Monocular Depth Estimation

Project Proposal

Rita Maatouk
Université Paris-Saclay
307 Rue Michel Magat, 91400 Orsay, France
rita.maatouk@universite-paris-saclay.fr

Maxence Gollier
Institut Polytechnique de Paris
19 Pl. Marguerite Perey, 91120 Palaiseau, France
maxence.gollier@ip-paris.fr

1. Motivation and Problem Definition

Monocular depth estimation based on deep learning is a fascinating area in computer vision. The goal is to predict the depth information of a scene using only a single input 2D image. Essentially, it's about teaching a model to understand the relative distances of different objects in a given image. This technology has various applications, such as: autonomous vehicles (monocular depth estimation helps self-driving cars perceive the depth of objects in their surroundings, aiding in navigation and obstacle avoidance), robotics (robots equipped with monocular depth perception can better navigate and interact with their environmen), photography (depth information can be used to create realistic depth-of-field effects in photos or assist in 3D scene reconstruction)... The project is inspired by the 3D stereo vision projects covered in the 3D Computer Vision course at M2 MVA. Stereo vision is a conventional method for estimating image depth. It relies on binocular cameras and involves calculating the disparity between two 2D images, captured simultaneously from slightly different perspectives, through stereo matching and triangulation to generate a depth map. Monocular depth estimation is gaining prominence as it operates with a single camera, eliminating the need for complex equipment and specialized techniques. The practicality of using just one camera aligns with the commonality of such setups in numerous application scenarios. this challenging task in computer vision, employs diverse deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs) to extract spatial and temporal features, pushing the boundaries of understanding depth from single images. Convolutional Neural Networks (CNNs) are commonly used for monocular depth estimation tasks. The network takes an image as input and outputs a corresponding depth map. During training, the network learns to minimize the difference between its predicted depth maps and the ground truth depth maps from the dataset. Recurrent Neural Networks (RNNs), with their sequential learning capabilities, introduce a temporal aspect to the depth estimation process. This can be particularly useful in scenarios where understanding the temporal evolution of depth information is crucial, such as in videos or dynamic environments. Generative Adversarial Networks (GANs) add another dimension to monocular depth estimation. GANs consist of a generator and a discriminator network that work in tandem, fostering a competitive learning process. This adversarial training enables the model to generate more realistic and accurate depth maps, pushing the boundaries of what can be achieved with monocular input. One popular approach is using encoder-decoder architectures, where the encoder extracts features from the input image, and the decoder generates the depth map. Skip connections are often employed to combine features from multiple scales, aiding in capturing both fine and coarse details. There are different ways to train these networks. Supervised methods use labeled data, where the network learns from images paired with known depth information. Unsupervised methods rely on the relationships between frames in sequences of images. Semi-supervised methods strike a balance, using both labeled and unlabeled data to improve accuracy. These approaches have their strengths and challenges, and their performance is often compared using datasets like KITTI and NYU Depth v2. Each method contributes to the ongoing effort to make accurate depth predictions from single images.

2. Methodology

Clearly, the first step of this project will be to make a larger literature review in order to solve the Monocular Depth Estimation problem. In a second stage, we aim to compare supervised techniques with unsupervised ones from *e.g.* [2, 3]. In both cases, there is a variety of different Deep Learning architectures that can be found in the litterature such as CNNs [2], GANs [1] or RNNs [5] can be used. Recently, transformer based methods have also been developed [4]. Some of these methods are already well-established in the computer vision community. A popular method [3] is to replace the depth finding problem with the problem of finding pixel-level correspondence between pairs of rectified stereo images that have a known camera baseline. We can then estimate depth by finding the pixelwise disparity from the image pairs. In this case, easier-to-find stereo images data-sets can be used during training. We might try to reproduce this work and to compare the results from these articles. This subject is still very active, hence, a lot of extensions can still be done.

3. Evaluation

For the data-set, we can use either the NYUv2 data-set which is widely used in the state-of-the-art literature and which is a set of RGB-D (RGB-Depth) images. This data-set will be useful for supervised tasks. For the stereo image data-set we choose the KITTI data set which contains traffic scenarios with stereo cameras. Since these scenes often look the same as they come from the same context, it would be interesting to experiment how they generalize for instance to scenes from the NYUv2 set.

[2] have evaluated a variety of cost functions and we can aim to use their results in ordrer to settle for a common evaluation function.

References

[1] Filippo Aleotti, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. *Generative Adversarial Networks for Unsupervised Monocular Depth Prediction: Munich, Germany, September 8-14, 2018, Proceedings, Part I*, pages 337–354. 01 2019. 1

[2] Marcela Carvalho, B. L. Saux, Pauline Trouvé, Andrés Almansa, and Frédéric Champagnat. Estimation de profondeur mono-image par réseaux de neurones et flou de défocalisation. 2018. 1, 2

[3] C. Godard, O. Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6602–6611, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. 1

[4] Dong-Jae Lee, Jae Young Lee, Hyounguk Shon, Eojindl Yi, Yeong-Hun Park, Sung-Sik Cho, and Junmo Kim. Lightweight monocular depth estimation via token-sharing transformer, 2023. 1

[5] Rui Wang, Stephen Pizer, and Jan-Michael Frahm. Recurrent neural network for (un-)supervised learning of monocular video visual odometry and depth. pages 5550–5559, 06 2019. 1