Course Paper for Academic Writing in English

Monocular Depth Estimation Based On Deep Learning: A Review

Author: 李宇潇

Student Number: 2021080907032

Date: 2023-5-22

Score: _____

Abstract

Depth information, referring to the distance of each pixel in a picture with the camera, is important for autonomous driving and robot controlling to perceive the environments and estimate their states. With the rapid development of the deep neural network, depth estimation has been widely studied and achieved promising accuracy in an end-to-end manner. Therefore, we first introduce a famous depth estimation method proposed by Gordard et al[2].and then we will focus on several representative existing depth estimation methods.

Main text

0 Introduction

The most commonly used method to obtain depth information is utilizing lasers, structured light, and other reflections on the object's surface. However, getting dense depth information by these methods usually requires an extremely heavy cost of manpower and computing resources. Therefore, image-based methods like monocular depth estimation have been the mainstream, in which unsupervised methods take the lead. To reduce the use of hard-to-obtain ground truth labels of the dense depth map, Zhou et al[1]. proposed the first monocular depth estimation method based on unsupervised learning with monocular video information. To deal with the non-rigid scene motion, an additional mask was added to predict pixels that violated the rigid camera motion. However, they disabled this term in their latest available code online, achieving superior performance, indicating that they did not explicitly address the issue of object occlusion. For dynamic objects and occlusion, Godard et al.[2] proposed an automatic occlusion method, Monodepth2, which minimized photometric error to reduce the artifacts at the object boundary, and improved the sharpness of the occlusion boundary.

1 monodepth2

Gordard et al.[2] sought to automatically infer a dense depth map from a single input image with the hypothesis that the model could learn depth information with monocular image sequences, similar to how humans learn through navigating and interacting in the real world.

The basic model for their method lay in the geometry constraints between the adjacent frames:

$$p_{n-1} \sim K T_{n \to n-1} D_n(p_n) K^{-1} p_n, \tag{1}$$

where p_n stands for the pixel on image I_n, and p_{n-1} refers to the corresponding pixel of p_n on image $I_{(n-1)}$. K is the camera intrinsics matrix, which is known. $D_n(p_n)$ denotes the depth value at pixel p_n and $T_{n \to n-1}$ represents the camera motion transformation matrix between I_n and I_{n-1}. Hence, with the depth and camera motion estimated by the network, the correspondence between the pixels on different images (I_n and I_{n-1}) is established by the projection function. To put it simply, one can project an image to its adjacent image with a different view with depth and camera motion. If the two elements both serve right, the projected image and the real image should be identical. Therefore, the photometric error between the corresponding pixels is calculated with the l1 loss:

$$L_{pl} = \frac{1}{N} \Sigma_p^N \left| I_n(p) - \hat{I}_n(p) \right|, \tag{2}$$

where pl stand for photometric loss and p indexes over pixel coordinates. $\hat{I}_n(p)$ denotes the reconstructed frame. Then the overall structure similarity can be evaluated by the Structural Similarity Index(SSIM), which takes the brightness, contrast, and structural information of an image into consideration, to quantify the differences between reconstructed and target images, the total loss is given by:

$$L_{total} = \alpha \frac{1 - SSIM(I_n - \hat{I}_n)}{2} + (1 - \alpha) \cdot L_{pl}, \tag{3}$$

where α is a balance factor.

By minimizing the photometric loss function in (3), forcing the projected image to be identical to the real one, the network learns to estimate the actual depth and camera motion in an end-to-end manner efficiently. However, the above method only works under the assumptions of a moving camera and a static scene. When these assumptions break down, for example when there are objects moving in the scene, performance can suffer greatly. Different from the work of Zhou[1], Gordard et al.[2] explicitly address this issue by explicitly computing the mask of pixels to be considered in the loss function, excluding pixels that violate the rigid camera motion and do not change appearance from one frame to the next in the sequence. It is worth noting that the mask of pixels is computed automatically in the forward pass, instead of adding an additional network to predict the mask in a memory-unfriendly way in Zhou's[1] method. The mask μ is set to only include the loss of pixels where the reprojection error of the warped image $I_{t'\to t}$ is lower than that of the original, unwarped source image I', i.e.

$$\mu = [\min_{t'} pe(I_t, I_{t'\to t}) < \min_{t'} pe(I_t, I_{t'})], \tag{4}$$

where [ ] is the Iverson bracket.

Gordard's method utilized the geometry constraints between the adjacent images in a sequence by projecting one to another with the estimated per-pixel depth and pose (camera motion) based on ResNet[4]. It has three major contributions:(i) Gordard et al. proposed a minimum reprojection loss, designed to robustly handle occlusions, (ii) a full-resolution multi-scale sampling method that reduced visual artifacts has been put forward, (iii) they introduced an auto-masking loss to ignore training pixels that violate camera motion assumptions, which further closed the performance gap between monocular and stereo-trained method. Furthermore, they pioneered a novel estimation system that eliminated the need for elusive ground truth depth maps. Instead, they trained models using only monocular video, making it a much more feasible approach. It has since become the standard approach and a solid foundation for monocular depth estimation[3][5]. However, due to the fact that Monodepth2 had to estimate two factors (depth and pose) simultaneously, it suffered from scale ambiguity and inferior robustness to dynamic objects and low-texture regions like walls and billboards. Based on Monodepth2, numerous current self-supervised monocular depth estimation approaches[6][7][8][9][16] are further researched.

2 Development of Monodepth2

In monodepth2, the current frame is projected to neighbor images using the estimated depth and pose. If both the depth and pose change in the same direction, such as the pose being estimated to be further by 1 meter and the depth being estimated to be closer by 1 meter, the projected image can still appear identical to the real one due to the compensating effects of the unchanged pair of depth and pose. The aforementioned phenomenon is also known as Scale ambiguity, posing a significant challenge in the context of monocular depth estimation. It should be pointed out that approaches in stereo training[12][14][15] and domain-adaptation training[7][10][11][12][13] do not suffer from the scale ambiguity issue because they can recover the metric depth with either known camera motion between binocular images or ground truth labels in virtual datasets, exhibiting a higher level of accuracy compared to monocular depth estimation. To close the performance gap with their counterparts, Chawla et al.[16] introduced additional GPS information into the training pipeline to address scale ambiguity and Guizilini et al. [6] proposed incorporating velocity information as weak supervision in the pose estimation stage to explicitly constrain the

estimated pose, showcasing superior accuracy. Focusing on scale consistency, Wang et al.[17] put forward a novel scale-aware geometric loss, which helped to conserve the scale, and a two-stream depth network to disentangle depth and scale prediction, further ensuring scale consistency. Jiang et al.[18] presented a scale-invariant approach in which scale-sensitive features were detached away while scale-invariant features were boosted further, making the model robust to scale changes. The vast majority of monocular networks based on monodepth2 [2] do not make use of sequence information in the form of video frames while it is available at most time. Based on this observation, Watson et al.[8] and Bangunharcana et al. [9] proposed novel methods that made use of sequence information at test time when available. In the following sections 2.1-2.3, we will introduce briefly how these methods have been optimized and improved in terms of scale ambiguity, robustness, and multi-frame testing.

2.1 Scale Ambiguity

Most self-supervised approaches on monocular videos suffer from scale ambiguity across long sequences due to the absence of scale information. Consequently, the depth of objects is inconsistent with the true camera motion. An additional way to elaborate on this issue is that the predicted distance moved forward of consecutive pairs of images in one video sequence varies drastically because of the compensating effects of depth and pose mentioned in the last section. Guizilini et al. [6] utilized velocity information to construct their velocity loss(L_v). During training, they imposed L_v between the magnitude of the pose-translation component of the pose network prediction $\hat{t}$ and the measured instantaneous velocity scalar v multiplied by the time difference between target and source frames $\Delta T_{t\to s}$, as shown below:

$$L_v(\hat{t}_{t\to s}, v) = \left|\left|\hat{t}_{t\to s}\right|\right| - |v|\Delta T_{t\to s}| \tag{5}$$

The aforementioned loss function explicitly constrains the dissimilarity between the predicted distance and the actual distance. Given the plausible pose estimation, the depth values have a unique solution that minimizes the photometric loss in (2) according to (1), addressing the issue of scale ambiguity in terms of pose orientation. However, access to instantaneous velocity may require the use of inertial measurement units (IMU) that are less ubiquitous. In contrast, GPS is more feasible, such as in dashboard cameras albeit with lower frequency, allowing training on more data [16]. Based on this observation, Chawla et al. [16] utilized the ratio of the relative distance measured by the GPS and the relative distance predicted by the network to additionally impose a novel loss function given by:

$$L_{g2s} = \sum_{s,t}\left(\frac{\left|\left|G_{s\to t}\right|\right|_2}{\left|\left|\hat{T}_{s\to t}\right|\right|_2} - 1\right)^2, \tag{6}$$

where s ∈ {-1,1} and t ∈ {0}. Minimizing the above loss function forced the estimated distance to be identical to the GPS information.

In the aforementioned methods, they both introduced additional information like GPS and velocity to supervise ego-motion, leading to scale-aware depth estimation. However, velocity or GPS information is not always available during training, it also requires additional hardware that is often prone to operational noise. Furthermore, fairness cannot be ensured when comparing methods that incorporate additional information such as GPS with those that do not. Therefore, predicting scale-aware and scale-consistent depth maps is a highly-discussed research area. In the absence of ground truth labels for dense depth maps, it is challenging to achieve absolute depth. However, there is a higher likelihood of ensuring consistency in such conditions.

Wang et al. [17] proposed a two-stream depth network to separately predict the depth and scale factor. Therefore, the scale-invariant depth $\overline{D}$ and scale factor μ should be multiplied together to get the scale-consistent depth $D = \mu\overline{D}$. Under the hypothesis that the scale factor between the estimated depth and the ground truth should be a constant in one sequence, after the depth map of the source image is projected onto the target view with the relation given in (1), two depth maps should stay identical. Because one cannot operate interpolation on depth maps, Guizilini et al. transferred depth maps to cloud points maps, which supportted similar transformations in (1). The transformation equation is given by:

$$P_t^{ij} = K^{-1} D_t^{ij} [i, j, 1]^T, \tag{7}$$

where K denotes the camera intrinsic; [i, j, 1] indicates the homogeneous coordinate of a pixel at location [i, j] of the image plane; while P_t^{ij} and D_t^{ij} represent the corresponding cloud point and depth of that pixel.

Given the correspondences between two point clouds P_s, $\widehat{P_s}$ are given, the least-square estimation can be formally described as follows,

$$\Phi, \Gamma, \tau = \arg\min_{\hat{\Phi}, \hat{\Gamma}, \hat{\tau}} \Sigma \left\| \hat{\tau} \hat{\Phi} \hat{P}_s^i + \hat{\Gamma} - P_s^{N(i)} \right\|_2^2 \tag{8}$$

where i and $N(i)$ indicate the indices of two corresponding points. The above least-square estimation can be solved in a closed-form as shown in [19]. If the predicted relative depth and ego-motion are accurate, the scaled point cloud $\tau\hat{P}_s$ should be perfectly aligned with P_s. Otherwise, the estimated rotation Φ and transformation Γ in (8) will imply their misalignment, thus, the loss function can be constructed to penalize the inaccurate predictions by forcing the estimated transformation Φ and Γ to approximate an identity mapping. However, the structure-from-motion constraints allow the method to learn depth and ego-motion only to an unknown scale, thus it still needs ground truth for evaluation on the standard dataset[11]. Therefore, obtaining scale-awareness depth in monocular depth estimation is a problem that remains unsolved.

2.2 Robustness

Monocular depth estimation is severely sensitive to scale changes especially when all the training samples are from one single camera. To be specific, the camera intrinsic stays the same in one dataset and the estimated depth varies drastically when cropping an image and testing the image again. To make the model robust to scale changes, Jiang et al.[18] presented a scale-invariant approach in which a simple but effective data augmentation by imitating camera zooming process was proposed to detach scale-sensitive features that can confuse the model in scale. Similar to the training pipeline of monodepth2, they also fed original image pairs into the pose and depth estimation network to get the relative pose $T_{t\to s}$ and per-pixel depth value. Additionally, original image pairs were cropped as the zoomed dataset I^z, which will be only fed into the depth net to get the augmented depth map D^t. Simultaneously, the camera intrinsic should multiply a cropping factor that controlled the cropping area in the original images to be used in the warping operation in (1). Then the estimated depth of zoomed depth map was required to be aligned with the original estimated camera pose with the relation equation in (1) by minimizing the photometric loss in (2). This ensured depth maps of cropped images to be identical to the original one, making the model robust to scale changes. The naïve concepts of scale-sensitive features and scale-invariant methods were introduced for the first time in self-supervised monocular depth estimation and also to a certain extent addressed the issue of scale ambiguity.

2.3 Multi-frame Testing

Recently, it is acknowledged that sequence information in the form of video frames can be utilized when it is available at test and training time, incorporating more information into the model when predicting the depth.

Based on multi-frame training, Watson et al. [8] introduced a novel approach that combined the strengths of monocular and multi-view depth estimation. They also proposed an adaptive cost volume to overcome the scale ambiguity arising from self-supervised training on monocular sequences. To be specific, by warping features extracted from images at different time points and different candidate depth to the target image view, an adaptive cost volume is constructed, which contains information of adjacent frames. The cost volume was concatenated with the target image feature and used as input to a convolutional decoder which regressed the depth map.

They involved the construction of cost volume from multiple views to compute pixel correspondences, bearing similarities to stereo models. By incorporating multi-frame data, geometric information was integrated to improve the performance as well as the robustness of their model. However, adaptive cost volume suffered from the unknown depth range to be chosen as candidates and relied on an estimated minimum and maximum depth to get scale information of the estimates[9].

Bangunharcana et al. [9] introduced an iterative update model that was based on epipolar geometry and direct alignment, which was also known as Deep equilibrium alignments (DEQ). At each iteration in the update step, matching costs based on the feature map of source and target images are computed and are used as the input of the Conv-GRU blocks, in which the predicted depth and pose are updated. DEQ-based alignments are performed to find the fixed point that minimizes the matching costs and outputs the final predictions. The iterative updates formulation integrated multi-view information and compute the local epipolar geometry as the foundation of refinement, achieving new State-Of-the-Art performance in monocular depth estimation. They also showed competitive results compared with stereo training and supervised training.

Due to the absence of scale information in monocular depth estimation, all existing methods in monocular depth estimation need to utilize the ground truth labels to scale the depth to a reasonable range, which is unacceptable theoretically speaking. The problem of scale ambiguity, coupled with the absence of ground truth in real-world scenarios, poses a significant barrier to the integration of monocular depth estimation methods into industrial applications. Therefore, scale ambiguity remains a persistent issue that urgently requires a solution.

References

[1] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1851–1858, 2017.

[2] Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J., 2019. Digging into self-supervised monocular depth estimation, in: Proceedings of the IEEE international conference on computer vision, pp. 3828–3838.

[3] Hang Zhou, David Greenwood, and Sarah Taylor. Selfsupervised monocular depth estimation with internal feature fusion. arXiv preprint arXiv:2110.09482, 2021.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[5] Zhang S, Zhao C. Dyna-DepthFormer: Multi-frame Transformer for Self-Supervised Depth Estimation in Dynamic Scenes[J]. arXiv preprint arXiv:2301.05871, 2023.

[6] Guizilini V, Ambrus R, Pillai S, et al. 3d packing for self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 2485-2494.

[7] Cheng R, Razani R, Taghavi E, et al. 2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 12547-12556.

[8] Watson J, Mac Aodha O, Prisacariu V, et al. The temporal opportunist: Self-supervised multi-frame monocular depth[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1164-1174.

[9] Bangunharcana A, Magd A, Kim K S. DualRefine: Self-Supervised Depth and Pose Estimation Through Iterative Epipolar Sampling and Refinement Toward Equilibrium[J]. arXiv preprint arXiv:2304.03560, 2023.

[10] Zheng C, Cham T J, Cai J. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 767-783.

[11] Swami K, Muduli A, Gurram U, et al. Do What You Can, With What You Have: Scale-aware and High Quality Monocular Depth Estimation Without Real World Labels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 988-997.

[12] Zhao S, Fu H, Gong M, et al. Geometry-aware symmetric domain adaptation for monocular depth estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9788-9798.

[13] Guizilini V, Li J, Ambruș R, et al. Geometric unsupervised domain adaptation for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 8537-8547.

[14] GonzalezBello J L, Kim M. Forget about the lidar: Self-supervised depth estimators with med probability volumes[J]. Advances in Neural Information Processing Systems, 2020, 33: 12626-12637.

[15] Wang R, Yu Z, Gao S. PlaneDepth: Plane-Based Self-Supervised Monocular Depth Estimation[J]. arXiv preprint arXiv:2210.01612, 2022.

[16] Chawla H, Varma A, Arani E, et al. Multimodal scale consistency and awareness for monocular self-supervised depth estimation[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 5140-5146.

[17] Wang L, Wang Y, Wang L, et al. Can scale-consistent monocular depth be learned in a self-supervised scale-invariant manner?[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 12727-12736.
[18] Jiang P, Yang W, Ye X, et al. Detaching and Boosting: Dual Engine for Scale-Invariant Self-Supervised Monocular Depth Estimation[J]. IEEE Robotics and Automation Letters, 2022, 7(4): 12094-12101.
[19] Umeyama S. Least-squares estimation of transformation parameters between two point patterns[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1991, 13(04): 376-380.