$ .  $A_p(n)$ ,  $c_n$  is the corresponding affinity and confidence.  $d_p$  is the initial disparity, and  $\tilde{d}_p$  is the refined disparity.

Unlike conventional spatial propagation using fixed propagation window, in NDP, we directly use the concatenation of pixel-level information to provide the guidance to learn an additional offset  $\Delta p$  and its affinities for each 8 closest neighbors. By applying the non-local offset, the propagation kernel can break through the limitation of the fixed window, and thus enhances the ability to determine the position of the propagation neighborhood adaptively which can be seen in Figure3.b. The new 8 neighbors can be defined as:

$$\tilde{N}_p = \{n + \Delta p_n | n \in N_p, \Delta p_n = f_\theta(I_l, I_r, S, E)\} \quad (4)$$

Here  $\tilde{N}_p$  is the non-local 8 closest neighbour,  $f$  is the residual block in Figure 3 with a learnable parameter  $\theta$ . The network takes left image  $I_l$ , right image  $I_r$ , estimated surface normal  $S$ , and warped error  $E$  as the input to provide the guidance information. Here the surface normal indicates the geometry consistency relations, and the warped error  $E$  reveals the locations of the inaccurate and occluded areas, which is beneficial to increase the directivity and efficiency of the spatial propagation process. Besides, the residual block also predicts the non-local affinities to manifest the similarities between non-local neighbor's disparities and the target disparity. To enhance the stability of the propagation, we apply confidence incorporated affinity normalization by multiplying the predicted non-local affinity values with their corresponding confidence value and then be divided by the absolute sum of all 8 non-local affinities at the target position. For one specific location  $p$ , the normalized affinity  $\tilde{A}_p^{norm}$  is:

$$\tilde{A}_p^{norm} = \left\{ \frac{a_i c_i}{\sum_{i=1}^8 |a_i c_i|} | a_i \in \tilde{A}_p, c_i \in C \right\} \quad (5)$$

By using the non-local neighbour  $\tilde{N}$  and normalized non-local affinity  $\tilde{A}_p^{norm}$ , we apply non-local disparity propagation to refine the disparities. Compared with Equation 3, the non-local propagation process becomes:

$$\tilde{d}_p = \sum_{n \in \tilde{N}_p} (\tilde{A}_p^{norm}(n) \times d_n) + (1 - \sum_{n \in \tilde{N}_p} \tilde{A}_p^{norm}(n)) \times d_p \quad (6)$$

As it shows that we also assign a dynamic weight to the initial disparity which equals 1 minus the summation of the affinities of its 8 non-local neighboring points.

Figure 3 (c)(d)(e) show how our proposed NDP module decides non-local neighbor of specific positions when dealing with edges, low texture, and occluded regions with the guidance of surface normal and warped error. Thus, the proposed NDP module can refine the disparities in these ill positions by non-local propagation.

### 3.4 Affinity-Aware Residual Learning

The non-local disparity propagation module primarily concerns the disparity estimation in bordered and occluded regions, but stereo matching in low texture regions remains a bottleneck. To mitigate this problem, we imitate the idea from EDNet[34] to apply error based spatial attention strategy for efficient residual learning and obtain the error concentrated feature map  $\hat{F}_e \in R^{H \times W \times C}$ . Given the concatenation of stereo images, warped error predicted disparity, and surface normal map, a block of 2D convolution network is introduced to obtain an error-aware feature map  $F_e$ . A sigmoid function is then used for calculating an error-based spatial attention map. The error concentrated feature map  $\hat{F}_e$  is computed as:

$$\hat{F}_e = F \odot (\sigma F_e), \quad (7)$$

where  $F$  is the original feature map,  $\odot$  represents element-wise product, and  $\sigma()$  is sigmoid function.

However, it's insufficient to know where mismatching occurs as the model still lacks intuitive guidance about where to refer for better refinement. Therefore, with the concatenation of stereo images, predicted disparity, warped error map, and normal estimation, we conduct a residual block to calculate a normalized local affinity matrix  $A \in R^{H \times W \times k \times k}$  within a  $k \times k$  neighboring region for each pixel. The error concentrated feature map is further re-weighted with this affinity matrix. The final local affinity-aware feature map is calculated as follows:

$$\tilde{F}_{i,j} = \sum_{a,b=-(k-1)/2}^{(k-1)/2} \hat{F}_{i+a,j+b} \odot A_{i,j}(a,b), \quad (8)$$

where  $\odot$  represents element-wise product. Assisted by this normal guided local affinity matrix, those mismatched pixels in low texture regions can be corrected by referring to their neighbouring regions with similar affinity. The re-weighted feature map is then feed into a deformable hourglass network for further aggregation and get an affinity-aware disparity  $r_a$  to obtain the updated disparity estimation.

### 3.5 Multi-Scale Loss Function

Our network is trained in an end-to-end manner supervised by ground truth disparity and generated ground truth surface normal

at multi scales. Disparity loss  $L_s^d$  and surface normal loss  $L_s^n$  at scale  $s$  are computed as:

$$L_s^d = \frac{1}{N} \sum_{i=1}^N \text{SmoothL1}(d_i^s - \hat{d}_i^s) \quad (9)$$

$$L_s^n = \frac{1}{N} \sum_{i=1}^N \text{SmoothL1}(n_i^s - \hat{n}_i^s) \quad (10)$$

where  $N$  is the number of elements for predicted map.  $\hat{d}_i^s$  and  $\hat{n}_i^s$  indicate the  $i^{th}$  element of the estimated disparity  $\hat{d}^s$  and estimated surface normal  $\hat{n}^s$  while  $d^s$  and  $n^s$  represent the ground truth for disparity and surface normal, respectively. The utilized Smooth L1 loss is defined as follows:

$$\text{SmoothL1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \quad (11)$$

To supervise the confidence  $c$  used in the NDP module, we apply the loss function in [22]:

$$L_s^c(c) = \max(1 - c, 0)_{\tau|d_i^s - \hat{d}_i^s| < C_1} + \max(c, 0)_{\tau|d_i^s - \hat{d}_i^s| > C_2} \quad (12)$$

where  $\tau$  is an indicator function which equals 1 when the condition is satisfied and 0 otherwise. The function aims at increasing the confidence if the distance between predicted disparity and ground truth is smaller than a threshold  $C_1$  and decreasing the confidence if the distance is larger than a threshold  $C_2$ . In our work, we set  $C_1 = 0.5$  and  $C_2 = 1.5$ .

The final loss function is a combination of losses for the disparity, surface normal, and confidence for all pixels across 4 scales:

$$L^{total} = \sum_{s=0}^3 \lambda_s (\lambda_d L_s^d + \lambda_n L_s^n + \lambda_c L_s^c) \quad (13)$$

where  $\lambda_s$  is a scalar for adjusting the loss weight at scale  $s$ .  $\lambda_d$ ,  $\lambda_n$  and  $\lambda_c$  are coefficients for disparity, surface normal and confidence losses, respectively. More details are provided in Section 4.3.

## 4 EXPERIMENT

In this section, we will first introduce the training datasets as well as the detailed training scheme. Then we provide a thorough quantitative analysis of our proposed model on the corresponding datasets.

### 4.1 Datasets and Evaluation Metrics

We validate the effectiveness on three public datasets which are Scene Flow [12], KITTI 2015 [13], and Middlebury [19]. The Scene Flow dataset consists of 39,824 pairs of synthetic stereo RGB images with a full resolution of 960×540 among which 35,454 pairs are split as the training set and the rest 4,370 pairs are split as the testing set. To facilitate the research of autonomous driving, the KITTI 2015 collects 400 pairs of real scenes with sparse depth ground truth annotated by LiDAR, where 200 pairs are for training and the other 200 pairs are utilized for testing. Only 33 pairs of stereo images are available for the Middlebury2014 dataset so we conduct a generalization evaluation on it. The standard end-point error (EPE) is reported on the Scene Flow dataset and we follow the standard evaluation protocol to submit our predicted results to the KITTI Benchmark from where the D1 error is obtained.





**Figure 4: Visualization of our generated surface normal ground truth on KITTI 2015.** Row one is the left image, row two is the gt sparse disparity, row three is the sparse surface normal generated with neighbour-existing disparities. The last row is our generated disparity with the assistance of pseudo disparity.

#### 4.2 Normal Estimation Supervision

For the Scene Flow dataset, the provided disparity labels are pixel-wise annotated which makes it possible to directly introduce the Sobel kernel for calculating gradient along  $x$  and  $y$  axis, respectively. The primary challenge comes from the KITTI datasets because of the sparse disparity annotations. To tackle this problem, we first leverage the state-of-the-art model LEAStereo [3] to generate the dense pseudo disparity labels following the Sobel kernel to calculate the pseudo normal ground truth  $N_{pseudo}$ . Then we search for the local region where the disparity ground truth is available for the neighboring 8 pixels to obtain the normal ground truth  $N_{sparse}$  as well as the ground truth position mask  $M_{sparse}$ . The final normal ground truth  $N_{gt}$  is then calculated as follows:

$$N_{gt} = N_{pseudo} * (1 - M_{sparse}) + N_{sparse} * M_{sparse} \quad (14)$$

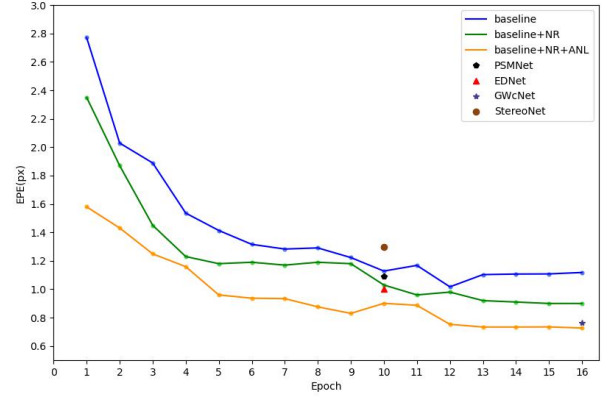
To fully exploit the limited sparse normal information and alleviate the influence of pseudo disparity of low quality, we further calculate the end-point error between the pseudo disparity labels and the sparse disparity ground truth to obtain an index  $E_{index}$  where the EPE is greater than 1 disparity. The loss of normal estimation  $L_N$  is computed as follows:

$$\begin{aligned} L_N = & S(N_{gt}, \hat{N}) * M_{sparse} * 1.5 \\ & + S(N_{gt}, \hat{N}) * (1 - M_{sparse}) * E_{index} * 1.0 \\ & + S(N_{gt}, \hat{N}) * (1 - M_{sparse}) * (1 - E_{index}) * 0.5, \end{aligned} \quad (15)$$

where  $S(x, y)$  denotes the SmoothL1 loss function mentioned before and  $\hat{N}$  represents the predicted normal. Our generated surface normal ground truth is shown in Figure 4

#### 4.3 Implementation Details

We implemented our NINet by PyTorch[16] and trained our model with Adam(momentum=0.9, beta=0.999) as the optimizer. We randomly cropped the input to 320×640 on the KITTI 2015 & 2012 and the Scene Flow datasets. The whole training process is performed on 2 NVIDIA RTX 2080ti GPUs with a batch size of 8. The max disparity search range is set as 192 pixels. The coefficients of four



**Figure 5: Comparison among our proposed methods with different settings on the Scene Flow testing set as well as other SOTA works.** It is obvious that our proposed modules significantly facilitate the speed of convergence.

disparity losses at each scale are correspondingly set as  $\lambda_0 = 1.0$ ,  $\lambda_1 = \lambda_2 = 0.8$ , and  $\lambda_3 = 0.6$ . The weights for disparity, surface normal and confidence are set as  $\lambda_d = 5$ ,  $\lambda_n = 50$ ,  $\lambda_c = 1$ .

For the Scene Flow dataset, we trained the networks for 150 epochs in total with an initial learning rate of 1e-3. After the first 20 epochs of training, we decrease our learning rate to 4e-4, 1e-4, 4e-5 at the epoch of 20, 40, 100. Note that to enhance the reliability of spatial propagation guided by normal, the normal estimation branch is trained independently for 40 epochs before the joint learning of disparity and normal.

For the KITTI 2015 and the KITTI 2012 datasets, we trained our networks based on the model pre-trained on the Scene Flow dataset. We first mixed these two datasets with 300 pairs of images in total for the first 200 epochs of training to obtain the model with the best performance on the validation set. Another 200 epochs of fine-tuning are employed on the KITTI 2015 and the KITTI 2012 datasets respectively. The learning rate is constantly set as 1e-5 for the whole training process.

#### 4.4 Ablation Study

Based on the performance shown in Table 1, it can be directly analyzed that our proposed non-local disparity propagation and affinity-aware residual learning modules effectively improve the performance. The joint learning scheme with surface normal improves the baseline with an obvious margin. The employment of the affinity-aware residual learning module which combines both attention and dynamic filtering operations significantly decreases the EPE to 0.60. The disparity level non-local propagation further improves the result as reflected by the EPE of baseline-ARL-NDP, which is 0.51. As shown in Figure 6, remarkable improvement can be achieved with the help of our proposed modules. The second column provides straightforward evidence that our NINet successfully preserves the detailed outline of the shelf with consistent disparity while other settings fail to reach this result. To further provide a more comprehensive interpretation of the superiority of our method, we visualize the converging speed of our methods as well as other competing approaches as shown in Figure 5.

**Table 1: Ablation study of our proposed method on the Scene Flow dataset. The best performance is stressed in bold. Our proposed affinity-aware residual learning and non-local spatial propagation modules significantly improve the disparity estimation. The best results are stressed in bold. Our proposed modules efficiently improve the stereo matching performance.**

Method	NR Branch	Affinity-aware Residual		Spatial Propagation		EPE	P1 Error	P3 Error
		attention	dynamic filtering	local	non-local			
Baseline						0.96	12.65%	5.94%
Baseline-R	✓					0.73	9.58%	4.72%
Baseline-RD	✓		✓			0.66	8.06%	4.42%
Baseline-RA	✓	✓				0.63	7.90%	4.18%
Baseline-ARL	✓	✓	✓			0.60	7.42%	4.00%
Baseline-LDP	✓			✓		0.65	8.02%	4.20%
Baseline-NDP	✓				✓	0.60	7.39%	3.98%
Baseline-ARL-LDP	✓	✓	✓	✓		0.54	6.65%	3.71%
Baseline-ARL-NDP	✓	✓	✓		✓	<b>0.51</b>	<b>6.17%</b>	<b>3.53%</b>

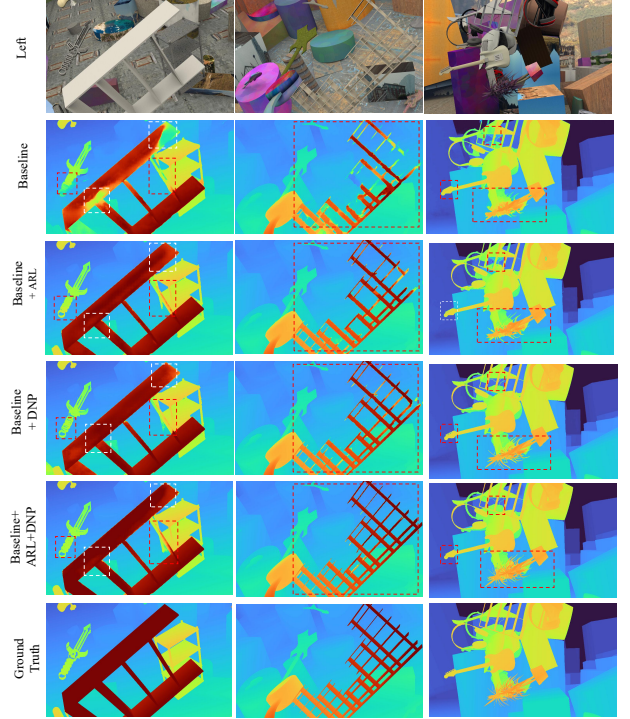
**Table 2: Comparison of our proposed method and other remarkable works on the Scene Flow dataset. Among all the competing methods, our NINet ranks the 3rd for end-point error. The best performance is emphasised in bold.**

Model	DispNetC[12]	PSMNet[2]	GwcNet[6]	EDNet[34]	HITNet[22]	LeaStereo[3]	CSPN[25]	ACVNet[27]	NINet (Ours)
EPE	1.68	1.09	0.76	0.63	<b>0.41</b>	0.78	0.78	0.48	0.51

**Table 3: Benchmark results on KITTI 2015 test sets. “Noc” and “All” indicate the percentage of outliers averaged over ground truth pixels of non-occluded and all regions, respectively. “fg” and “all” indicate the percentage of outliers averaged over the foreground and all ground truth pixels, respectively. Our NINet achieves the best performance on the foreground pixels with a competitive inference speed.**

Method	Noc (%)		All (%)		Time (s)
	fg	all	fg	all	
GANet [33]	3.37	1.73	3.82	1.93	0.36
GCNet [7]	5.58	2.61	6.16	2.87	0.9
PSMNet [2]	4.31	2.14	4.62	2.32	0.41
GwcNet [6]	3.49	1.92	3.93	2.11	0.32
SegStereo [30]	3.70	2.08	4.07	2.25	0.6
HD <sup>3</sup> [32]	3.43	1.87	3.63	2.02	0.14
Bi3D [1]	3.11	1.79	3.48	1.95	0.48
AANet [28]	4.93	2.32	5.39	2.55	0.075
DispNetC [12]	3.72	4.05	4.41	4.34	0.06
FADNet [24]	3.07	2.59	3.50	2.82	0.05
EDNet[34]	3.33	2.31	3.88	2.53	0.05
ACVNet[27]	2.84	1.52	3.07	<b>1.65</b>	0.20
CSPN[25]	2.67	1.61	2.88	1.74	1.0
HITNet [22]	2.72	1.74	3.20	1.98	0.02
LEAStereo[3]	2.65	<b>1.51</b>	2.91	<b>1.65</b>	0.30
NINet (Ours)	<b>2.53</b>	1.83	<b>2.80</b>	1.98	0.17

With the assistance of the proposed modules, our method not only achieves the lower initial EPE but outperforms other works at the corresponding epochs as well. Note that, the competitive result of GwcNet is mainly attributed to the application of 3D convolution which is discarded in our work.



**Figure 6: Visualization of the ablation study on Scene Flow dataset. With the assistance of our proposed modules, thin structures and more consistent disparity can be preserved.**

## 4.5 Performance Evaluation

In this subsection, we will mainly focus on the comparison between our method and other top-performing methods.

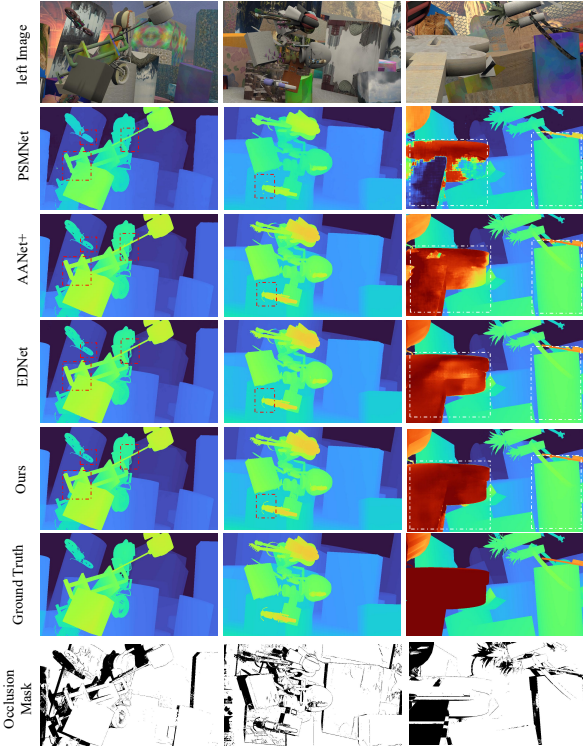


Figure 7: SceneFlow performance compares with other superior works.

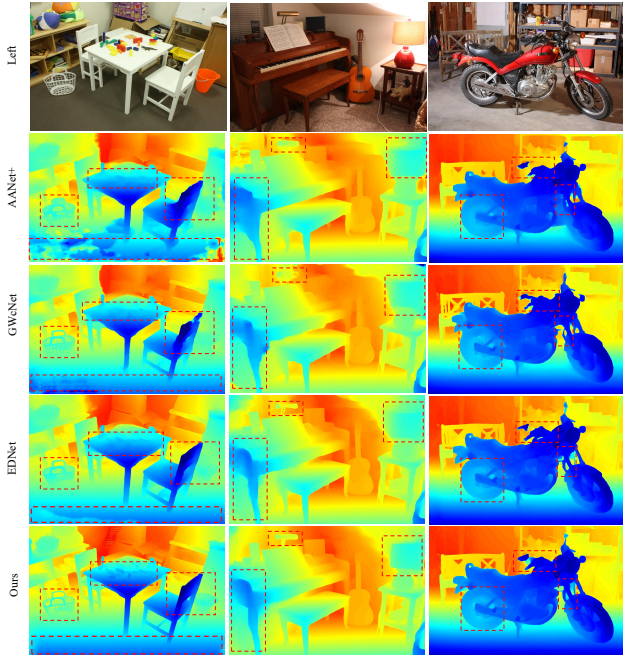


Figure 8: Generalization ability on MiddledyBurry 2014 dataset.

Table 4: Evaluation for bordered and occluded regions on the Scene Flow dataset with our proposed modules.

Method	Bordered EPE	Occluded EPE
Baseline	1.95	2.31
+ARL	1.63	1.92
+NDP	1.44	1.81
+NDP+ARL	1.31	1.71

Table 5: Generalization evaluation on the Middlebury 2014 dataset. Our NINet achieves the best generalizability.

Method	Bad2.0		Bad4.0		RMSE		AvgErr	
	Occ	Non	Occ	Non	Occ	Non	Occ	Non
AANet	38.6	42.5	25.3	28.9	23.3	21.4	12.4	10.7
EDNet	69.5	66.8	57.1	54.0	21.9	20.4	44.2	41.2
GwcNet	34.9	30.8	22.7	19.4	27.0	25.4	13.4	11.8
FADNet	31.3	27.0	20.9	17.1	21.8	19.6	11.2	9.5
NINet(Ours)	<b>24.5</b>	<b>20.2</b>	<b>15.8</b>	<b>12.5</b>	<b>18.8</b>	<b>18.3</b>	<b>8.0</b>	<b>7.8</b>

**Scene Flow** As shown in Table 2, our method outperforms most of the superior approaches including LEAStereo [3] and ranks 3rd among all the competing works. As illustrated in Figure 7, our method produces sharper and more detailed disparity estimation. A representative instance comes from the third column that our NINet achieves consistent disparity estimation for the plank while results from other methods are less convincing. Table 4 specifically reveals the effectiveness of proposed modules in both bordered and occluded regions.

**KITTI 2015** Table 3 demonstrates the performance of several remarkable methods in the KITTI 2015 leaderboard. When evaluating disparity estimation in foreground areas, our approach even outperforms the HITNet by 6.99% in non-occluded regions and ranks 1st among all the published works, which strongly validates the effectiveness of our model.

**Middlebury 2014** Generalizability is also a necessary indicator when comprehensively evaluating learning-based stereo matching methods and thus we conduct a generalization experiment on Middlebury 2014 dataset. As shown in Table 5, our proposed method obtains the best performance when transferring to a new data domain without finetuning. Compared with 3D convolution-based GwcNet, our method decreases the average error from 13.4 to 8.0 and 11.8 to 7.8 in occluded regions and non-occluded, respectively.

## 5 CONCLUSION

In this paper, we propose a normal incorporated joint learning framework named NINet to explicitly leverage the plane information for intuitive geometric guidance. A non-local spatial propagation (NDP) module is also introduced to dynamically select points with a similar affinity for further propagation at the disparity level. An affinity-aware residual learning (ARL) module is then designed to impose consistency within the same plane according to a local affinity matrix. Extensive experiments on several datasets have demonstrated the effectiveness as well as the generalization ability of our proposed method. By the time we finished it, our method ranked 1st on the KITTI 2015 dataset for foreground objects.



## REFERENCES

- [1] A. Badki, A. Troccoli, K. Kim, J. Kautz, P. Sen, and O. Gallo. 2020. Bi3D: Stereo Depth Estimation via Binary Classifications. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1597–1605.
- [2] J. Chang and Y. Chen. 2018. Pyramid Stereo Matching Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5410–5418.
- [3] Xuelian Cheng, Yiran Zhong, Mehrtaash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. 2020. Hierarchical Neural Architecture Search for Deep Stereo Matching. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 22158–22169. <https://proceedings.neurips.cc/paper/2020/file/fc146be0b230d7e0a92e66a6114b840d-Paper.pdf>
- [4] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox. 2015. FlowNet: Learning Optical Flow with Convolutional Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 2758–2766.
- [5] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. 2020. Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2492–2501. <https://doi.org/10.1109/CVPR42600.2020.00257>
- [6] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li. 2019. Group-Wise Correlation Stereo Network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3268–3277.
- [7] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry. 2017. End-to-End Learning of Geometry and Context for Deep Stereo Regression. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 66–75.
- [8] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. 2018. StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- [9] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. 2020. Normal Assisted Stereo Depth Estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2186–2196. <https://doi.org/10.1109/CVPR42600.2020.00226>
- [10] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang. 2018. Learning for Disparity Estimation Through Feature Constancy. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2811–2820.
- [11] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Wei Li, Christian Theobalt, Ruigang Yang, and Wenping Wang. 2021. Adaptive Surface Normal Constraint for Depth Estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 12829–12838. <https://doi.org/10.1109/ICCV48922.2021.01261>
- [12] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. 2016. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4040–4048.
- [13] Moritz Menze and Andreas Geiger. 2015. Object scene flow for autonomous vehicles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3061–3070. <https://doi.org/10.1109/CVPR.2015.7298925>
- [14] Christoph Rhemann Michael Bleyer and Carsten Rother. 2011. PatchMatch Stereo - Stereo Matching with Slanted Support Windows. In *Proceedings of the British Machine Vision Conference*. 14.1–14.11. <http://dx.doi.org/10.5244/C.25.14>.
- [15] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. 2020. Non-local Spatial Propagation Network for Depth Completion. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 120–136.
- [16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., 8026–8037. <https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf>
- [17] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiayia Jia. 2018. GeoNet: Geometric Neural Network for Joint Depth and Surface Normal Estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 283–291. <https://doi.org/10.1109/CVPR.2018.00037>
- [18] Mattia Rossi, Mireille El Gheche, Andreas Kuhn, and Pascal Frossard. 2020. Joint Graph-Based Depth Refinement and Normal Estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12151–12160. <https://doi.org/10.1109/CVPR42600.2020.01217>
- [19] D. Scharstein, H. Hirschmuller, Y. Kitajima, G. Krathwohl, N. Nesić, and P. Westling X Wang. 2014. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition (GCPR)*.
- [20] D. Scharstein, R. Szeliski, and R. Zabih. 2001. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*.
- [21] Zhelun Shen, Yuchao Dai, and Zhibo Rao. 2021. CFNet: Cascade and Fused Cost Volume for Robust Stereo Matching. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13901–13910. <https://doi.org/10.1109/CVPR46437.2021.01369>
- [22] Vladimir Tankovich, Christian Häne, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. 2021. HITNet: Hierarchical Iterative Tile Refinement Network for Real-time Stereo Matching. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14357–14367. <https://doi.org/10.1109/CVPR46437.2021.01413>
- [23] Fangjinhuo Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. 2021. PatchmatchNet: Learned Multi-View Patchmatch Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14194–14203.
- [24] Qiang Wang, Shaohuai Shi, Shizhen Zheng, Kaiyong Zhao, and Xiaowen Chu. 2020. FADNet: A Fast and Accurate Network for Disparity Estimation. In *2020 IEEE International Conference on Robotics and Automation (ICRA 2020)*. 101–107.
- [25] Tianyuan Wang, Can Ma, Haoshan Su, and Weiping Wang. 2021. CSPN: Multi-Scale Cascade Spatial Pyramid Network for Object Detection. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1490–1494. <https://doi.org/10.1109/ICASSP39728.2021.9414883>
- [26] Xiaolong Wang, David F. Fouhey, and Abhinav Gupta. 2015. Designing deep networks for surface normal estimation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 539–547. <https://doi.org/10.1109/CVPR.2015.7298652>
- [27] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. 2022. ACVNet: Attention Concatenation Volume for Accurate and Efficient Stereo Matching. *arXiv e-prints*, Article arXiv:2203.02146 (March 2022), arXiv:2203.02146 pages. arXiv:2203.02146 [cs.CV]
- [28] H. Xu and J. Zhang. 2020. AANet: Adaptive Aggregation Network for Efficient Stereo Matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1956–1965.
- [29] Zheyuan Xu, Hongche Yin, and Jian Yao. 2020. Deformable Spatial Propagation Networks For Depth Completion. In *2020 IEEE International Conference on Image Processing (ICIP)*. 913–917. <https://doi.org/10.1109/ICIP40778.2020.9191138>
- [30] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiayia Jia. 2018. SegStereo: Exploiting Semantic Information for Disparity Estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- [31] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. 2019. Enforcing Geometric Constraints of Virtual Normal for Depth Prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 5683–5692. <https://doi.org/10.1109/ICCV.2019.00578>
- [32] Z. Yin, T. Darrell, and F. Yu. 2019. Hierarchical Discrete Distribution Decomposition for Match Density Estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6037–6046.
- [33] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr. 2019. GA-Net: Guided Aggregation Net for End-To-End Stereo Matching. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [34] Songyan Zhang, Zhicheng Wang, Qiang Wang, Jinshuo Zhang, Gang Wei, and Xiaowen Chu. 2021. EDNet: Efficient Disparity Estimation with Cost Volume Combination and Attention-based Spatial Residual. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5429–5438. <https://doi.org/10.1109/CVPR46437.2021.00539>
- [35] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. 2019. Pattern-Affinitive Propagation Across Depth, Surface Normal and Semantic Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4101–4110. <https://doi.org/10.1109/CVPR.2019.00423>
- [36] Wang Zhao, Shaohui Liu, Yi Wei, Hengkai Guo, and Yong-Jin Liu. 2021. A Confidence-based Iterative Solver of Depths and Surface Normals for Deep Multi-view Stereo. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 6148–6157. <https://doi.org/10.1109/ICCV48922.2021.00611>

## A NORMAL ESTIMATION SUB-NET

Our Normal Estimation Sub-Net is illustrated in Figure 11. We directly use the shared encoder part to provide the feature and adopt the network structure of DispNetC[1] to estimate the surface normal in a coarse-to-fine manner. For residual learning(Residual Blocks in blue ), we use a shallow U-Net architecture to take features at the current scale together with upsampled predicted surface normal at the previous scale and left-right image pairs as input to predict the surface normal residual at the current scale. Besides, we normalized the output of each scale directly on the feature channels(3), making it more in line with the definition of the surface normal. Figure 9 illustrate visualization results on SceneFlow and KITTI 2015 and MiddleBurry 2014 datasets.

## B NON-LOCAL DISPARITY PROPAGATION AT DIFFERENT SCALES.

In this section, by visualizing the disparities estimation results at different resolutions, we show how our proposed NDP module improves the quality of the predicted disparity through spatial propagation, especially in thin structures, edges, and occluded regions. As is shown in Figure 12, the non-local disparity propagation based on surface normal and other semi-context information performs well at thin structures and edges especially at low scales to alleviate the disparity discontinuity and blurring issues. After spatial propagation, our disparity preserves more structural details, which facilitates more accurate depth results for downstream tasks and thus constructs more accurate AR/VR applications.

## C SPATIAL ATTENTION AND AFFINITIES DISTRIBUTION IN ARL MODULE

As mentioned in the main paper, we have proposed an ARL module that manages to optimize disparity estimation results at the feature level. There are two main tricks in the ARL module for efficient residual learning. One is to apply the dynamic spatial attention mechanism to generate attention maps at different refinement scales. The attention maps imply where should be highlighted as well as where should be ignored. It promotes the network to efficiently optimize the disparity estimation results for different locations at different scales.

As shown in Figure 10, the dynamic spatial attention mechanism tends to assist residual learning at different levels and focus on regions with different attributes. The attention distribution at the 1/4 scale mainly lies in the background objects, while the distribution of attention at 1/2 resolution is highly consistent with that of the occlusion mask. This enables the ARL module to be occlusion-aware when refining the disparity. Besides, the attention map at full scale shows bigger attention to foreground object and planes that is texture-less, which benefit the refinement in such regions.

The other trick of the ARL module is the local affinity filtering module whose intermediate affinity visualization result of the local 8 neighbors can be seen in Figure 13. As the figure illustrated, the red

arrow in reference images shows the plane direction, as we can see that the affinity value at the left-top value is bigger than the other direction, which indicates the direction of the feature aggregation should be. By applying the local affinity filtering module, our ARL module can make full use of surface normal information for better feature aggregation.

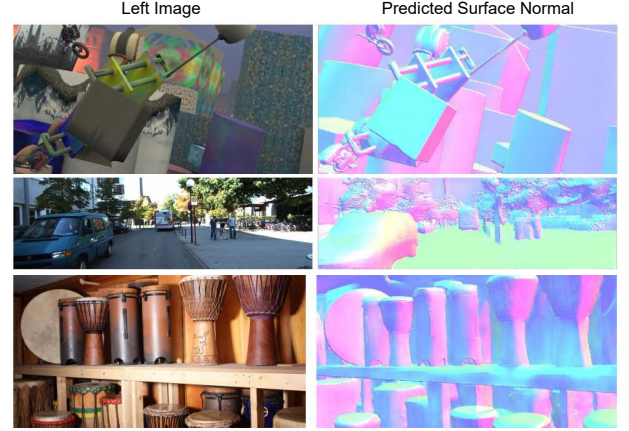


Figure 9: Normal Estimation Result on SceneFlow, KITTI2015 and MiddleBurry testing set.

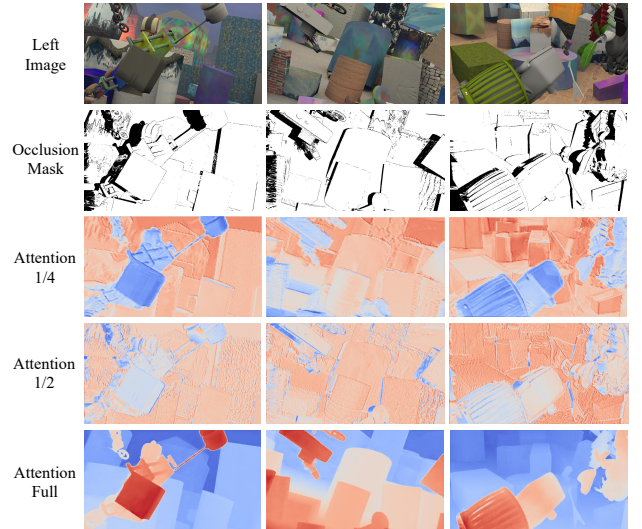
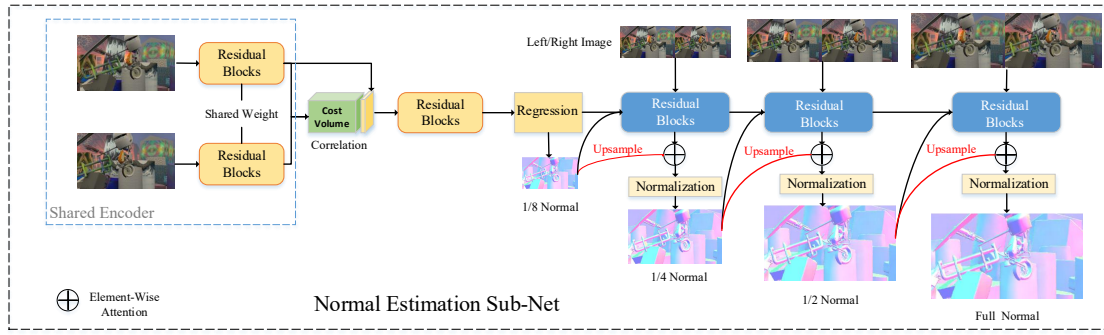
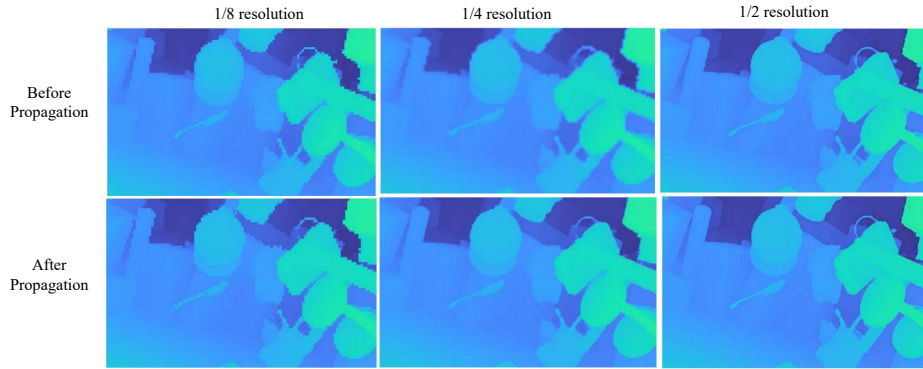


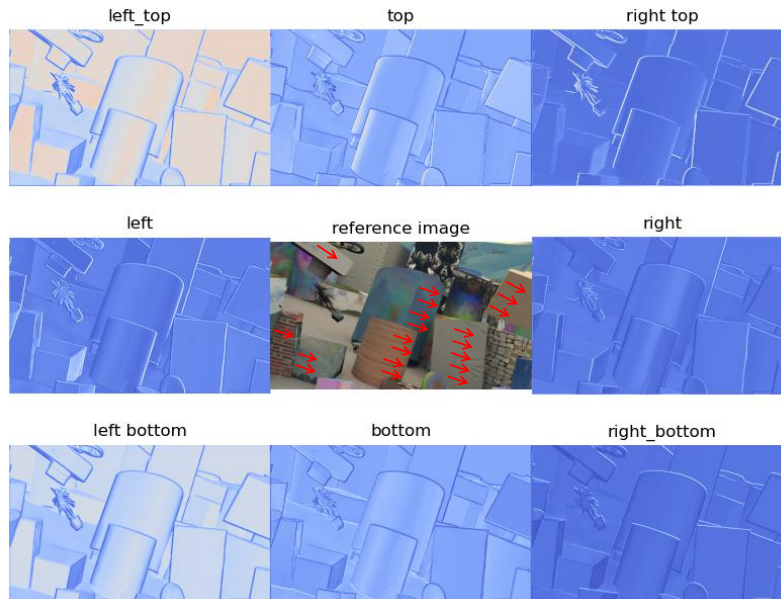
Figure 10: Attention Maps at different scales(Range from 1/4 full scale). Different scales pay different attention to different regions, such as foreground, texture-less planes, occluded areas and background.



**Figure 11: Normal Estimation Sub-Net Architecture.** The feature encoder is shared with the disparity estimation branch. The whole Sub-Net adopts the DispNetC [12] architecture with a residual learning module. We normalized the predicted disparity to satisfy the geometry constrain.



**Figure 12: Disparity refinement at different scales.** It clearly shows that our proposed non-local propagation witness a clear edges and structures at different disparity scales, which alleviates the blurring and breakage issues at the edges of the image.



**Figure 13: A sample of affinity distribution of local 8 neighbors in ARL module.**