Depth Refinement for Improved Stereo Reconstruction

Amit Bracha  Noam Rotstein  David Bensaïd  Ron Slossberg
Ron Kimmel
Technion - Israel Institute of Technology
{amit.bracha, snoamr, dben-said, ronslos, ron}@cs.technion.ac.il

Abstract

Depth estimation is a cornerstone of a vast number of applications requiring 3D assessment of the environment, such as robotics, augmented reality, and autonomous driving to name a few. One prominent technique for depth estimation is stereo matching which has several advantages: it is considered more accessible than other depth-sensing technologies, can produce dense depth estimates in real-time, and has benefited greatly from the advances of deep learning in recent years. However, current techniques for depth estimation from stereoscopic images still suffer from a built-in drawback. To reconstruct depth, a stereo matching algorithm first estimates the disparity map between the left and right images before applying a geometric triangulation. A simple analysis reveals that the depth error is quadratically proportional to the object’s distance. Therefore, constant disparity errors are translated to large depth errors for objects far from the camera. To mitigate this quadratic relation, we propose a simple but effective method that uses a refinement network for depth estimation. We show analytical and empirical results suggesting that the proposed learning procedure reduces this quadratic relation. We evaluate the proposed refinement procedure on well-known benchmarks and datasets, like Sceneflow and KITTI datasets, and demonstrate significant improvements in the depth accuracy metric.

1. Introduction

Depth estimation is a well-studied task in computer vision with a central role in increasingly important areas such as robot navigation, augmented reality, and autonomous driving. Depth sensing can be achieved either by actively illuminating the environment and measuring various aspects of the reflected light or by passively sensing the ambient light from one or more viewpoints. The former method is considered more accurate and complex laser-based systems, such as LiDAR, are the gold standard in autonomous driving. However, the high cost of modern LiDAR systems as well as their limitations in real-time configurations restrain their wide application. Stereoscopic depth estimation relies on estimating the relative displacement of an object between two views. This displacement is termed disparity and is inversely proportional to the object’s distance. For this purpose, a dense correspondence between pairs of images is performed, followed by a disparity refinement phase.

In essence, stereo reconstruction trades hardware for software complexity, as the correspondence task is particularly challenging within textureless, poorly illuminated, and occluded regions. However, human high depth estimation capabilities with a natural stereo configuration suggest that a simple two-camera setup is sufficient for accurate spatial reasoning. In recent years, stereo correspondence methods have greatly benefited from modern deep learning techniques and their quality is constantly improving. Current state-of-the-art methods achieve a 3-pixel outlier rate of less than 2% on standard benchmarks such as Middlebury, ETH3D, or Kitti [17, 22, 23].

Despite the rapid improvements in the quality of stereo matching, range-from-stereo pipelines based on disparity estimation suffer from an inherent flaw. As detailed in Section 3.1, the depth error is quadratically proportional to the depth of the observed object; for a constant error in disparity, the perceived error in depth is much larger for distant objects. The previous quadratic relation is a major drawback of the stereo methods, making them unsuitable for applications such as autonomous driving where accurate distance estimation of far objects is essential. In PseudoLidar++ [29], You et al. have pointed out this fundamental problem and addressed it by directly optimizing a model based on PSMNet [2] on depth instead of disparity. However, this approach also suffers from a significant drawback. Modern range-from-stereo models, including PSMNet, are hinging on stereo matching algorithms where the disparity is the natural metric to optimize. The transition to a direct depth-based approach negatively affects the performances of the model. We propose a novel two-step approach to resolve the dichotomy between the essential metric for applications, depth, and the natural metric for the...
optimization of stereo matching algorithms, disparity. The two main steps comprising our methodology are the following: Firstly, state-of-the-art stereo range estimation models are trained on disparity, as usual, to maximize the performance of the stereo matching algorithm. Secondly, a refinement model is depth-wisely-optimised using the computed disparity map and the left and right images. A comparison between unrefined and refined depth estimates is depicted in Fig. 1. We observe that the refined depth depicts finer detail and sharper object boundaries.

Our main contributions are the following:

• We propose a novel approach to improve the accuracy of stereoscopic depth estimation. This procedure can easily be implemented on top of any existing depth-from-stereo model.

• We demonstrate a significant improvement for several state-of-the-art methods by recasting the classical disparity refinement stage into a depth refinement problem.

• We present an ablation study demonstrating a clear advantage of our two-step refinement approach over frameworks trained end-to-end with a depth-based loss.

• We successfully leverage analytical relations between disparity and depth errors to efficiently correct discrepancies in-depth estimation.

2. Related work

2.1. Deep stereo matching

The problem of finding the stereo correspondence between a pair of rectified images is well studied in the field of computer vision. In recent years, this task has benefited from the introduction of modern neural networks specifically adapted for the stereo correspondence task and is now largely dominated by such data-driven rather than classical model-driven methods.

The first to tackle stereo correspondence in the era of DNN were [30] who proposed a patch matching scheme based on a triplet hinge loss fed by matching and non-matching examples. The learned patch descriptors provide the initial disparity estimates which are further refined using classical methods. This approach, however, is somewhat limited due to its local nature focusing on the matching scheme while disregarding the global geometric structure. This was remedied by later efforts [5], who provide end-to-end frameworks for matching as well as disparity refinement. In a later publication by [9], termed GC-Net, 3D convolutions kernels were applied to the feature correspondence space, facilitating a more general framework for end-to-end correspondence and disparity refinement. This new approach provided a significant improvement over previous ones.

In the network proposed by [2] termed PSMnet, the authors further improve the feature extraction with a multi-
scale feature extraction network [13] typically used for object detection. In LEA-Stereo [3], a neural architecture search is performed to further streamline the matching pipeline. Another notable improvement was GA-net [31]. There, an architecture inspired by the semi-global-matching (SGM) algorithm [7] was incorporated into the deep stereo learning pipeline. This is achieved by introducing semi-global as well as local guided aggregation layers which mimic the classical SGM method with tunable weights.

The work perhaps closest to our line of thought is [29] who argue that depth, rather than disparity, is a more suitable metric for training and evaluation of geometry from stereo pipelines as the desired output from the model is depth estimation and not disparity. To this end, they suggest altering the model output to calculate the depth from the obtained disparity and adapt the cost aggregation layer accordingly. However, their adjustment is not based on an axiomatic method, and it does not have any geometric foundations as the stereo matching algorithm. Here, we propose to refine the output obtained from the stereo matching algorithm under metrics that evaluate the estimated depth rather than disparity. This subtle yet important difference allows us to retain geometric fidelity and at the same time significantly improve the more meaningful depth estimation metrics.

2.2. Refinement for stereo matching

In several recent papers, a refinement component has been introduced in addition to an existing pipeline to improve the disparity estimation quality. For example, in [18], the authors utilize a cascade residual learning model as a refinement stage applied to DispNet [16]. The MSMD-Net [24] proposed a multi-scale approach where left-right pairs were processed by a multi-resolution pyramid network and then refined, reducing inference complexity and subsequently run-time. In a recent paper [14], the RAFT-stereo architecture was proposed. This model consists of a stereo estimation network coupled with a GRU-based multi-scaling refinement stage. All models mentioned in this section emphasize the refinement step of the matching process; however, they all address the problem as a disparity refinement process. Hence, these methods still suffer from the quadratic relation between the distance from the observed object and the depth error.

2.3. Monocular depth estimation

Monocular depth estimation methods attempt to estimate depth from a single image. Although this problem is ill-posed, by obtaining a good prior knowledge of the surrounding scenes and objects it is possible to obtain reasonable results. Such methods were previously proposed by [1, 6, 10, 20, 26, 27] to name a few. Here, we propose to combine elements from both monocular and stereo-reconstruction to derive our depth refinement scheme.

In [6], the authors incorporate a multi-scale network architecture coupled with an ordinal training loss. This loss effectively reduces the errors stemming from large depth values during training. In addition, the authors of [10] utilized an encoder-decoder architecture with an additional local layer as a refinement stage. Furthermore, in [10] the authors propose an end-to-end object-detection from a stereo model similar to [29] supplemented by an additional refinement stage improving previous results obtained by similar approaches.

In contrast to monocular depth estimation schemes which solve an ill-posed problem, we retain geometric information captured by the stereo-matching pipeline and utilize it to better estimate the depth from the viewpoint of one or both of the stereo cameras. In addition, our goal is to refine depth estimates of existing models, hence, residual learning is more appropriate in our case than previously discussed monocular reconstruction losses.

3. Depth Refinement for Stereo Reconstruction

Typical pipelines for depth reconstruction from stereoscopic image pairs generally rely on finding correspondences between the left and right images and are optimized according to a disparity criterion. However, such optimization is error-prone for more distant objects and leads to a distorted reconstruction of the geometry, see Section 3.1. Here, we propose to mitigate this bias by introducing a refinement network trained to minimize a depth-based loss.

3.1. Depth and disparity

The relation between depth and disparity is given by a simple geometric triangulation [4],

$$Z_a = \frac{b \cdot f_x}{d_a} \propto \frac{1}{d_a^2}. \quad (1)$$

Where $Z_a$ is the depth estimate, $b$ is the baseline of the camera, $f_x$ is the horizontal focal length, and $d_a$ is the disparity approximation. Denote $d_{gt}$ the disparity ground truth, $\epsilon_d$ the error in disparity and $\epsilon_Z$ the error in depth. Assuming $d_{gt} \gg \epsilon_d$, $\epsilon_Z$ can be approximated by,

$$\epsilon_Z = Z_{gt} - Z_a = \frac{b f_x}{d_{gt}} \frac{d_{gt} + \epsilon_d}{d_{gt}} \approx \frac{b f_x \epsilon_d}{d_{gt}^2} \approx \frac{\epsilon_d}{b f_x} Z_{gt}^2. \quad (2)$$

The above equation reveals that the quadratic relation between the error in depth and the distance of the observed object mentioned in Sec. 1. Fortunately, this relation only holds as long as the range is determined by the disparity. In
PseudoLidar++ [29]. You et al. promote training directly for depth estimation while skipping the initial correspondence stage. On one hand, by abandoning the classical stereo matching paradigm, we can avoid the problem exposed in Eq. 2. On the other hand, reconstructing depth from disparity is a natural approach with geometric foundations.

Here, we propose a new method for processing depth estimation while also preserving the geometric guarantees stemming from correspondence and triangulation. In contrast to the end-to-end approach proposed by [29], we propose a two-step method that splits the overall pipeline into a disparity estimation stage followed by a dedicated depth refinement model. The disparity approximation stage is obtained by adopting several successful stereo matching pipelines such as the ones discussed in the previous section. The initial disparity approximation is then converted into a depth estimation that is further refined via a dedicated DNN model, (see Sec. 3.2).

### 3.2. Multiplicative residual refinement

Motivated by modern denoising procedures [32], we consider a residual architecture. Denote Z the final depth estimate, \( Z_e \) the estimate before refinement, \( R \) the output of the refinement process, and \( \oplus \) the pixel-wise addition. We have,

\[
Z = Z_e \oplus R. \tag{3}
\]

According to Eq. 2, the refinement \( R \) should be proportional to \( Z^2 \). However, approximating a polynomial function of the input values by a neural network is not trivial [8, 12, 28]. Theoretically, representing a polynomial function of the input requires deep architectures. The network’s complexity is bounded from below by the desired error and the degree of the polynomial [12, 28]. Beyond simply increasing the number of parameters, the introduction of domain-specific prior knowledge typically has a positive influence on model performance. A successful example of such utilization of prior knowledge in 3D Computer Vision is MomeNet [8], a neural network dedicated to processing point clouds based on the previously introduced PointNet [19]. MomeNet demonstrates improved performance over PointNet by exploiting the importance of moments in shape analysis. This is achieved by providing the model with the second-order moments of the points in addition to their coordinates. Along the same line of thought, we leverage the analytical knowledge of the relation between the depth error and the distance to the observed object (Eq. 2). This is achieved by multiplying the model output by the depth estimate \( Z_e \) instead of directly learning \( R \).

\[
R = f(Z_a, I_l, I_r; \theta) \odot Z_a, \tag{4}
\]

Where \( \odot \) is the Hadamard product and \( f(Z_a, I_l, I_r; \theta) \) is the refinement network with weights \( \theta \). Multiplying the network’s output by the depth estimates allows the network to easily converge towards a \( R \) proportional to \( Z^2 \). We quantitatively show the benefit of the former operation in Sec. 4.3.

### 3.3. Depth Refinement (DR) Network

The refinement model proposed is a CNN based on UNet [21] that learns depth priors from images to ameliorate the existing depth map. The inputs to the DR network are the depth estimate computed by pretrained stereo matching pipeline, the left image, and the warping [18] of the right image with respect to the disparity approximation. That is,

\[
R = f(Z_a, I_l, I_r^l; \theta) \odot Z_a, \tag{5}
\]

and the output of our method is,

\[
Z = Z_a \oplus (f(Z_a, I_l, I_r^l; \theta) \odot Z_a). \tag{6}
\]

Where \( I_r^l \) is the right image warped into the left image plane. As previously discussed, we utilize a depth based loss during the training process.

\[
\mathcal{L} = \|Z_{gt} - Z_a \oplus f(Z_a, I_l, I_r^l; \theta) \odot Z_a\|_1, \tag{7}
\]

For numerical stability, the DR network is trained with a slightly modified version of the loss presented in Eq. 7.

\[
\mathcal{L} = \| \frac{Z_{gt} - Z_a}{Z_a} - f(Z_a, I_l, I_r^l; \theta) \|_1. \tag{8}
\]

This modified loss produces range invariant gradients. To improve the numerical stability during training, we ignore objects at a distance greater than 100m. This assumption is often applied in standard benchmarks such as KITTI to assess results.

### 4. Experiments

In this section, we evaluate the proposed refinement procedure by applying it to three recent stereo matching pipelines, the CFSNet, the LEA-Stereo, and the RAFT-Stereo. Experiments are performed on two classical benchmarks: the synthetic dataset SceneFlow [15] and KITTI 2015 [17]. We demonstrate that the proposed refinement methodology substantially improves the depth estimate of each model, see tables 2 and 1. In addition, we observe an improvement in the performance of the studied stereo methodologies in several common disparity-related metrics as well. We performed an ablation study in which we examine a similar architecture and training procedure but vary the loss function between an additive or multiplicative residual loss, see Table 4. Finally, we compare our approach to PseudoLidar++ [29], see Table 3, which proposes a line of thought close to ours and also focus on depth.
### 4.1. Datasets

**SceneFlow.** As previously discussed, we evaluate the proposed refinement pipeline on two well known datasets, SceneFlow and KITTI 2015. SceneFlow is a synthetic dataset presented in [15], containing roughly 35K images for training and 4.3K images as a test-set. The main benefit of such synthetic datasets is their dense and precise ground truth and their large size making them suitable for the training of DNNs. On the other hand, these types of datasets present unique challenges when transferring trained models to real domains due to many factors such as sensor noise, lens deformations, scene composition, and realism. Example scenes from the dataset as well as their estimated and refined depth outputs are depicted in Fig. 1. Nevertheless, we utilize this dataset to achieve several goals. (i) We demonstrate our ablation experiment using this data. (ii) We examine our model performance without the confounding factors of ground truth noise, misalignment, miscalibration, etc. (iii) We utilize its vast size to perform transfer learning to the smaller real KITTI dataset.

**KITTI 2015.** To evaluate our model on real-world data, we turn to the KITTI 2015 dataset [17]. This widely used benchmark includes a relatively small training-set containing 200 images paired with sparse ground-truth depth maps as a similarly sized test-set with held-out ground-truth. The KITTI dataset was obtained by mounting cameras as well as a Velodyne LIDAR onto a car and driving it on roads and city streets while recording. The LIDAR was calibrated and synchronized with the stereo cameras thus producing the corresponding depth to each dataset image. The depth sparsity stems from the scanning laser technology employed by the Velodyne scanner. The sparse nature of the ground-truth images creates some difficulty when evaluating highly accurate stereo correspondence pipelines as the ground-truth is typically inaccurate near object boundaries which are often the differentiating factor between the various tested methods. Therefore, methods that perform well at object boundaries and may receive high accuracy scores on benchmarks containing dense depth maps might under-perform on the KITTI benchmark. We note that the proposed refinement process based on depth and image data is especially successful at object boundaries and therefore suffers from the sparse nature of the KITTI ground-truth depth during evaluation.

### 4.2. Depth Refinement Evaluation

We perform an extensive evaluation of the proposed depth refinement methodology using the aforementioned datasets. First, we evaluate our approach on the synthetic SceneFlow dataset. We conduct our evaluation on three recent stereo correspondence pipelines, namely: LEAStere [3], CFNet [25], and RAFT-Stereo [14]. To perform a fair evaluation, we load the author-provided weights for each model trained specifically for each dataset, and use the same data preprocessing procedures appropriate for each method.

We test the refined depth output under both depth as well as disparity-related metrics. Since our refinement scheme is trained according to a depth $L_1$ loss, it is, therefore, possible that the increase in depth accuracy is achieved at the cost of disparity degradation. To evaluate the disparity, we transform the refined depth estimate back to disparity through

$$d_a = \frac{b \cdot f_x}{Z}, \quad (9)$$

where $d_a$ is the refined disparity estimate. All the results are evaluated only in unmasked areas where the disparity is not greater than the max-disparity hyper-parameter of the model, and the depth is not greater than 100 meters. The obtained refined depth and disparity errors are presented in Tab. 1. We find that our refined depth estimates demonstrate a 15% to 39% improvement in depth accuracy over the baseline.

Furthermore, the refined disparity derived from the estimated depth is improved as well relative to the baseline, even though the refinement process does not aim to directly improve disparity errors. These results demonstrate that while our methodology achieves lower depth error rates as expected, the disparity estimation does not suffer and even manages to benefit from it as well.

To examine our method on the KITTI 2015 dataset, we use the same three deep learning matching pipelines used
Table 2. Evaluation of our model on the KITTI 2015 dataset \[17\]. To evaluate our model on the training-set, we used a 10-fold validation scheme. We examine our model on three different stereo matching algorithms, where LEAStereo \[3\] is the first ranked algorithm on KITTI 2015 among methods with published code. The results demonstrate a performance increase of roughly 7% in depth estimation over the baseline. We observe that the EPE metric, the disparity error, and D1-1px, the percentage of pixels with more than 1px error in disparity, are also improved.

Table 3. We conduct a comparison between two pretrained models: stereo depth network (SDN) \[29\], and PSMNet \[2\], with and without depth refinement (DR). We measure average absolute depth error (Depth error), end-point error (EPE), and percentage of pixels with more than 3 pixels of disparity error (D1). We observe that the proposed refinement improves all tested metrics with relation to the baseline. Note that reported results might differ from published figures due to different masking protocols.

### 4.3. Ablation Study

**Benefits of the two-step pipeline.** To evaluate the benefit of performing a two-step refinement in contrast to end-to-end learning of depth estimation, we compare our results to the single-step SDN model \[29\]. In \[29\], the authors used as a backbone to their network, the PSMNet model \[2\], and replace the disparity-based training loss with a depth estimation loss. We compare the results of our proposed pipeline applied to PSMNet with the reported one-step training reported by \[29\]. We used the pretrained network weights from Sceneflow published by the authors for both the PSMNet as well as SDN. We trained our refinement scheme for 20 epochs on the Sceneflow dataset and present the result in Tab. 3. The results demonstrate that our proposed method outperforms both the PSMNet baseline and the version trained on depth as reported by \[29\]. In addition, we observe that our method with the PSMNet reached a lower error in depth in comparison to the SDN model.

**Naive model.** To refine the depth estimate, the output of our model is multiplied by the depth estimate produced by the stereo matching algorithm. In this part, we wish to examine this multiplicativc residual approach in contrast to a more naive additive approach as discussed in Sec. 3.2. To this end, we test the refinement model with the naive additive residual approach. Instead of multiplying by the depth estimate of the stereo model, we simply add the refinement output to the depth estimate. Denoting the DR network output \(f(Z_a, I_l, I_l^l; θ)\) and the pre-computed depth map \(Z_a\), the multiplicative residual approach refers to the computation of the final depth estimates \(Z = Z_a ⊕ f(Z_a, I_l, I_l^l; θ)\). The naive refinement refers to the results obtained by opting for a straightforward approach where the refinement output of the DR network is directly added to the pre-computed depth map, i.e \(Z = Z_a ⊕ f(Z_a, I_l, I_l^l; θ)\). We reevaluate the for-
mer method with the three stereo correspondence pipelines: CFNet, LEAStereo, and RAFT-Stereo. We substitute only the residual component and retain all the training procedures as before while training on the Sceneflow train-set. We report the results on the Sceneflow test-set comparing the two approaches. As before, each model has been trained for 30 epochs. The results presented in Tab. 4 demonstrate that while both approaches improve depth estimation, the multiplicative approach outperforms the naive additive one. This aligns with our expectation as discussed in Sec. 3.2.

| Methods      | LEAStereo | RAFT-Stereo | CFNet  |
|--------------|-----------|-------------|--------|
| Ours         | 39.7%     | 15.5%       | 22.5%  |
| Naive refinement | 33.9%   | 13.6%       | 18.2%  |

Table 4. The results of the ablation study regarding two styles of residual learning. Indeed, the multiplicative approach consistently outperforms the naive additive approach across all tested correspondence methods. This result supports our expectations regarding quadratic functions and neural networks.

Quadratic error mitigation. Our proposed method for depth refinement of stereo matching algorithms attempts to tackle and reduce the quadratic relation between the depth and the error within the depth estimate. In this experiment, we assess the reduction of the quadratic relation after refinement when compared to the unrefined depth error.

We first obtain the depth estimation resulting from the LEAStereo model evaluated on the Sceneflow test-set. We measure the depth error with respect to the ground-truth with our refinement and without it. To this end, we divided the pixels into bins according to their ground-truth distance with one meter intervals. We then take the median depth error for each bin and plot the medians before and after the refinement process. We present the result with a fitted second-degree polynomial trend line in Fig. 2. The polynomial coefficients demonstrate a substantial reduction in the quadratic component when refining the depth estimate with our proposed method.

![Figure 2. A comparison between the depth estimate of the baseline stereo algorithm and the refined estimate and fit a curve representing the quadratic relation of the error with the distance. The x-axis is the ground-truth distance while the y-axis represents the median depth error in mm. By examining the coefficients of the fitted polynomials, we notice that the quadratic relation between the error and the distance is reduced due to the refinement process.](image)

5. Discussions and Conclusions

In this paper, we demonstrate experimentally that processing the depth rather than the disparity, we can refine depth recovery for state-of-the-art depth from stereo pipelines. In addition, we find that the gained benefit in depth estimation quality is independent of the performance of the tested stereo pipeline and that even highly successful methods are amenable to our refinement approach. Moreover, we demonstrate that, while our approach is geared towards improved depth estimation, the disparity-related metrics do not suffer a degradation in quality. This suggests that the depth domain, where natural geometric relations are preserved, is more tractable for learning via DNN’s.

When attempting to refine disparity values, a DNN ideally models the relation between the disparity and the scene geometry; however, this relation is indirect as several transformations including homography transform, lens distortion, and depth triangulation are necessary mathematical transformations relating between perceived disparity and physical scene geometry. Our ablation study demonstrates that even a simple residual denoising-based refinement model significantly improves depth estimation accuracy. Additionally, it also improves disparity estimate fidelity, thus, validating the powerful notion of processing the model output in a more natural domain. Furthermore, we show that multiplying the refinement model output by the input depth estimate before applying the loss function helps the model to converge in a shorter time, and improves the final estimation quality significantly. This result aligns with our mathematical reasoning regarding the quadratic relation between the depth and the distance to the observed object as well as previous observations regarding the ability of networks to process quadratic functions.

Finally, we assess the relation between distance and depth error before and after our proposed refinement procedure. By doing so, we can demonstrate that the refined output exhibits a weaker quadratic relation. The former points highlight that our framework is more robust to the object distance than previous disparity-based approaches.
Thus, it potentially opens the door to more complex and accurate depth-based refinement methodologies for stereo reconstruction, possibly leading to a linear relation in future work.

References

[1] Shubhra Aich, Jean Marie Uwabeza Vianney, Md Amirul Islam, and Mannat Kaur Bingbing Liu. Bidirectional attention network for monocular depth estimation. In 2021 IEEE International Conference on Robotics and Automation (ICRA), pages 11746–11752. IEEE, 2021. 3

[2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5410–5418, 2018. 1, 2, 6

[3] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. Advances in Neural Information Processing Systems, 33:22158–22169, 2020. 2, 3, 5, 6

[4] Robert Cormack and Robert Fox. The computation of disparity and depth in stereograms. Perception & Psychophysics, 1985. 3

[5] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 2758–2766, 2015. 2

[6] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2002–2011, 2018. 3

[7] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), volume 2, pages 807–814. IEEE, 2005. 3

[8] Mor Joseph-Rivlin, Alon Zvirin, and Ron Kimmel. Moments: Flavor the moments in learning to classify shapes. In Proceedings of the IEEE/ICCV International Conference on Computer Vision (ICCV) Workshops, Oct 2019. 4

[9] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In Proceedings of the IEEE International Conference on Computer Vision, pages 66–75, 2017. 2

[10] 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 preprint arXiv:1907.10326, 2019. 3

[11] Sihang Lee, Janghyeon Lee, Byungiu Kim, Eojindl Yi, and Junmo Kim. Patch-wise attention network for monocular depth estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 1873–1881, 2021. 3

[12] Shiyu Liang and R Srikanth. Why deep neural networks for function approximation? In 5th International Conference on Learning Representations, ICLR 2017, 2019. 4

[13] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 213–2125, 2017. 3

[14] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. arXiv preprint arXiv:2109.07547, 2021. 3, 5

[15] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2016. arXiv:1512.02134. 2, 4, 5

[16] Nikolaus Mayer, Eddy Ilg, Philip Hauser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4040–4048, 2016. 3

[17] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. In ISPRS Workshop on Image Sequence Analysis (ISA), 2015. 1, 4, 5, 6

[18] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengyi Yang, and Qiong Yang. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 887–895, 2017. 3, 4

[19] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 652–660, 2017. 4

[20] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3997–4008, 2021. 3

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015. 4

[22] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In German conference on pattern recognition, pages 31–42. Springer, 2014. 1

[23] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 1
[24] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Msmd-net: Deep stereo matching with multi-scale and multi-dimension cost volume. arXiv preprint arXiv:2006.12797, 2020.

[25] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13906–13915, 2021.

[26] Jean Marie Uwabeza Vianney, Shubhra Aich, and Bing-bing Liu. Refinedmpl: Refined monocular pseudolidar for 3d object detection in autonomous driving. arXiv preprint arXiv:1911.09712, 2019.

[27] Yifang Xu, Chenglei Peng, Ming Li, Yang Li, and Sidan Du. Pyramid feature attention network for monocular depth prediction. In 2021 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE, 2021.

[28] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. Neural Networks, 94:103–114, 2017.

[29] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. In ICLR, 2020.

[30] Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. Journal of Machine Learning Research, 17:1–32, 2016.

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

[32] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE transactions on image processing, 26(7):3142–3155, 2017.