# Adaptive Fusion of Single-View and Multi-View Depth for Autonomous Driving

Junda Cheng[1,2][*], Wei Yin[2][*], Kaixuan Wang[2], Xiaozhi Chen[2], Shijie Wang[1], Xin Yang[1][†]

[1] Huazhong University of Science and Technology    [2] DJI Technology

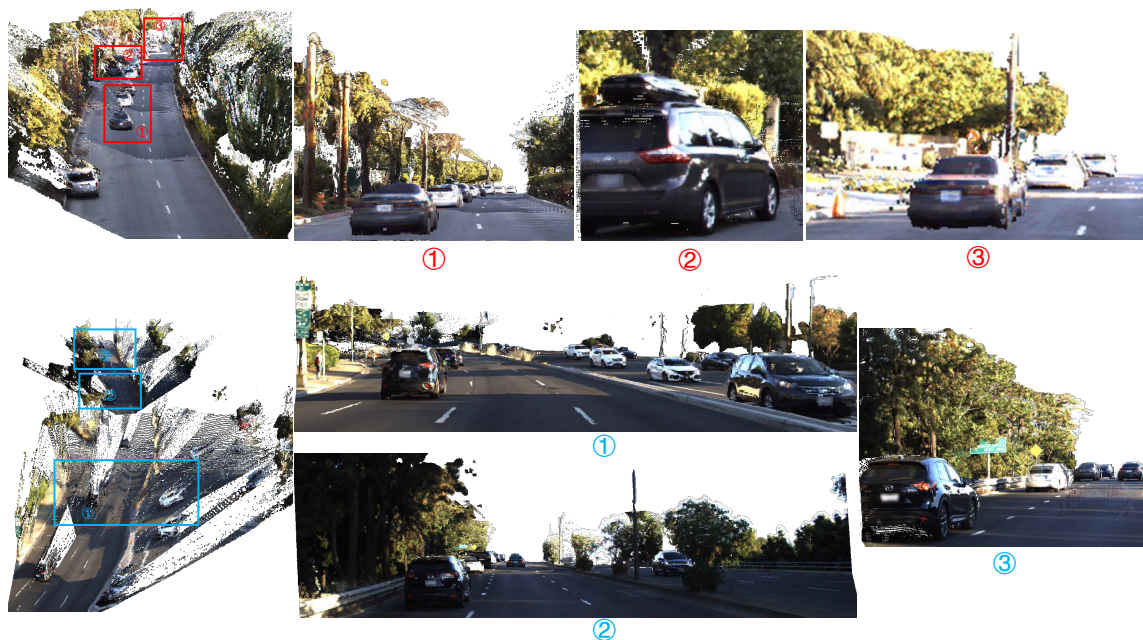{Junda Cheng, sjw, xinyang2014}@hust.edu.cn, yvanwy@outlook.com

Figure 1. Visualization of reconstructed 3D point clouds of DDAD [12] scenes. We fuse the results of 10 frames (including the dynamic object cars) and zoom in on some details for visualization. It shows that our method can achieve high-quality results on both static and dynamic parts.

## Abstract

*Multi-view depth estimation has achieved impressive performance over various benchmarks. However, almost all current multi-view systems rely on given ideal camera poses, which are unavailable in many real-world scenarios, such as autonomous driving. In this work, we propose a new robustness benchmark to evaluate the depth estimation system under various noisy pose settings. Surprisingly, we find current multi-view depth estimation methods or single-view and multi-view fusion methods will fail when given noisy pose settings. To address this challenge, we propose a single-view and multi-view fused depth estimation system, which adaptively integrates high-confident multi-view and single-view results for both robust and accurate depth estimations. The adaptive fusion module performs fusion by dynamically selecting high-confidence regions between two branches based on a wrapping confidence map. Thus, the system tends to choose the more reliable branch when facing textureless scenes, inaccurate calibration, dynamic objects, and other degradation or challenging conditions. Our method outperforms state-of-the-art multi-view and fusion methods under robustness testing. Furthermore, we achieve state-of-the-art performance on challenging benchmarks (KITTI and DDAD) when given accurate pose estimations. Project website: https://github.com/Junda24/AFNet/.*

---

[*]Equal contribution.
[†]Corresponding author.

# 1. Introduction

Depth estimation from images is a long-standing problem in computer vision with wide applications. For vision-based autonomous driving systems, perceiving depth is an indispensable module for understanding the correlation of road objects and modeling 3D environment maps. Since deep neural networks are applied to solve various vision problems, CNN-based methods [2–5, 13, 39–42, 44, 46, 48, 52] have dominated various depth benchmarks.

According to the input formats, they are mainly categorized into multi-view depth estimation [3, 13, 23, 26, 44, 45, 51, 53] and single-view depth estimation [14–16, 19, 37, 38]. Multi-view methods estimate depth under the assumption that given correct depth, camera calibration, and camera poses, the pixels should be similar across views. They rely on epipolar geometry to triangulate high-quality depth. However, the accuracy and robustness of multi-view methods heavily rely on the geometric configuration of the camera and the correspondences matching among views. First, the camera is required to translate sufficiently for triangulation. In autonomous driving scenarios, the car may stop at traffic lights or turn around without moving forward, which causes failure triangulation. Furthermore, the multi-view methods suffer from dynamic objects and textureless regions, which are ubiquitous in autonomous driving scenarios. Another problem is SLAM pose optimization on moving vehicles. Noises are inevitable in existing SLAM methods, not to mention challenging and inevitable situations. For example, one robot or autonomous car can be deployed for years without re-calibration, causing noisy poses. In contrast, as single-view methods [14–16, 19, 37, 38] rely on the semantic understanding of the scene and the perspective projection cues, they are more robust to textureless regions, dynamic objects, and not rely on camera poses. However, its performance is still far from the multi-view methods because of scale ambiguity. Here, we tend to think about if both methods' benefits can be well combined for robust and accurate monocular video depth estimation in autonomous driving scenarios.

Although the fusion-based systems have been explored in previous work [1, 9], they all assume ideal camera poses. The consequence is fusion system's performance is even worse than single-view depth estimation given noise poses. To solve this problem, we propose a novel adaptive fusion network to exploit the advantages of both the multi-view and single-view methods and mitigate their disadvantages, which maintain high precision and also improve the robustness of the system under noisy poses. Specifically, we propose a two-branch network, i.e. one targets monocular depth cues, while the other one leverages the multi-view geometry. Two branches both predict a depth map and a confidence map. To supplement the semantic cues and edge details lost in the cost aggregation of the multi-view branch,

we first fuse the monocular features in the decoder part. We further design an adaptive fusion (AF) module to achieve the final merged depth. Apart from the predicted confidence, we design another wrapping confidence map by performing the multi-view texture consistency check with the predicted depth and provided camera poses. We can easily notice the inconsistency projection when the pose or the depth is inaccurate, or dynamic objects appear. By using such proposed confidence maps to perform the pixel-wise fusion, we can achieve a much more robust depth finally. Our contributions are summarized below.

- We propose AFNet to adaptively fuse the single-view and multi-view depth for more robust and accurate depth estimation. It achieves the state-of-the-art performance on both KITTI [11] and DDAD [12] datasets.
- We are the first to propose the multi-view and single-view depth fusion network for alleviating the defects of the existing multi-view methods which will fail under noisy pose. We propose a new robustness testing benchmark to explore the effectiveness of the multi-view methods given noisy poses to prove this. Our method outperforms all other classical multi-view methods under noisy poses.
- Our AF module can improve the performance of the dynamic object regions which cannot be well processed by the classical multi-view depth estimation methods.

# 2. Related work

**Monocular depth estimation.** Monocular depth estimation is a classical problem in computer vision. Recent CNN-based methods [2, 20, 21, 46, 48, 52] mainly formulate it as a per-pixel classification [2, 10, 46] or regression problem [14, 18, 19, 38]. To boost the performance, some methods [28, 52] propose to aggregate stronger vision features, some methods propose various losses [2, 46, 48], and some also propose to leverage mix-data training [7, 47, 49]. Although their performance on various benchmarks has been improved continuously, the state-of-the-art accuracy is still far from the multi-view geometry-based methods. In our work, we integrate the single-view depth estimation module in our system because of such methods' robustness to low-texture regions and dynamic objects.

**Multi-View depth estimation.** A variety of works have been proposed to estimate the depth based on multi-view observation with known intrinsics and camera poses. [53] is the first one to bring the power of feature learning into multi-view stereo, but they process the matching costs with the traditional aggregation method. [44] proposed to first construct a differentiable cost volume and then use 3D CNNs to regularize the cost volume obtaining the most advanced accuracy at that time. Most recent state-of-the-art methods follow such a paradigm [3, 23, 26, 51]. But these methods have a strong dependency on high-parallax motion and heterogeneous-texture scenes for high accuracy, and

can't handle dynamic objects, etc. In addition, these methods require accurate relative pose between multiple frames and have poor anti-interference ability to pose noise.

**Single and multi-View depth fusion.** There are also some methods [1, 9, 43] for integrating single-view and multi-view prediction to exploit the properties of these two methods. But they basically focus on improving accuracy and efficiency. [9] improves the accuracy by combining the local structure of single-view and the reliable prediction of multi-view in high-parallax and high-gradient regions. [1] first predicts a depth through a single-view network, then builds a thin cost volume based on this initial depth to reduce the high computation overhead and achieve a certain accuracy improvement. However, the above two methods all ignore a problem: multi-view branches are highly dependent on the accuracy of the pose. With noisy poses, the accuracy decreases seriously, even worse than with single-view methods. [43] proposed an Epipolar Attention Module to fuse single-view and multi-view geometric information, and installing attention modules at different resolutions can alleviate the problem of pose inexact to a certain extent. However, this approach can only slightly alleviate the pose noise problem, the wrong matching information extracted through the wrong poses is still integrated into the single-view branch, leading to the decay of accuracy.

Thus, we propose AFNet which adaptively combines the reliable results of the two branches through the AF module to enhance the robustness of the system under noisy poses.

## 3. Method

### 3.1. Method Overview

We aim to estimate the depth $d \in \mathcal{R}^{H \times W}$ for a reference image $\boldsymbol{I}_0 \in \mathcal{R}^{H \times W \times 3}$, given $n$-1 source images $\{\boldsymbol{I}_i\}_{i=1}^{n-1}$, camera intrinsics and camera poses. Figure 2 overviews our AFNet, which consists of three parts, i.e. single-view depth module, multi-view depth module, and adaptive fusion module. Furthermore, we propose a pose correction module to ensure robustness under large noise poses, the details can be found in the supplementary materials.

### 3.2. Single-view and Multi-view Depth Module

In our system, we use ConvNeXt-T [22] as the backbone to extract 4 scales features $\boldsymbol{F}_{i,l}$ ($l$ = 1, 2, 3, 4), where $i$ is the index of the image and $l$ is the scale. The extracted 4-scale feature dimensions are $C = 96, 192, 384, 768$ respectively.

**Single-view branch.** Following [43], we construct a multi-scale decoder to merge backbone features and obtain the depth feature $\boldsymbol{D}_s \in \mathcal{R}^{H/4 \times W/4 \times 257}$. By applying the softmax along the channel dimension for the first 256 channels of the $\boldsymbol{D}_s$, we get the depth probability volume $\boldsymbol{P}_s \in \mathcal{R}^{H/4 \times W/4 \times 256}$. The last channel of the feature serves as the confidence map $\boldsymbol{M}_s \in \mathcal{R}^{H/4 \times W/4}$ for the

single-view depth. Finally, the single-view depth is computed by the soft-weighted sum. It is as follows.

$$\boldsymbol{d}_s = \sum_{d \in \mathbb{B}} d \cdot p_d, \qquad (1)$$

where $\mathbb{B}$ denotes the bins uniformly sampled in the log space from $d_{min}$ to $d_{max}$, which represents the depth search range, $p_d$ denotes the corresponding probability in $\boldsymbol{P}_s$.

**Multi-view branch.** The multi-view branch shares the backbone with the single-view branch to extract features $\boldsymbol{F}_{i,l}$ for the reference and source images. We adopt deconvolutions to deconvolve the low-resolution features to the quarter resolution and combine them with the initial quarter features $\boldsymbol{F}_{i,1}$, which is used to construct the cost volume. By warping the source features into hypothesis planes of the reference camera follow [44], the feature volumes are formed. For robust matching information without much computation, we retain the channel dimension of the feature and construct 4D cost volume, and then reduce the channel number to 1 through two 3D convolution layers. The sampling method of the depth hypothesis is consistent with the single-view branch, but the sampling number is only 128, i.e. the initial cost volume $\boldsymbol{C}_{init} \in \mathcal{R}^{H/4 \times W/4 \times 128}$. Then we use a stacked 2D hourglass network for regularizing to obtain the final multi-view cost volume $\boldsymbol{C}_m \in \mathcal{R}^{H/4 \times W/4 \times 256}$. To supplement the rich semantic information of the single-view feature and the details lost due to cost regularization, we use a residual structure to combine the single-view depth feature $\boldsymbol{D}_s$ and cost volume to obtain the fuse depth feature $\boldsymbol{F}_{fuse}$ as follows:

$$\boldsymbol{F}_{fuse} = \text{Conv}\left\{\text{Concat}\left\{\boldsymbol{C}_m, \boldsymbol{D}_s\right\}\right\} + \boldsymbol{C}_m. \qquad (2)$$

After using a 2D hourglass network for aggregating the $\boldsymbol{F}_{fuse}$, the subsequent operation is exactly the same as the single-view branch. Confidence map of multi-view branch $\boldsymbol{M}_m \in \mathcal{R}^{H/4 \times W/4}$ and the final depth prediction $\boldsymbol{d}_m$ are obtained in the same regression way.

### 3.3. Adaptive Fusion Module

To obtain the final accurate and robust prediction, we design the AF module to adaptively select the most accurate depth between the two branches as the final output as shown in Figure 2. We conduct fusion through three confidence maps, two of which are the confidence maps $\boldsymbol{M}_s$ and $\boldsymbol{M}_m$ generated by the two branches respectively (Section 3.2), and the most critical one is the confidence map $\boldsymbol{M}_w$ generated by forward warping to judge whether the prediction of the muti-view branch is reliable. We use the camera pose and the multi-view depth $\boldsymbol{d}_m$ as the input to warping the source images to the reference camera space to obtain $\left\{\boldsymbol{I}_i^{'}\right\}_{i=1}^{n-1}$, and concat with $\boldsymbol{I}_0$. No matter the multi-view depth is inaccurate, the pose is noisy, or in the dynamic object area, the warping source images will be dissimilar with
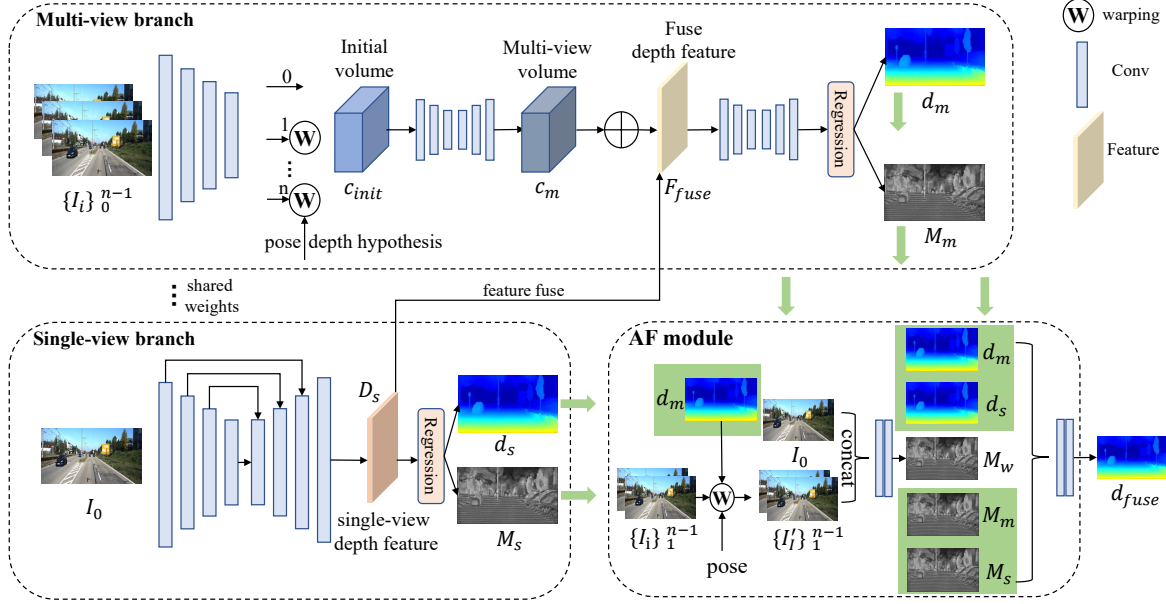
Figure 2. Overview of the AFNet, which consists of three parts: single-view branch, multi-view branch, and the adaptive fusion (AF) module. Two branches share the feature extraction network and have their own prediction and confidence map, i.e. $\boldsymbol{d}_s$, $\boldsymbol{M}_s$, $\boldsymbol{d}_m$ and $\boldsymbol{M}_m$, and then fused by the AF module to obtain the final accurate and robust prediction $\boldsymbol{d}_{fuse}$. The green background in AF module represents the outputs of the single-view branch and multi-view branch.

the corresponding pixels in the reference image. So we can build a warping confidence map $\boldsymbol{M}_w \in \mathcal{R}^{H/4 \times W/4}$ in this way as follows:

$$\boldsymbol{M}_w = \text{Conv}\left\{\text{Concat}\left\{\boldsymbol{I}_0, \boldsymbol{I}_1', \boldsymbol{I}_2' \cdots \boldsymbol{I}_{n-1}'\right\}\right\}. \quad (3)$$

$\boldsymbol{M}_s$ and $\boldsymbol{M}_m$ in single-view branch and multi-view branch reflect the overall matching ambiguity, while $\boldsymbol{M}_w$ reflects the subpixel accuracy. Thus, we take these three confidence maps as guidance for the fusion of single-view depth $\boldsymbol{d}_s$ and multi-view depth $\boldsymbol{d}_m$, and obtain the final fusion depth $\boldsymbol{d}_{fuse} \in \mathcal{R}^{H/4 \times W/4}$ by two 2D convolution layers.

## 3.4. Loss Function

The loss function during AFNet training is mainly composed of two parts, i.e. depth loss and confidence loss. Depth loss uses a simple L1 loss:

$$L_d = \|\boldsymbol{d}_s - \boldsymbol{d}_{gt}\|_1 + \|\boldsymbol{d}_m - \boldsymbol{d}_{gt}\|_1 + \|\boldsymbol{d}_{fuse} - \boldsymbol{d}_{gt}\|_1. \quad (4)$$

For confidence loss, to prevent outliers from interfering with training, we first calculate the valid mask as follows:

$$\begin{aligned}\boldsymbol{\Omega}_s &= |\boldsymbol{d}_s - \boldsymbol{d}_{gt}| < \boldsymbol{d}_{gt}. \\ \boldsymbol{\Omega}_m &= |\boldsymbol{d}_m - \boldsymbol{d}_{gt}| < \boldsymbol{d}_{gt}.\end{aligned} \quad (5)$$

The final confidence loss is calculated as follows:

$$L_c = \frac{1}{N_s} \sum_{p \in \boldsymbol{\Omega}_s} |\boldsymbol{M}_s(p) - (1 - |\boldsymbol{d}_s(p) - \boldsymbol{d}_{gt}(p)|/\boldsymbol{d}_{gt}(p))|$$

$$+ \frac{1}{N_m} \sum_{q \in \boldsymbol{\Omega}_m} |\boldsymbol{M}_m(q) - (1 - |\boldsymbol{d}_m(q) - \boldsymbol{d}_{gt}(q)|/\boldsymbol{d}_{gt}(q))|$$

$$(6)$$

$N_s$ and $N_m$ respectively represent the total number of valid points in $\boldsymbol{\Omega}_s$ and $\boldsymbol{\Omega}_m$. The total loss is the sum of the above two losses $L_d$ and $L_c$.

## 4. Experiment

### 4.1. Datasets and Evaluation Metrics

**DDAD** (Dense Depth for Automated Drivin) [12] is a new autonomous driving benchmark for dense depth estimation in challenging and diverse urban conditions. It is captured with 6 synchronized cameras and contains accurate ground-truth depth (across a full 360-degree field of view) generated from high-density LiDARs. It has 12650 training samples and 3950 validation samples in a single camera view, of which the resolution is 1936×1216. The whole data from 6 cameras are used in training and testing.

**KITTI** [11] is a dataset that provides stereo images and corresponding 3D laser scans of outdoor scenes captured on a moving vehicle. The resolution is around 1241×376. We train and test our method on KITTI Eigen split [8]. Metrics defined in Table 1 are used for evaluation.
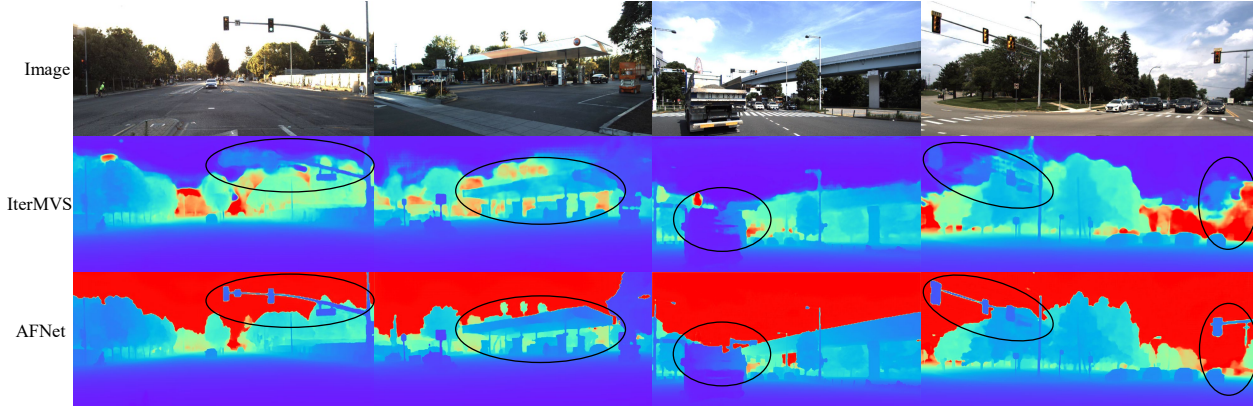
Figure 3. Qualitative results on DDAD [12] test set. Black ellipses highlight obvious improvements achieved by our method.

| AbsRel | SqRel | RMSE |
|---|---|---|
| $\frac{1}{N}\sum_i \frac{(d_i-d_i^*)^2}{d_i^*}$ | $\frac{1}{N}\sum_i \frac{(d_i-d_i^*)^2}{d_i^*}$ | $\sqrt{\frac{1}{N}\sum_i (d_i-d_i^*)^2}$ |

Table 1. Quantitative metrics for depth estimation. $d_i$ is the predicted depth; $d_i^*$ is the ground truth depth; $N$ corresponds to all pixels with the ground-truth label.

## 4.2. Implementation Details

We implement our methods with PyTorch [27] and perform experiments using NVIDIA RTX 3090 GPUs. We use AdamW optimizer [24] and schedule the learning rate using one-cycle policy [31] with $lr_{max} = 1.0\times 10^{-4}$. We trained 30 epochs on DDAD [12] and 40 epochs on KITTI [8]. During training, our system consumes consecutive 3 frames as input, i.e, $n$=3. For the details of multi-view setup, the depth hypothesis number of planes is 128, the weight of the images from different views is the same.

## 4.3. Main Results

To demonstrate the outstanding performance of our method, we evaluate AFNet on DDAD [12] and KITTI [8].

**DDAD [12].** Since most of classical multi-view and single-view methods are not trained and tested on DDAD, We employ the same training scheme to all methods (see Section 4.2 for details). Note that all methods have converged well. In the testing, we evaluate all ring cameras instead of the front-view camera. Quantitative comparisons are reported in Table 11, our AFNet achieves the state-of-the-art (SOTA) performance on DDAD. Compared with current SOTA methods [34] and [1], our AFNet can achieve over 15% improvement on AbsRel error. Qualitative comparisons are shown in Figure 3. The proposed AFNet achieves better results both on dynamic objects and static objects.

**KITTI [11].** KITTI Eigen split [8] is an important benchmark for single-view and multi-view depth estimation. We compare with the state-of-the-art methods on it and show results in Table 11. Our AFNet achieves 0.039 AbsRel, outperforming recent methods by a large margin.

It is worth mentioning that the AbsRel error of AFNet is reduced by 27.8% compared with the SOTA method [1].

## 4.4. Ablation Study

To demonstrate the effectiveness of each component of AFNet, we ablation on DDAD. Following [43], we design a two-branch network and employ the ConvNeXt-T [22] as the backbone. In this section, the following variants are discussed to verify the effectiveness of the proposed AFNet:

- **Base**: Baseline is a multi-view depth estimation model [43], which has the same backbone as ours.
- **Results Fusion**: We add a decoder to the baseline model (Base) for the single-view branch to get single-view prediction and use Ground Truth for supervision, and $d_{fuse}$ is obtained by directly fusing the results of the two branches through two convolutional layers, denoted as Base-RF.
- **Feature Fusion**: We integrate the depth features of the single-view branch into the multi-view branch for complementing semantic and depth cues, denoted as Base-FF.
- **Adaptive Fusion**: Our proposed AF module adaptively selects the most accurate depth between the two branches, demoted as Base-AF.

**Ablation on results fusion.** The comparison results are shown in Table 3. Compared with 'Base', adding a decoder to the single-view branch and supervising the output can extract more robust features for the epipolar attention, leading to the SqRel error of the multi-view prediction ('Base (Multi)' v.s. 'Base-RF (Multi)') reduced by 6.3%. However, a naive fusion of the single-view and the multi-view branch will drop the final performance, i.e. the accuracy of 'Base-RF (Fuse)' is lower than 'Base-RF (Multi)'.

**Effectiveness of the feature fusion.** According to the previous analysis, single-view and multi-view extract feature through different modes, which can be complementary. Therefore, by introducing the single view depth feature into the multi-view branch, 'Base-FF (Multi)' reduces the SqRel error of multi-view branch prediction by 10.7% compared

| Type | Model | DDAD [12] | | | KITTI [11] | | |
|------|-------|-----------|---|---|------------|---|---|
| | | AbsRel↓ | SqRel↓ | RMSE↓ | AbsRel↓ | SqRel↓ | RMSE↓ |
| Single View | Monodepth2 [12] | 0.194* | 3.52* | 13.32* | 0.106 | 0.806 | 4.630 |
| | FeatDepth [30] | 0.189* | 3.21* | 12.45* | 0.099 | 0.697 | 4.427 |
| | DORN [10] | - | - | - | 0.088 | 0.806 | 3.138 |
| | BTS [19] | 0.169* | 2.81* | 11.85* | 0.059 | 0.245 | 2.756 |
| | AdaBins [2] | 0.164* | 2.66* | 11.08* | 0.058 | 0.190 | 2.360 |
| | Metric3D [50] | 0.183* | 2.92* | 12.15* | 0.053 | 0.174 | 2.243 |
| Multi View | PMNet [33] | 0.141* | 2.23* | 10.56* | - | - | - |
| | Deepv2d [32] | - | - | - | 0.091 | 0.582 | 3.644 |
| | CasMVS [13] | 0.129* | 2.01* | 9.87* | 0.066* | 0.228* | 2.567* |
| | MVSNet [44] | 0.109* | 1.62* | 8.21* | - | - | - |
| | IterMVS [34] | 0.104* | 1.59* | 7.95* | 0.057* | 0.178* | 2.234* |
| | MVS2D [43] | 0.132* | 2.05* | 9.82* | 0.058* | 0.176* | 2.277* |
| | SC-GAN [36] | - | - | - | 0.063 | 0.178 | 2.129 |
| | MaGNet [1] | 0.112* | 1.74* | 9.23* | 0.054 | 0.162 | 2.158 |
| | AFNet | **0.088** | **1.41** | **7.23** | **0.039** | **0.121** | **1.743** |

Table 2. Quantitative evaluation on DDAD [12] and KITTI [11]. Note that the ∗ marks the result reproduced by us using their open-source code, other reported numbers are from the corresponding original papers.

| Model | RF | FF | Single | Multi | Fuse |
|-------|-----|-----|--------|-------|------|
| Base | | | - | 1.90 | - |
| Base-RF | Conv | | 2.68 | 1.78 | 1.99 |
| Base-FF | Conv | ✓ | **2.66** | **1.59** | 1.79 |
| Base-AF | Adaptive | ✓ | **2.66** | **1.59** | **1.55** |

Table 3. Results of ablation experiments for each strategy in our method on DDAD [12]. Single represents the result of single-view branch prediction, Multi represents the result of multi-view branch prediction, Fuse represents the fusion result $d_{fuse}$. The reported numbers are SqRel error.

| Model | Single | Multi | Fuse |
|-------|--------|-------|------|
| BASE | 2.66 | 1.59 | 1.55 |
| BASE-PS | 2.62 | 1.53 | 1.49 |
| AFNet (BASE-PS-cost) | 2.62 | **1.44** | **1.41** |

Table 4. Ablation results on feature extraction network parameter sharing and methods for extracting matching information. The reported numbers are SqRel error.

with 'Base-RF (Multi)'.

**Effectiveness of the Adaptive Fusion.** As shown in Table 3, the accuracy of the fusion result 'Base-FF (Fuse)' is lower than 'Base-FF (Multi)' since the previous method of direct convolution fusion was too crude. Thus, we propose an adaptive fusion (AF) module to replace this naive fusion way. Comparing 'Base-AF (Fuse)' with 'Base-FF (Fuse)', using the AF module for results fusion has a 13.4% performance improvement. In addition, the fusion result 'Base-AF (Fuse)' reduces the SqRel error by 2.5% compared with multi-view branch prediction 'Base-AF (Multi)'.

**Ablation on other strategies.** Compared with Base-AF (denoted as BASE for ease of description in this section), our AFNet mainly consists of the following two adjustments: 1) parameter sharing of feature extraction network of single-view and multi-view branch, denoted as BASE-PS; 2) replace Epipolar Attention Module in [43] with cost volume illustrated in Section 3.2 to extract matching information, denoted as BASE-PS-cost. As shown in Table 4, compared with BASE, our BASE-PS-cost has a 9.0% per-

formance improvement for the final result, which is also our final network model, denoted as AFNet.

## 4.5. Discussions

In this section, we discuss the robustness of AFNet in dynamic regions, on zero-shot datasets, and under noise poses.

### 4.5.1 Performance in Dynamic Object Region

By adaptively fusing monocular depth, our AF module can alleviate the problem that multi-view methods cannot handle dynamic objects. This is because the dynamic object region does not satisfy the projection relation, warping confidence map $M_w$ in AF module has certain ability to recognize the dynamic object region, so it can adaptively fuse the single-view result in this region for better prediction. The method of obtaining the mask of the dynamic object region can be found in the supplementary material. As shown in Table 5, our fusion result 'AFNet($d_{fuse}$)' has a 21.0% improvement on SqRel error compared with multi-view prediction 'AFNet($d_m$)'. We also compare with the SOTA single-view and multi-view fusion method [1], [1] improve depth accuracy by fusing single-view depth probability with

multi-view geometry, but this kind of multi-view geometry in dynamic object region is inaccurate which could bring into wrong guidance, and it can be seen that our system performs better in the dynamic object region.

| Model | AbsRel↓ | SqRel↓ | RMSE↓ |
|---|---|---|---|
| MVS2D [43] | 0.163 | 2.362 | 7.325 |
| MaGNet [1] | 0.169 | 2.500 | 7.783 |
| AFNet($d_m$) | 0.158 | 1.879 | 6.024 |
| AFNet($d_{fuse}$) | **0.145** | **1.484** | **5.675** |

Table 5. Performance comparison in dynamic object region on DDAD. AFNet($d_m$) denotes the results of multi-view branch. AFNet($d_{fuse}$) denotes the fusion result.

| Model | AbsRel↓ | SqRel↓ | RMSE↓ |
|---|---|---|---|
| IterMVS[34] | 0.123 | 0.056 | 0.332 |
| MVS2D [43] | 0.098 | 0.044 | 0.276 |
| MaGNet [1] | 0.112 | 0.051 | 0.314 |
| AFNet | **0.091** | **0.039** | **0.253** |

Table 6. Zero-shot performance on ScanNet [6]. All the models are trained on the DDAD dataset and tested on ScanNet. The proposed adaptive fusion network shows a better cross dataset generalization ability.

### 4.5.2 Generalization

The integration of single-view feature into multi-view branch for complementary and the use of adaptive fusion module to select more accurate depth operations are conducive to network robustness and generalization. To observe the generalization of our network, we evaluated AFNet through cross-dataset testing, i.e. testing our model trained on DDAD directly on ScanNet, as shown in Table 6, it shows our model has better performance than the current SOTA methods on ScanNet [6].

### 4.5.3 Robustness under Noise Poses

Noise is inevitable in various SLAM methods to retrieve pose, but whether the pose is accurate or not greatly affects the accuracy of multi-view depth estimation. Therefore, in practical applications, the robustness of the depth estimation network under noisy poses is critical. We provide two kinds of noise to test the robustness of the proposed AFNet, including different levels of synthetic noise and real-world noise generated by SLAM systems. To reflect the anti-noise ability more comprehensively, we added noise to the input pose in the training process of all the methods in Table 7 and Table 10 to enable the networks to adapt to various modes of accurate poses and noise poses simultaneously.

**Synthetic noise.** We propose a new robustness testing benchmark to explore the effectiveness of the multi-view methods given noisy poses. One model is required to be evaluated under different noise levels of the pose, which is

| Method | $\delta = 0$ | $\delta = 0.01$ | $\delta = 0.025$ | $\delta = 0.05$ | ID | R-Rel |
|---|---|---|---|---|---|---|
| PMNet [33] | 0.144 | 0.235 | 0.354 | 0.382 | 0.246 | 0.359 |
| CasMVS [13] | 0.131 | 0.165 | 0.195 | 0.215 | 0.181 | 0.206 |
| MVSNet [44] | 0.112 | 0.160 | 0.189 | 0.208 | 0.177 | 0.202 |
| MVS2D [43] | 0.133 | 0.159 | 0.179 | 0.184 | 0.173 | 0.184 |
| IterMVS [34] | 0.107 | 0.154 | 0.178 | 0.187 | 0.175 | 0.189 |
| MaGNet [1] | 0.115 | 0.162 | 0.185 | 0.191 | 0.183 | 0.195 |
| AFNet($d_m$) | 0.095 | 0.137 | 0.163 | 0.175 | 0.170 | 0.178 |
| AFNet | **0.092** | **0.125** | **0.155** | **0.164** | **0.165** | **0.168** |

Table 7. Performance comparison of AFNet and some state-of-the-art networks under noisy poses on DDAD [12]. The reported numbers are AbsRel error at different settings, $\delta$ represents the intensity of the noise which increases from 0 gradually. "ID" means that we set the input relative pose to the identity pose, and the input source images are the same as the reference image. R-rel is the proposed robustness-aware relative error. AFNet($d_m$) denotes the results of multi-view branch $d_m$, i.e, without AF module.

| Sequence | | 04 | | | 05 | |
|---|---|---|---|---|---|---|
| Pose | GT | ORB2 (ATE=0.15) | ORB1 (ATE=0.98) | GT | ORB2 (ATE=0.4) | ORB1 (ATE=5.6) |
| MoRec[35] | 0.075 | 0.088 | 0.345 | 0.068 | 0.085 | 0.381 |
| IterMVS[34] | 0.068 | 0.078 | 0.182 | 0.073 | 0.080 | 0.169 |
| MaGNet[1] | 0.066 | 0.071 | 0.077 | 0.069 | 0.077 | 0.082 |
| AFNet | **0.059** | **0.061** | **0.064** | **0.063** | **0.066** | **0.070** |

Table 8. Performance comparison under Ground Truth poses and SLAM system poses (ORB1 and ORB2 represents the monocular version and stereo version of ORBSLAM2[25] respectively) on KITTI [11] Odometry dataset. ATE represents the absolute trajectory error between the estimated poses and the Ground Truth poses. The reported numbers are AbsRel error.

obtained by converting the relative poses into Euler angles and translations and then multiplying the Euler angles and translations by a disturbance coefficient. Different coefficient corresponds to different noise level, respectively: 1) accurate pose, denoted as $\delta = 0$; 2) the disturbance coefficient of 50% test samples is $(1 + 0.01)$, and the remaining 50% samples is $(1 - 0.01)$, denoted as $\delta = 0.01$; 3) the disturbance coefficient of 50% samples is $(1 + 0.025)$, and the remaining 50% samples is $(1 - 0.025)$, denoted as $\delta = 0.025$; 4) the disturbance coefficient of 50% test samples is $(1 + 0.05)$, and the remaining 50% samples is $(1 - 0.05)$, denoted as $\delta = 0.05$; 5) set input relative pose as identity pose, and the input source images are the same as the reference image, denoted as ID. The mean relative error $(\mu)$ and the standard deviation $(\sigma)$ are calculated based on all noise pose levels. Finally, a robustness-aware relative error (R-Rel) is defined as $\mu + \sigma$. The lower the better.

As shown in Table 7, our AFNet outperforms other classical multi-view methods for the ability to resist input pose noise. When pose has larger noise, the accuracy of classical
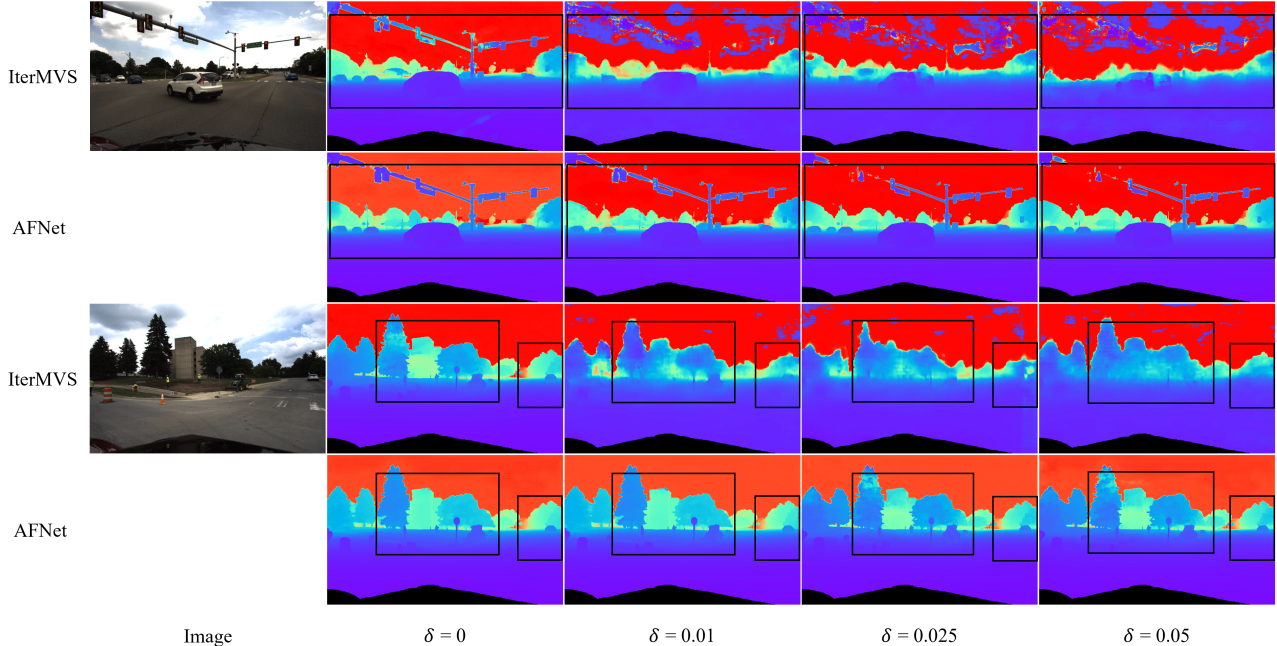
Figure 4. Visualization comparison results on DDAD [12]. The black boxes show the robustness of our AFNet. With the gradual increases of pose noise, the accuracy of [34] which is mainly based on multi-view matching decreased dramatically, while we remained stable.

cost volume-based multi-view methods [13, 33, 44] is even much lower than that of single-view methods, such as Ada-Bins [2] and BTS [19] whose AbsRel errors are 0.164 and 0.169 as shown in Table 11. This is because [13, 33, 44] obtain depth by regression based on the matching information which is false when poses fail. Some current SOTA multi-view methods also integrate single-view information, such as MVS2d [43] and MaGNet [1]. However, since they do not have the structure to autonomously judge when they can rely on the results of the multi-view methods, thus, when the pose is seriously disturbed, the multi-view branch will introduce wrong guidance. We overcome this with the proposed adaptive fusion module, AFNet has the highest precision in all noisy pose settings and always remains above the level of the single-view method. There is also an appreciable improvement in robustness-aware relative error (R-Rel) compared with the SOTA fusion system[1, 43], which is an overall assessment of the robustness of the system.

Whats more, by comparing the result of the multi-view branch (denoted as AFNet($d_m$)) with the result after fusing (denoted as AFNet), it can be seen that the effect of AF module will be more obvious when under noisy poses.

**Real-world noise poses generated by SLAM systems.** To more comprehensively evaluate the robustness of our system under noise poses, we tested the depth accuracy under poses obtained by different SLAM systems on KITTI [11] Odometry dataset. We selected a representative slam system, i.e. ORBSLAM2[25], and used two versions of it to obtain poses respectively, one monocular version (denoted as ORB1) and one stereo version (denoted as ORB2). The monocular version will sometimes crush, and we compare the depth accuracy of the image sequences before crush. The pose noise of the monocular version is larger than that of the stereo version, which can be reflected by the absolute trajectory error. As shown in Table 10, our AFNet has the highest accuracy under different poses, and is also the most robust. The results of the remaining sequences in the KITTI Odometry dataset can be found in the supplementary materials.

## 5. Conclusion

In this paper, we propose a new multi-view and single-view depth fusion network AFNet for alleviating the defects of the existing multi-view methods, which will fail under noisy poses in real-world autonomous driving scenarios. We propose a new robustness evaluation metric and testing benchmark to explore the effectiveness of the multi-view methods under different noise levels. We fuse the single-view and multi-view depth by the proposed adaptive fusion module, improving the accuracy and robustness of the system. AFNet achieves state-of-the-art performance on both KITTI [11] and DDAD [12] datasets with accurate pose while also outperforms all other classical multi-view methods on robustness testing benchmark under noisy poses.

# References

[1] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Multi-view depth estimation by fusing single-view depth probability with multi-view geometry. In *CVPR*, 2022. 2, 3, 5, 6, 7, 8

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, 2021. 2, 6, 8

[3] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *ICCV*, 2019. 2

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

[5] Junda Cheng, Gangwei Xu, Peng Guo, and Xin Yang. Coatrsnet: Fully exploiting convolution and attention for stereo matching by region separation. *International Journal of Computer Vision*, 132(1):56–73, 2024. 2

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 7

[7] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *CVPR*, pages 10786–10796, 2021. 2

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 2014. 4, 5

[9] José M Fácil, Alejo Concha, Luis Montesano, and Javier Civera. Single-view and multi-view depth fusion. *RA-L*, 2017. 2, 3

[10] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 2, 6

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 2, 4, 5, 6, 7, 8

[12] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 1, 2, 4, 5, 6, 7, 8

[13] 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*, 2020. 2, 6, 7, 8

[14] Zhixiang Hao, Yu Li, Shaodi You, and Feng Lu. Detail preserving depth estimation from a single image using attention guided networks. In *3DV*, 2018. 2

[15] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *WACV*, 2019.

[16] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkilä. Guiding monocular depth estimation using depth-attention volume. In *ECCV*, 2020. 2

[17] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, 2015. 1

[18] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016. 2

[19] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv:1907.10326*, 2019. 2, 6, 8

[20] Rui Li, Dong Gong, Wei Yin, Hao Chen, Yu Zhu, Kaixuan Wang, Xiaozhi Chen, Jinqiu Sun, and Yanning Zhang. Learning to fuse monocular and multi-view cues for multi-frame depth estimation in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21539–21548, 2023. 2

[21] Rui Li, Danna Xue, Shaolin Su, Xiantuo He, Qing Mao, Yu Zhu, Jinqiu Sun, and Yanning Zhang. Learning depth via leveraging semantics: Self-supervised monocular depth estimation with both implicit and explicit semantic guidance. *Pattern Recognition*, 137:109297, 2023. 2

[22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 3, 5, 1

[23] Xiaoxiao Long, Lingjie Liu, Wei Li, Christian Theobalt, and Wenping Wang. Multi-view depth estimation using epipolar spatio-temporal networks. In *CVPR*, 2021. 2

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

[25] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017. 7, 8, 1, 2

[26] Guang-Yu Nie, Ming-Ming Cheng, Yun Liu, Zhengfa Liang, Deng-Ping Fan, Yue Liu, and Yongtian Wang. Multi-level context ultra-aggregation for stereo matching. In *CVPR*, 2019. 2

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[28] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *CVPR*, 2018. 2

[29] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 1

[30] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *ECCV*, 2020. 6

[31] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of residual networks using large learning rates. 2018. 5

[32] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605*, 2018. 6

[33] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *CVPR*, 2021. 6, 7, 8

[34] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Itermvs: Iterative probability estimation for efficient multi-view stereo. In *CVPR*, 2022. 5, 6, 7, 8, 2

[35] Felix Wimbauer, Nan Yang, Lukas Von Stumberg, Niclas Zeller, and Daniel Cremers. Monorec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6112–6122, 2021. 7, 2

[36] Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Song Wang, and Lili Ju. Spatial correspondence with generative adversarial network: Learning depth from monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7494–7504, 2019. 6

[37] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *CVPR*, 2017. 2

[38] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *CVPR*, 2018. 2

[39] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12981–12990, 2022. 2

[40] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023.

[41] Gangwei Xu, Yun Wang, Junda Cheng, Jinhui Tang, and Xin Yang. Accurate and efficient stereo matching via attention concatenation volume. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[42] Gangwei Xu, Huan Zhou, and Xin Yang. Cgi-stereo: Accurate and real-time stereo matching via context and geometry interaction. *arXiv preprint arXiv:2301.02789*, 2023. 2

[43] Zhenpei Yang, Zhile Ren, Qi Shan, and Qixing Huang. Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions. In *CVPR*, 2022. 3, 5, 6, 7, 8, 1, 2

[44] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 2, 3, 6, 7, 8

[45] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *CVPR*, 2019. 2

[46] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *ICCV*, 2019. 2

[47] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. page 2002.00569, 2020. 2

[48] Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. 2021. 2

[49] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3d scene shape from a single monocular image. 2022. 2

[50] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 6

[51] Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *CVPR*, 2020. 2

[52] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. *arXiv preprint arXiv:2203.01502*, 2022. 2

[53] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *CVPR*, 2015. 2