Deep Learning - Project proposal
Monocular Depth Estimation for *in-the-wild* Auto-focus application

Younes Belkada Antoine Cadiou
École Normale Supérieure Paris-Saclay
firstname.lastname@ens-paris-saclay.fr

Abstract

*Recent works have shown that in the real world, humans rely on the image obtained by their left and right eyes in order to estimate depths of surrounding objects. Thus, depth estimation is a classic task in computer vision, which is of great significance for many applications such as augmented reality, target tracking and autonomous driving. We firstly summarize the deep learning models for monocular depth estimation. Secondly, we will implement a recent Vision Transformers based architecture for this task. We will seek to improve it by adding a segmentation head in order to perform multi-task learning using a customly built dataset. Thirdly, we will implement our model for in-the-wild images (*i.e. *with no control on the environment, the distance and size of objects of interests, and their physical properties (rotation, dynamics, etc.)) for Auto-focus application on humans and will give qualitative comparison across other methods.*

1. Introduction

Depth estimation task could be achieved with 2 inputs/images for the same scene but taken from 2 different places. This method corresponds to a binocular Depth-estimation. Actually, it is difficult to deal with this kind of model because we need a specific dataset, and also the applications are limited because we need the appropriate equipment (2 cameras for instance). This limitation is at the origin of our work: given a single view of a scene (monocular camera) be able to do the same thing.
This Monocular Depth estimation work will finally be applied to a specific task: the auto-focus on humans. But the latest has a challenge, ideally we need given an image with a human the segmentation mask of the image and the depth map of the image in order to blur all the object that do not have the same depth distribution as the human. Ideally at inference time, we would have a single model that outputs both maps given a single image.

2. Related Work

Active depth estimation methods usually utilize lasers, structured light and other reflections on the object surface to obtain depth point clouds, complete surface modeling and estimate scene depth maps. However, for some specific applications such as autonomous driving, robotics and auto-focus, the hardware constraint is a hard constraint. Thus, in this context the depth estimation needs to be computed given a single image.

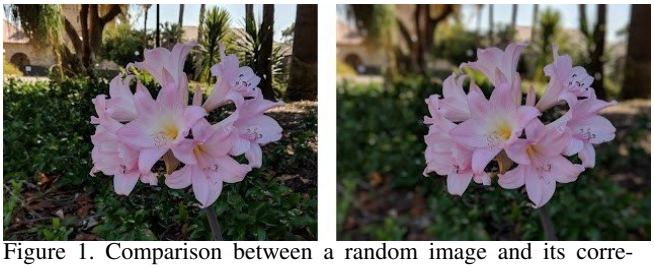

Figure 1. Comparison between a random image and its corresponding *focuses* image. We would like to achieve this task using Deep Learning based methods

To the best of our knowledge, there are 2 different methods that have already been explored to estimate the depth with a monocular point of view: CNNs based models[2] [6](*i.e.* GANs or Unets) and Vision Transformers[3] [5]. Both of these architecture could be learned following supervised or unsupervised strategies. [1].

3. Proposed Solution

We will work on a Vision Transformer based model entitled *Dense Prediction Transformers* [3] for 2 main reasons. Firstly, because the latest is currently the new state of the art for the Monocular Depth estimation task on various datasets. Secondly, we are enthusiasts on learning about the new *trend* in Computer Vision which are the Transformers. We would like to follow the trend and try to learn how the black box works by implementing by our own, and trying to improve the current state of the art method. We would like to achieve this goal by combining image segmentation and

Figure 2. Examples of depth estimated pictures using the method presented in[5]

Dense Prediction Transformers[4], and try our approach on a dataset based on video frames for auto-focus tracking. Our solution would be summarized by the following steps:

- Creating a dataset that is adapted for the task (*e.g.* by combining several open-source datasets (**??**)). For this dataset, each image would contain its associated depth map, and the segmentation mask that contains the label *human*. We can build this dataset by using current state-of-the-art methods (such as [4]) for image segmentation on a dataset that contains depth estimation ground truths or vice-versa. To achieve this task, we will aim to build a dataset that will be a mix of 2-3 datasets following [5] (*e.g.* 3D Movie, Kitti, NYUv2, MegaDepth).

- Try to implement and modify the DPT[4] model **from scratch**. Train the model on our dataset for image segmentation and depth estimation separately.

- Train our DPT[4] model with 2 heads, one head for human segmentation and another on depth estimation.

- Compare the performance of our implementation against other methods (preferably a CNN based model) or the original method on a specific test set for depth estimation (*e.g.* NYUv2 test set). Following [4] we will attempt to use simple and common metrics for this task such as pixel-wise RMSE for evaluating our methods.

- Implement a simple image focus tool using the trained model.

- **Optional:** Implement an auto-focus system (*i.e.* capable to track something previously selected) from a video stream. We will be using our model for depth-estimation mixed with a solution to track the selected instance (*e.g.* an image segmentation model)

References

[1] Shir Gur and Lior Wolf. Single Image Depth Estimation Trained via Depth from Defocus Cues. *arXiv:2001.05036 [cs]*, Jan. 2020. arXiv: 2001.05036. 1

[2] Yue Ming, Xuyang Meng, Chunxiao Fan, and Hui Yu. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33, May 2021. 1

[3] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. *arXiv:2103.13413 [cs]*, Mar. 2021. arXiv: 2103.13413. 1

[4] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 2

[5] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 1, 2

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 1