Advanced Ranking Sampling Strategies for Monocular Depth Estimation

Master Thesis Proposal & Work Plan

Praneeth Balakrishna

matriculation number: 6868873

praneeth@mail.uni-paderborn.de

June 18, 2021

1 Motivation

Depth estimation is a well studied problem in computer vision. It is an important step in the 3-D reconstruction of scenes and is crucial in fields like robot-vision and self driving cars. Monocular depth estimation aims to solve this problem using images obtained from a single camera. Humans can easily infer depth from single images with sufficient samples learnt(how near and far objects appear) over the lifetime. In contrast, for a computer vision system, monocular depth estimation is an ill posed problem due to its inherent ambiguity in mapping RGB pixel values to depth.

Typically, this is solved as a regression problem where the learning model learns to predict the depth map of the given RGB image. Deep learning models, especially Convolutional Neural Networks(CNN) have achieved good success in learning such models, as demonstrated in [EPF14]. However, not all applications require exact metric depth prediction. Instead, prediction of an ordinal relation is sufficient in such cases, *e.g.*, to detect occlusion boundaries in the image for an augmented reality application. This also provides an opportunity for models to learn from pseudo depth data [ZIKF15, CFYD16]. Furthermore, such an ordinal classification problem can be solved as a ranking problem [Liu11]. [XSC$^+$18, XZW$^+$20] employ learning-to-rank methods on pairwise training samples, minimizing pairwise ranking loss in order to predict a dense pseudo-depth map. However, pairwise prediction leads to loss of information, particularly information about the transitivity of the order relation [LHEN21]. In order to overcome this issue, list-wise ranking was proposed [CQL$^+$07], where an arbitrary number of samples are chosen as a list and the model is trained to rank the samples in the order of their relative depth [LHEN21]. This is achieved by minimizing a permutation loss function which computes the inconsistency between the output ranking predicted by the model and the ground truth permutation. In the monocular depth estimation methods described in [XSC$^+$18, CFYD16, LHEN21], sampling methods utilized for training the model is predominantly random sampling of pairs or lists. Random sampling leads to two potential issues. Firstly, it leads to imbalanced ordinal relations [XSC$^+$18] and secondly, it misses out the fact that certain pixels in the image provide better depth cues than others.

In order to further improve the performance of list-wise learning-to-rank method for monocular depth estimation [LHEN21], two potential methods are explored in this thesis. Firstly, an informativeness score is computed over the training data and the learning model is trained on highly informative samples that provide better cues for depth prediction. Secondly, a query strategy on a partially trained model is developed such that, the model queries samples from the unlabeled train-

ing data about which it is most uncertain with respect to relative depth prediction. Thereby, choosing to learn from training samples that it is most uncertain of. Such a technique lies in the realm of active learning [Set09].

2 Related Work

2.1 Metric depth estimation

Experiments conducted by Saxena *et al.* [SCN+05] are one of the first studies using Markov Random Field(MRF) to predict depth from monocular images. Since then, various deep neural network(DNN) based methods have been introduced. Training large RGB datasets on deep learning based CNNs have resulted in commendable performance in monocular depth estimation. Typically, it was solved as a regression problem in a supervised learning format. In [EPF14], metric depth is regressed using a CNN in two stages. The first stage estimates the global structure of the scene and the second refines it using local information. In the works of Liu *et al.* [LSLR15], a deep convolutional neural field model was proposed which combines CNN models with continuous conditional random field(CRF). Furthermore, attention models that were previously used in natural language processing was introduced in monocular depth estimation. Huynh *et al.* [HNHM+20] devised depth-attention volume model which enhanced depth estimation by capturing non-local depth dependencies between co-planar points. The contemporary state of the art method is based on vision transformers. [RBK21] proposed dense prediction transformers(DPT) for monocular depth estimation which results in more fully grained coherent predictions when compared to CNN based architectures. In contrast to supervised learning methods, [GBCR16] predicts a depth map using unsupervised learning techniques. Recent works by Zhu *et al.* [ZBL20] utilize the border constraint between semantic segmentation and depth estimation such that the inconsistency between the segmentation edge and the depth edge is used as the loss function.

2.2 Relative depth perception

Learning to Rank

Learning to rank is a popular machine learning technique used in Information Retrieval(IR) [CQL+07]. The task of such models is to construct a ranking model using training data, such that it can sort new samples based on their degree of preference. Ranking methods mainly fall under two categories: pair-wise methods and list-wise methods. For a given query, pair-wise methods order/classify a pair of samples according to its importance. It is trained using a pair-wise loss function

as in [CFYD16]. A query in list-wise method consists of a list of samples and the method sorts the elements of the list according to a degree of preference. List-wise methods define losses over the set of permutation of the samples in the list [Liu11]. It has been observed that list-wise methods yield a better performance than pairwise ranking methods [CQL+07].

Depth Estimation

Certain applications like 2D-to-3D conversion and depth-of-field do not demand the exact metric depth prediction. Models for such applications can be trained on pseudo-depth information. Such models can be designed as ranking models described above, which rank pixels/objects based on their pseudo depth. Zoran *et al.* [ZIKF15] uses a CNN model that is trained on pseudo-depth samples taken pair-wise on the basis of superpixel segmentation and later derives a global depth map by solving an energy optimization problem. Further, Chen *et al.* [CFYD16] proposed a "Depth in the Wild"(DIW) dataset with images containing manually annotated sample point pairs with their relative depth. A CNN model was trained with this data by minimizing a ranking loss function to predict a dense metric depth map. However, the DIW dataset only provides labels for only one pair of points in every image, hence missing out on a large amount of perception relevant information. To overcome this drawback, Xian *et al.* [XSC+18] introduced a method to automatically produce dense relative depth annotations from web stereo images. Now, multiple sample pairs were randomly sampled in order to train a ResNet based CNN model to minimize an improved ranking-loss function. Chen *et al.* [CQD19] proposed a method to automatically generate monocular depth data by integrating structure from motion(SfM) and a quality assessment network.

In the depth estimation methods described so far, training sample pairs were sampled randomly from images without considering depth cues or structure in the image which contains higher information and have the potential to yield a better performance in depth estimation. In this regard, Xian *et al.* [XZW+20] proposed a novel structure guided sampling technique. The authors stated that samples in the image across edges and segments contain higher structural information and are hence, more relevant in depth prediction. Therefore, the ResNet50 based model was trained on such selective samples to minimize a novel structure guided ranking-loss [XZW+20].

Relative depth prediction models trained on pair-wise samples in [ZIKF15,CFYD16, XSC+18, XZW+20] have proven to be effective. These models solve depth estimation as a pair-wise ranking problem by reducing a pair-wise ranking loss function. However, the fact that a large number of sample pairs can be constructed, renders

these methods to be rather inefficient. In order to overcome this problem, list-wise ranking [XLW+08] was introduced as an alternate to pair-wise ranking. The inherent feature of list-wise ranking is that, it allows for higher order ranking of arbitrary length data to be training samples. It is also demonstrated that list-wise ranking yields higher performance than pairwise ranking [XLW+08].

Monocular relative depth estimation was formulated as a list-wise ranking problem in [LHEN21]. Training samples were sampled randomly as lists and trained on PLDepthEffNet model [LHEN21] by optimizing the ListMLE loss function [XLW+08]. In ListMLE, the permutations of the list is modeled as a probability distribution using the Plackett Luce model. The ListMLE loss is the negative log likelihood of the Plackett Luce probability distribution.

3 Goals

The fundamental goal of the thesis is to solve the monocular depth estimation (of pixels) problem as a learning-to-rank task. However, the challenge is to achieve this by weakly supervised training data where the learning model does not learn from accurate metric depth, but instead learn from relative-depth training data. The methods described in Sec 2 achieve laudable results in solving such a problem. The focus here is to improve those methods in order to achieve better results.

3.1 Improved sampling strategy

In methods [XSC+18, CFYD16, LHEN21], the learning models are trained with samples(pairs or lists) that are randomly sampled from the training images. There are certain aspects of the image, like depth-edges, texture, occlusion, etc. that provide depth cues and random sampling misses out on entirely utilizing such information and hence, the model learns mostly from less informative samples. Therefore, the aim is to propose an information measurement score on the training samples and develop a superior sampling strategy such that informative samples are more likely picked during training.

3.2 Active learning

Active learning is a case of machine learning where the learning algorithm chooses the data on which it learns [Set09]. The model poses queries on unlabeled data to an oracle(eg. a human annotator), for the label. By this method, the learning model can learn effectively on fewer number of training samples.
Therefore, the second aim of the thesis is to develop an active learning algorithm

to query uncertain samples for labeling, during prediction. Additionally, an oracle which can label the queried samples is to be formulated.

Finally, an analysis of the performance improvement with both the above techniques is to be conducted in comparison to the earlier methods.

4 Approach

4.1 Sampling based on informativeness score

The dataset consists of RGB-D images with pseudo depth annotation which is represented as a tuple $(I, D) \in (\mathbb{R}^{h \times w \times 3}, \mathbb{R}^{h \times w})$ and ρ is the ranking assigned to a particular image sample. The training samples are in the form of lists of length n, whose elements are pixels of the image (I, D). n such pixels $M = \{p_1, p_2 .... p_n\}$ are sampled and ranked according to the pseudo depth values, i.e., $\rho_\pi : l_{\pi(1)} > l_{\pi(2)} > \cdots > l_{\pi(n)}$, where $\pi \in \mathbb{S}_n$ is a permutation of $[n] := 1, ..., n$ such that $D[p_{\pi(1)}] < D[p_{\pi(2)}] < \cdots < D[p_{\pi(n)}]$, where $D[p_{\pi(i)}]$ is the depth of the element ranked at the i'th position, as described in [LHEN21]. N such ranked lists with randomly sampled pixels are initially obtained for every image to form a pool of sample lists S. The following two strategies compute an informativeness score on the samples in S. Based on the combined score, k most informative samples can be selected, where $(k, N \in \mathbb{N}$ with $k < N)$.

Figure 1 Image consists of pixels with diverse depths.

9
10
10
12
15
80
85
90

10
20
30
40
50
60
70
80

Figure 2 Comparing informativeness of sample lists containing pixel-depth values for $n = 8$. Left list is a (less informative) and the right list is b (more informative)

Heterogeneity in distance

Heterogeneity in distance can be used as an informativeness measure on the samples in S. This assigns a higher informativeness score to sample-lists M, which have pixel samples with diverse depths. For example, the image in Figure 1 contains objects at different depths. Let sample-lists a and $b \in S$ contain the depth values of the sampled pixels as shown in Figure 2. Assuming the range of depths of the pixels/objects in the image is $[0, 100]$ units, it is observed that sample-list a consists of pixels elements that do not represent the entire range of depths. Whereas sample-list b is a better representative of the entire depth range. Hence, in this example, sample-list b is more informative than a. In order to measure heterogeneity, popular goodness of fit tests like the chi-square test [MBN11] can be used to compare the elements of the list $M \in S$ to a list E of n elements that are uniformly distributed within the range of possible depths in the image, which can also be termed as the expected depth of the pixel. The test statistic for a chi-squared test is:

$$\chi^2 = \sum_{1}^{n} \frac{(D_i - E_i)^2}{E_i}$$

where D_i is the depth of the element in the list M, at position i and E_i's are evenly spaced values between (min_depth, max_depth) in the corresponding image. χ^2 can be used as the informativeness score I_d ($I_d = \chi^2$), $\forall\ M \in S$, for heterogeneity in distance.

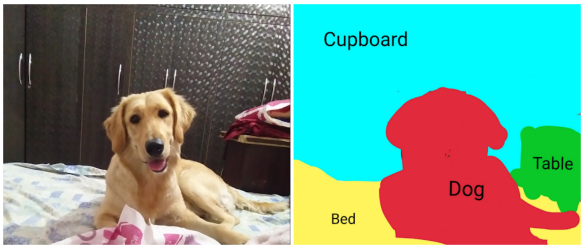

Figure 3 To demonstrate samples with heterogeneity across instance segments [Nis21]

C
C
C
C
C
C
D
T

C
C
B
B
D
D
T
T

Figure 4 Comparing informativeness of sample lists for $n = 8$. Left list c (less informative) and Right list d (more informative)

Heterogeneity across instance segments

Here, a higher informativeness score is assigned to sample lists M which have pixel samples from diverse instance segments. Instance segmentation is the detection of each distinct object of interest in the image. For example, Figure 3 contains four distinct objects namely $\{Cupboard, Bed, Dog, Table\} \leftrightarrow \{C, B, D, T\}$. Let sample-list c and $d \in S$ contains the segment-id of the sampled pixels as shown in Figure 4. The elements of sample-list c consists mainly of pixel representation from the $Cupboard$ instance segment (C) whereas sample-list d conists of a uniform representation of all the instance segments in the image (C, B, D, T) which is therefore a more informative sample-list. To achieve this, first, an instance segmentation is performed on the training images using Mask R-CNN [HGDG17] to generate a segmentation map for each image in the training dataset. A chi-squared test similar to the one described in Sec 4.1 can be used to derive an informativeness score I_s, based on heterogeneity across instance segments.

Total informativeness score

The total informativeness score for each sample $M \in S$ could be a weighted sum of I_d and I_s.

$$I_t = a \cdot I_d + b \cdot I_s \qquad a, b \in \mathbb{R} \text{ and } a, b > 0$$

I_t is computed for each sample $M \in S$ and the top k samples are chosen for training.

Algorithm 1: Usage of Informativeness score I_t

Require: Learning model L, training images I and list-size n
Initialize: Dataset $D = \{\}$, weights W
for $i \leftarrow 1$ **to** $|I|$ **do**
    Randomly sample N sample-lists of size n each ;
    Compute I_d, I_s and total Informativeness score I_t for each sample ;
    Select top k most informative samples K based on I_t ;
    Append $D \leftarrow D + K$;
end
`/* Training model` L `on the informative samples dataset` D `*/`
for $e \leftarrow 1$ **to** *epochs* **do**
    Sample batch of sample-lists b from D ;
    Train model L on the sampled batches: Compute loss value on b ;
    Optimize loss by updating weights W ;
end

4.2 Active learning

In active learning, the model queries the most uncertain samples among the unlabeled pool, to be labeled by an oracle. However this requires an uncertainty measure to be formulated. Drawing inspiration from [ZBL20], where the inconsistency between the segmentation edge and depth edge obtained from the depth map predicted by the model is used to construct a loss function, a similar approach could be used to measure the uncertainty in the model's prediction.

Let E_s be the edges extracted from the segmentation map and E_d be the edges extracted from the depth map predicted by the model. Choose r samples $q = \{q_1, q_2...q_r\} \in E_s$. For each segmentation edge point q_i, compute its nearest distance d_i, from the depth edge map E_d.

$$d_i = \min_{p \in E_d} ||p - q_i||$$

The elements of q are sorted with respect to d_i and the top $m(m < r)$ samples are considered the most uncertain samples and hence, sent as queries to the oracle. These samples are as a result of the assumption that an edge in the segmentation map must also appear as an edge in the depth map. Once labeled, these samples are re-trained, expecting a better performance in depth prediction.

5 Time-Schedule and Subtasks

1. **Dataset: Weeks: 1-4**
   a. Evaluate a suitable dataset with pseudo-depth annotation for images-in-the-wild.
   b. Implement the algorithm to measure informativeness of each pixel/segment.
   c. Implement the sampling strategy based on the informativeness score to sample ranking lists.

2. **ML model: Weeks: 5-7**
   a) Implement the PLDepthEffNet learning model as in [LHEN21].
   b) Implement the list-wise ranking loss function: ListMLE [XLW$^+$08]

3. **Active Learning: Weeks 8-11**
   a) Implement a query strategy to select the most uncertain samples.
   b) Implement an oracle for labeling the queried samples.

4. **Analysis: Weeks: 12-17**
   a) Evaluate the performance of the proposed methods on metrics like ordinal error and Discounted Cumulative Gain(DCG).
   b) Compare the performance against baseline models such as the ones in [LHEN21] and [XSC$^+$18].

5. **Documentation: Weeks: 16-21**
   The documentation is planned to be carried out in parallel to the tasks and finalized in the last 3 weeks.

6 Preliminary Document Structure

1. Introduction
2. Fundamentals
   - 2.1. Monocular Depth Estimation
   - 2.2. Learning to Rank
   - 2.3. Image Depth Learning Models
   - 2.4. Active Learning
3. Implementation
   - 3.1. Depth Estimation using List-wise Ranking
   - 3.2. Sampling Strategy for List-wise Ranking Model
     - 3.2.1. Random Sampling
     - 3.2.2. Sampling based on Informativeness Measure
   - 3.3. CNN Learning Model with Active Learning
     - 3.3.1. Query Strategy
4. Evaluation
   - 4.1. Accuracy of Depth Prediction
   - 4.2. Assessment of the Performance
     - 4.2.1. Improved Sampling
     - 4.2.2. Active Learning based Querying
5. Related Work
6. Conclusion
7. Literature
8. Appendix

References

[CFYD16] Chen, Weifeng; Fu, Zhao; Yang, Dawei; Deng, Jia: Single-image depth perception in the wild. In: *arXiv preprint arXiv:1604.03901* (2016)

[CQD19] Chen, Weifeng; Qian, Shengyi; Deng, Jia: Learning single-image depth from videos using quality assessment networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, S. 5604–5613

[CQL+07] Cao, Zhe; Qin, Tao; Liu, Tie-Yan; Tsai, Ming-Feng; Li, Hang: Learning to rank: from pairwise approach to listwise approach. In: *Proceedings of the 24th International Conference on Machine learning*, 2007, S. 129–136

[EPF14] Eigen, David; Puhrsch, Christian; Fergus, Rob: Depth map prediction from a single image using a multi-scale deep network. In: *arXiv preprint arXiv:1406.2283* (2014)

[GBCR16] Garg, Ravi; B.G., Vijay K.; Carneiro, Gustavo; Reid, Ian: Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. In: Leibe, Bastian (Hrsg.); Matas, Jiri (Hrsg.); Sebe, Nicu (Hrsg.); Welling, Max (Hrsg.): *Computer Vision – ECCV 2016*. Cham : Springer International Publishing, 2016, S. 740–756

[HGDG17] He, Kaiming; Gkioxari, Georgia; Dollár, Piotr; Girshick, Ross: Mask r-cnn. In: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, S. 2961–2969

[HNHM+20] Huynh, Lam; Nguyen-Ha, Phong; Matas, Jiri; Rahtu, Esa; Heikkilä, Janne: Guiding Monocular Depth Estimation Using Depth-Attention Volume. In: Vedaldi, Andrea (Hrsg.); Bischof, Horst (Hrsg.); Brox, Thomas (Hrsg.); Frahm, Jan-Michael (Hrsg.): *Computer Vision – ECCV 2020*. Cham : Springer International Publishing, 2020. – ISBN 978–3–030–58574–7, S. 581–597

[LHEN21] Lienen, Julian; Hullermeier, Eyke; Ewerth, Ralph; Nommensen, Nils: Monocular Depth Estimation via Listwise Ranking Using the Plackett-Luce Model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, S. 14595–14604

[Liu11] Liu, Tie-Yan: Learning to rank for information retrieval. (2011)

[LSLR15] Liu, Fayao; Shen, Chunhua; Lin, Guosheng; Reid, Ian: Learning depth from single monocular images using deep convolutional neural fields. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38 (2015), Nr. 10, S. 2024–2039

[MBN11] Meesad, Phayung; Boonrawd, Pudsadee; Nuipian, Vatinee: A chi-square-test for word importance differentiation in text classification. In: *Proceedings of International Conference on Information and Electronics Engineering*, 2011, S. 110–114

[Nis21] Nishad, Garima. *Semantic Segmentation: The easiest possible implementation in code!* Mar 2021

[RBK21] Ranftl, René; Bochkovskiy, Alexey; Koltun, Vladlen. *Vision Transformers for Dense Prediction.* 2021

[SCN+05] Saxena, Ashutosh; Chung, Sung H.; Ng, Andrew Y. et al.: Learning depth from single monocular images. In: *NIPS* Bd. 18, 2005, S. 1–8

[Set09] Settles, Burr: Active learning literature survey. (2009)

[XLW+08] Xia, Fen; Liu, Tie-Yan; Wang, Jue; Zhang, Wensheng; Li, Hang: Listwise approach to learning to rank: theory and algorithm. In: *Proceedings of the 25th International Conference on Machine learning*, 2008, S. 1192–1199

[XSC+18] Xian, Ke; Shen, Chunhua; Cao, Zhiguo; Lu, Hao; Xiao, Yang; Li, Ruibo; Luo, Zhenbo: Monocular relative depth perception with web stereo data supervision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, S. 311–320

[XZW+20] Xian, Ke; Zhang, Jianming; Wang, Oliver; Mai, Long; Lin, Zhe; Cao, Zhiguo: Structure-guided ranking loss for single image depth prediction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, S. 611–620

[ZBL20] Zhu, Shengjie; Brazil, Garrick; Liu, Xiaoming: The edge of depth: Explicit constraints between segmentation and depth. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, S. 13116–13125

[ZIKF15] Zoran, Daniel; Isola, Phillip; Krishnan, Dilip; Freeman, William T.: Learning ordinal relationships for mid-level vision. In:

Proceedings of the IEEE International Conference on Computer Vision, 2015, S. 388–396

Hereby supervisor and student confirm that this proposal is the basis for the topic assignment of the described work. The timetable and topic description are accepted by both sides as laid out in this proposal.

Supervisor	Student
(Prof. Dr. Eyke Hüllermeier)	(Praneeth Balakrishna)