Monocular Depth Estimation via Listwise Ranking using the Plackett-Luce Model

Julian Lienen[1,*]   Eyke Hüllermeier[2]   Ralph Ewerth[3,4]   Nils Nommensen[3,4]

[1] Paderborn University   [2] University of Munich (LMU)

[3] L3S Research Center, Leibniz University Hannover   [4] TIB Hannover

Abstract

In many real-world applications, the relative depth of objects in an image is crucial for scene understanding. Recent approaches mainly tackle the problem of depth prediction in monocular images by treating the problem as a regression task. Yet, being interested in an order relation *in the first place, ranking methods suggest themselves as a natural alternative to regression, and indeed, ranking approaches leveraging pairwise comparisons as training information ("object A is closer to the camera than B") have shown promising performance on this problem. In this paper, we elaborate on the use of so-called* listwise *ranking as a generalization of the pairwise approach. Our method is based on the Plackett-Luce (PL) model, a probability distribution on rankings, which we combine with a state-of-the-art neural network architecture and a simple sampling strategy to reduce training complexity. Moreover, taking advantage of the representation of PL as a random utility model, the proposed predictor offers a natural way to recover (shift-invariant) metric depth information from ranking-only data provided at training time. An empirical evaluation on several benchmark datasets in a "zero-shot" setting demonstrates the effectiveness of our approach compared to existing ranking and regression methods.*

1. Introduction

Estimating depth in monocular images constitutes a problem of practical importance when aiming to understand the geometry of a scene, e.g., in autonomous driving systems or for augmented reality applications. Due to its ill-posed nature, methods approaching this problem nowadays typically incorporate complex models, trained on large amounts of data using machine learning methods.

The majority of existing approaches tackles depth estimation (whether per-pixel or per-object) as a regression problem, i.e., as the problem of learning a model to predict a (pseudo-)metric map (e.g., [1, 9, 17, 18]). However, on the one hand, accurate prediction of metric depth actually depends on the intrinsic camera parameters, which are often not available. On the other hand, instead of predicting absolute depth, it is often enough to predict the *relative* depth of pixels or higher level concepts (such as objects), that is, to sort them from closest to farthest away from the camera.

One may then argue that regression is solving an unnecessarily difficult task, and rather advocate a formalization of depth estimation as a *ranking* task [12]. So-called "learning-to-rank" methods can be used to minimize suitable performance metrics based on relative errors. As absolute depth measurements are not necessarily needed, ranking has the additional advantage that it potentially allows for learning from weaker training information. This includes depth annotations that are not metric but can be regarded as pseudo-metric data, e.g., disparity maps constructed from stereo images or videos [6, 14, 20], or human-annotated data [5, 7]. Without the need for metric RGB-D data produced by depth sensors, the diversity of training datasets can be drastically increased due to cheaper data acquisition [33].

Existing ranking methods are essentially based on pairwise comparisons of the form "object A is closer to the camera than B" [5, 33, 34, 35]. Pairwise relations of that kind are sampled from a depth map as training information, and predictive models are induced by minimizing pairwise ranking losses. While these approaches have proven effective, the quadratic number of possible pairs that can be constructed renders them rather inefficient and necessitates sophisticated sampling strategies to eliminate less informative pairs [34]. Besides, breaking a linear order into pairwise comparisons necessarily comes with a certain loss of information. In par-

---

ticular, information about the transitivity of order relations, which is implicitly contained in a linear order, will be lost.

To avoid these drawbacks, so-called "listwise ranking" [32] has been proposed as an alternative to pairwise methods. In the listwise approach, higher order rankings of arbitrary length can be considered as training information. In this paper, we elaborate on the use of listwise ranking for depth estimation in images. More specifically, we propose a listwise ranking method based on the well-known Plackett-Luce (PL) model [22, 24], which allows for learning probability distributions over rankings from pseudo-metric data. Moreover, taking advantage of the representation of PL as a random utility model [27], we suggest a natural way to recover translation-invariant approximations of the underlying metric depth information. Along with that, we propose a state-of-the-art neural network architecture as a backbone, together with a simple sampling strategy to construct training examples from raw pseudo-depth data.

In a zero-shot evaluation, where we compare models on data not considered for training, we study the cross-dataset performance of our model and compare it with state-of-the-art approaches. Thereby, we demonstrate that listwise ranking is an effective approach for rank-based error minimization, and our model constitutes an appropriate choice for the prediction of depth orders in unseen scenes, as well as providing promising results in recovering metric depth.

2. Related Work

In learning to rank, the goal is to infer ranking models from training data in the form of rankings (permutations) of individual items. According to a rough categorization of methods, one can distinguish between pointwise, pairwise, and listwise approaches [21]. While single items are considered as training examples in pointwise learning-to-rank methods, relations between items are typically used as training examples in the other categories, either relations of order two (pairwise) or arbitrary length (listwise). In the case of pointwise learning-to-rank, examples are usually annotated by a score that determines their individual usefulness, from which, for instance, regression models can be induced. For pairwise approaches, where examples are typically given as single relations among two items, existing methods range from SVM-based classifiers [15] to boosting methods [11] and ranking networks [2]. Similarly, several listwise ranking methods have been proposed, in which examples are represented by higher order (potentially partial) item rankings. One of the most well-known representative is ListMLE [32], a maximum likelihood estimation method to infer Plackett-

Luce probability distributions over rankings.

Several approaches to tackle the problem of estimating depth in images using relative depth information for training have been proposed. Among the first, Zoran et al. [35] classify individual point pairs from an image, which are then combined into a global solution for a complete dense map over all image pixels. Following a similar motivation, Chen et al. [5] train a deep neural network architecture by using a pairwise ranking loss, directly predicting a dense-map in an end-to-end fashion. This approach has also been adopted in subsequent works and improved in various directions, for example by using a different model architecture [33], additional data [6], or improved sampling strategy [34]. Furthermore, Ewerth et al. [10] propose a method to estimate relative depth using a RankBoost model. Alternative approaches also exploit ordinal depth information [20], either directly or to pretrain regression models [4].

To learn models that work well for arbitrary scenes, e.g., in both indoor and outdoor scenarios, diversity of training data is crucial. Commonly used metric data produced by depth sensors typically provide limited diversity, e.g., NYUD-v2 [26] with indoor-only or KITTI [13] with only street scenes. Since maximal depth capacities of sensors constrain the recognizable depth, they fail to capture scenes "in the wild". This is why Chen et al. [5] propose a human-annotated dataset with pairwise point samples, for which the "closer-to-camera" relation is captured. However, as it provides ground truth information for only two points in each image, and the human annotation process is quite costly, other strategies aiming to automatically extract depth information have been proposed. For instance, stereo images [33] or sequences of images in videos [6] have been facilitated to predict structural disparity maps from the motion of elements. Combinations of such methods have been considered, too [31]. As none of them delivers metric information per pixel, the information produced must be considered as pseudo-depth, which, as previously explained, is still sufficient for depth relations. Although scale-invariant regression methods are also capable of learning from such data [20, 25], their ability to generalize to new datasets with structurally different scenes is fairly limited, at least for the task of depth ordering, as our empirical evaluation will confirm later on.

3. Plackett-Luce Model for Depth Estimation

In the following, we introduce our proposal of a Plackett-Luce model for depth estimation as illustrated in Fig. 1, along with a description of the model architecture and sampling strategy to construct training examples from raw depth data.
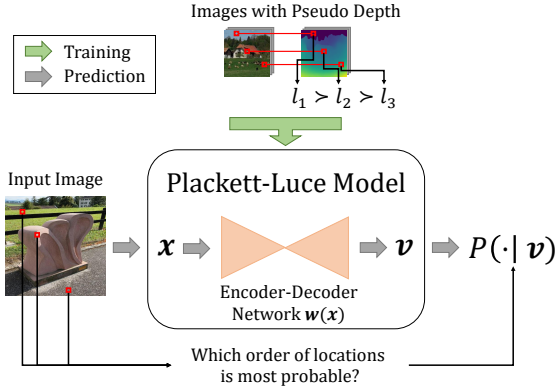
Figure 1. Overview of our method: The PL model incorporates a deep neural network to predict scores for each pixel in an input image, which are then turned into probabilities for rankings of queried image locations. For training, we sample rankings from images annotated by pseudo depth.

3.1. Problem Formulation

We assume training information in the form of RGB images I together with (pseudo-)depth annotations D, i.e., tuples $(I, D) \in \mathbb{R}^{h \times w \times 3} \times \mathbb{R}^{h \times w}$, where h and w denote the image height and width, respectively. Moreover, $D[l]$ denotes the (pseudo-)depth of a position $l \in \{1, \ldots, h\} \times \{1, \ldots, w\}$ identified by a height and width coordinate. Without loss of generality, lower values $D[l]$ encode shorter distances to the camera.

We are mainly interested in the order relation of the locations in an image I as induced by the (pseudo-)depth D. Formally, the relation between n locations $M = \{l_1, l_2, \ldots, l_n\}$ can be represented in terms of a permutation π of $[n] := \{1, \ldots, n\}$ such that $D[l_{\pi(i)}] < D[l_{\pi(i+1)}]$ for $i \in \{1, \ldots, n-1\}$. This permutation encodes the ranking $l_{\pi(1)} \succ l_{\pi(2)} \succ \cdots \succ l_{\pi(n)}$, i.e., location $l_{\pi(1)}$ is closest, then $l_{\pi(2)}$, etc. At query time, when I is given but D is not, the task of a rank-based depth estimation model is to predict the "closer-to-camera" relation $\succ$, that is, to produce an accurate order-preserving estimate of D. Formally, this estimate can again be represented in terms of a permutation, which is then compared to the ground truth permutation π.

3.2. Listwise Depth Ranking

We model information about rankings in a *probabilistic* way, which has several advantages, especially from a learning point of view (for example, it makes the problem amenable to general inference principles such as maximum likelihood estimation). A well-known probability model on rankings is the Plackett-Luce (PL) model, which is param-

eterized by a vector $\boldsymbol{v} = (v_1, \ldots, v_K) \in \mathbb{R}^K_+$, where K is the number of items (length of the ranking). Referring to the interpretation of a ranking in terms of a preferential order, the value v_i is also called the (latent) utility of the ith item — subsequently, we shall use the more neutral notion of PL score or parameter. The probability of a permutation π of $[K]$ is then given by

$$P(\pi \mid \boldsymbol{v}) = \prod_{i=1}^{K-1} \frac{v_{\pi(i)}}{\sum_{k=i}^{K} v_{\pi(k)}}, \qquad (1)$$

where $\pi(i)$ is the index of the item on the ith rank. One easily verifies that, the larger the score v_i, the higher the probability that the ith item will show up on a top rank. Moreover, the mode of the distribution, i.e., the ranking with the highest probability, is obtained by sorting the items in decreasing order of their scores.

The PL model has the appealing property that each marginal of a PL model is again a PL model (with the same parameters). More specifically, if $J = \{j_1, \ldots, j_k\} \subseteq [K]$ is a subset of the K items, then the corresponding marginal of (1) is a PL model with parameters $v_{j_1}, \ldots, v_{j_k}$. This property greatly facilitates learning and inference from possibly incomplete rankings that do not comprise all K items. In fact, learning to rank with the PL model essentially comes down to estimating the score vector $\boldsymbol{v} = (v_1, \ldots, v_K)$.

In the case of depth estimation, items correspond to the pixels of an image, and the task of the learner is to predict the scores of these pixels. To make this possible, we assume that the score of a pixel can be expressed as a function of its context on the image. Thus, a parameter v_i is defined through a function $\phi_i : \mathcal{X} \longrightarrow \mathbb{R}$ on an input space $\mathcal{X}$ [8], where $\mathcal{X} = \mathbb{R}^{h \times w \times 3}$ corresponds to the space of all possible images of size $h \times w$. Assuming all images to have the same size, we set the overall number of alternatives K to $h \times w$.

In the domain of depth estimation, the most obvious way to represent the functions $\phi_1, \ldots, \phi_K$ is to model them as a (joint) deep convolutional neural network. Thus, each function ϕ_i is represented in terms of a set of network parameters $\boldsymbol{w}_i$, a subset of the parameters $\boldsymbol{w}$ of the entire (joint) network. In the experimental section, different state-of-the-art model architectures will be assessed for that purpose.

For an image $\boldsymbol{x} \in \mathcal{X}$, let $\boldsymbol{w}(\boldsymbol{x})$ denote the output of the neural network under parameterization $\boldsymbol{w}$ and

$$(v_1, \ldots, v_K) = (\phi_1(\boldsymbol{x}), \ldots, \phi_K(\boldsymbol{x})) = \exp(\boldsymbol{w}(\boldsymbol{x})) \quad (2)$$

the induced (non-negative) PL parameters. Thus, the entire PL model for the image $\boldsymbol{x}$ is eventually specified by the network parameters $\boldsymbol{w}$. Given a ranking π of (a subset of)

the pixels of $\boldsymbol{x}$ as training information, one can thus determine the probability $P(\pi \mid \boldsymbol{x}, \boldsymbol{w})$ of that ranking under $\boldsymbol{w}$ according to (1). More generally, given training information in the form of a collection of images with rankings, $\{(\boldsymbol{x}_i, \pi_i)\}_{i=1}^{L}$, learning an optimal model can be realized as maximum likelihood estimation [32]:

$$\boldsymbol{w}^* \in \arg\min_{\boldsymbol{w}} -\sum_{i=1}^{L} \log P(\pi \mid \boldsymbol{x}, \boldsymbol{w}). \qquad (3)$$

3.3. Metric Depth Estimation

Going beyond the prediction of rankings, one may wonder whether there is any possibility to recover metric depth information from a learned PL model. At first sight, this would be surprising, because the model is only trained on qualitative information in the form of rankings, and predicts probabilities instead of metric depth. Yet, the PL model also comprises a quantitative part, namely the scores v_i, which, as will be explained in the following, are in direct correspondence with the underlying metric information.

The PL model is a specific random utility model (RUM) [23]. In this class of models, it is assumed that the true order $z_1 < z_2 < \ldots < z_n$ of n real numbers — think of them as the true depth values of the pixels in a image — is "randomized" through (independent) measurement noise: Each value z_i is replaced by the measurement $X_i = z_i + \epsilon_i$, where ϵ_i is an error term, and what is observed as a ranking is the order of the measurements $X_1, \ldots, X_n$. In particular, the true order relation $z_i < z_j$ between two items is reversed if the corresponding error terms satisfy $\epsilon_i - \epsilon_j > z_j - z_i$, and the smaller the distance $|z_i - z_j|$, the more likely such a mistake is going to happen. Thus, the probability of a ranking error is indicative of the distance between z_i and z_j.

The PL model is obtained for the special case where the error terms ϵ_i follow a Gumbel distribution with fixed shape parameter [27]. More specifically, the so-called Thurstone model with parameters $z_1, \ldots, z_n$ is equivalent to the PL model (1) with parameters $v_i = \exp(z_i)$, $i = 1, \ldots, n$. In the context of depth estimation, the model can thus be interpreted as follows: The true depth of the ith image object (pixel) is given by z_i, but due to measurement noise, these distances are not observed precisely. Accepting the assumption of a Gumbel distribution[1], a PL model fitted to the observed (noisy) rankings of image objects yields estimates $\hat{v}_i$ of $v_i = \exp(z_i)$. Thus, a natural estimate of the underlying metric depth is given by $\hat{z}_i = \log(\hat{v}_i)$.

[1]This distribution looks similar to the normal distribution. Even if not provably correct, it is certainly not implausible.

We note that, since the PL model (1) is invariant toward multiplicative scaling (i.e., $P(\pi \mid \boldsymbol{v}) \equiv P(\pi \mid \lambda \boldsymbol{v})$ for $\lambda > 0$), the parameter $\boldsymbol{v}$ can only be determined up to a multiplicative factor. Correspondingly, the parameter $\boldsymbol{z}$ can only be determined up to an additive constant. This is indeed plausible: Assuming that the probability of reversing the order of two image objects only depends on their true distance $|z_i - z_j|$, this probability will not change by shifting the entire scene (i.e., moving the camera closer or farther away). In addition to this shift invariance, there is also a scaling effect, albeit of a more indirect nature. This effect is caused by fixing the shape parameter of the Thurstone model to 1. Therefore, instead of a simple log-transformation, we shall use an affine transformation of the form $\hat{z} = s \log(\hat{v}) + t$, with $s, t \in \mathbb{R}$ fitted to the image at hand.

3.4. Model

Regarding the underlying neural network, taking an image $\boldsymbol{x}$ as input and producing $\boldsymbol{w}(\boldsymbol{x})$ as used in (2) as output, we suggest two variants of our listwise ranking approach. The first one, dubbed *PLDepthResNet*, uses the same model architecture as suggested by Xian et al. [33]. As a second model, by consideration of recent neural architecture research, we propose *PLDepthEffNet* as a closely related architecture relying on EfficientNet [29] as backbone. Without further notice, the variant EfficientNetB5 is used as encoder, while the decoder part is a stack of repeating convolutional, BatchNormalization, ReLU and bilinear upsampling layers until the original shape is recovered. Similar to the model in [33], different scale features from the encoder branch are fed into the corresponding levels of the decoder part. Instead of fusing these features by addition, we concatenate at the respective layers. As a result, we obtain a model with approximately 45 million parameters for PLDepthEffNet, which is similar to the size of PLDepthResNet with 42 million parameters, while increasing the model performance at the same time (cf. empirical evaluation).

For both PLDepthResNet and PLDepthEffNet, we use encoders pretrained on ImageNet. Consequently, we standardize input images to match the preprocessing on ImageNet. During training, we freeze the encoder part and only allow the BatchNormalization layers to adjust to the new input data as typically done in transfer learning.

3.5. Sampling

In the past, different strategies to construct pairwise relations from raw depth data have been proposed, including superpixel sampling [35], random sampling [5], and combinations of multiple structure-guided strategies [34]. Ac-

cording to Xian et al. [34], random sampling of pairwise relations from raw depth data may harm the model's performance, due to training on uninformative or even misleading examples. Even worse, due to imprecision in the ground truth data, the risk of incorrectly ordered items increases with larger samples.

To address these issues, we propose a random sampling strategy that is almost as simple as pure random sampling, and which allows for incorporating the depth structure of the given image while leading to a relatively low training complexity. For R n-ary rankings to be queried per training tuple (I, D), $N \cdot R$ item sets M with n individual image locations are sampled, where $N > 1$ is a parameter. For each ranking set M, we order all image locations l by $D[l]$ to construct a ground truth permutation π. Given π, we sum up all pairwise depth differences $|D[l_{\pi(i)}] - D[l_{\pi(i+1)}]|$, $i \in [n-1]$. Afterwards, we sort all $N \cdot R$ rankings per image in a decreasing order according to this sum of depth difference and select the top R rankings as training examples. This way, we consider those rankings that seem to be most informative, since their relative depth values are maximized among the samples. Other strategies, such as the minimum among all pairwise depth differences in a ranking, are of course also possible as a proxy of the amount of information.

It is worth to mention that the Plackett-Luce model does not support partial rankings, i.e., neither allows for ties nor incomparability between items. Thus, as opposed to strategies incorporating equality relations, as e.g. [5], such relations are not explicitly considered here. To avoid sampling point pairs that are almost equally far away from the camera, we add a penalty of -10 to the depth difference sum for each compared image location pair l_1 and l_2 if their depth difference is such that $\max\left\{\frac{D[l_1]}{D[l_2]}, \frac{D[l_2]}{D[l_1]}\right\} < 1 + \tau$, where the parameter τ is set to $\tau = 0.03$ in our experiments.

4. Experiments

To demonstrate the effectiveness of our method, we conduct an exhaustive empirical evaluation on several benchmark datasets. Before presenting the results, we first introduce the datasets, followed by a brief description of the baseline methods and metrics used for assessment.

4.1. Datasets

To train our models, we use the recently introduced pseudo-metric "High-Resolution Web Stereo Image" (*HR-WSI*) dataset [34]. It consists of $20,378$ diverse, high resolution images annotated with pseudo-depth maps generated from flow predictions. For hyperparameter optimization,

a separate set of 400 images was used. Since the flow predictions provided as depth annotation failed for some image regions, a consistency mask is attached to each prediction to allow for sampling only from pixels that provide a reasonable depth value. To this end, a forward-backward flow consistency check has been applied. Furthermore, the annotations have been preprocessed to also assign a constant depth value to sky regions. Despite its relatively small size, we found this dataset to provide highly informative image and depth pairs to learn from.

In the experiments, we compare our model to various baselines in a "zero-shot" generalization study on datasets that were not used within the training processes. Thus, we follow the basic evaluation scheme by Ranftl et al. [25]. As datasets, we consider *Ibims* [16], *Sintel* [3], *DIODE* [30], and *TUM* [28]. In the supplementary material, we detail the characteristics of each dataset, such as their data diversity. With this choice of benchmark targets, we capture indoor, outdoor, and computer generated scenes, which provides a good basis for assessing the generalization performance of different models, and their ability to predict depth orders in a wide variety of applications.

4.2. Baselines

We compare our PL-based approach to state-of-the-art depth estimation models using depth relations as training information. To this end, we consider the ResNet-based model trained on "Relative Depth from Web" (ReDWeb), "Depth in the Wild" (DIW), and YouTube3D as described by Chen et al. [6], hereinafter referred to as YouTube3D, and the same model as used by Xian et al. [34] trained on HR-WSI (referred to as Xian 2020). Both approaches have shown compelling generalization performance, corroborating our motivation to use relative data for supervision.

Besides models trained on relative depth information, regression models are obviously also capable of inferring rankings, simply by sorting the image locations based on their values in a predicted dense depth map. Therefore, we consider state-of-the-art (pseudo-)regression methods as additional baselines, namely, DenseDepth [1], BTS [18], MegaDepth [20], MannequinChallenge (MC) [19], and MiDaS [25]. Furthermore, we also evaluated MonoDepth2 [14] as a completely unsupervised resp. self-supervised method.

While we considered most baselines as described in the related work, let us note that the authors of MiDaS provide a model trained on approximately 2 million examples, which is far more than most of the other methods we compare with. To account for this, we re-implemented their approach and

retrained the model on HR-WSI for a fairer comparison. For a complete overview of all baselines, including a categorization of the respective training data diversity, we refer to the supplementary material.

4.3. Metrics

To evaluate our models, we report the "ordinal error" on sampled point pairs as done by Xian et al. [34]. For two points l_1 and l_2 sampled from an example (I, D), with I being the image and D a dense (pseudo-)depth map as specified before, the ground truth ordinal relation $r(l_1, l_2, D)$ is given by $+1$ for $D[l_1] > D[l_2]$, -1 for $D[l_2] > D[l_1]$ and 0 otherwise. The ordinal error is then given by

$$ord(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(I, D, l_1, l_2) \in \mathcal{D}} \mathbb{1}\big(r(l_1, l_2, D) \neq r(l_1, l_2, f(I))\big),$$

$$(4)$$

where f is the function predicting depth or, in the case of a PL model, scores for each pixel of the input image I, resulting in a dense depth map just as given by D, and $\mathcal{D}$ denotes the set of all point pairs sampled from the test dataset images and depth maps.

As already noted, we omit all equal pairs, i.e., relations with $r(\cdot, \cdot, \cdot) = 0$. Hence, we report ord on unequal pairs only without any equality thresholding. Thus, there is no need to rely on re-scaling and -translating as done in [25] and [34] to identify reasonable equality thresholds, which comes with additional complications for the evaluation process.

Often, depth orders have varying priorities, i.e., closer elements are more critical for correct ordering than elements far away from the camera. For example, an autonomous vehicle has less time to react to elements very close to the car and must rely on valid input for safe interactions. This is reflected by metrics like the discounted cumulative gain (DCG), which measures the usefulness of rankings by accumulating graded relevances of ranking items discounted with decreasing rank. More precisely, for every image location l associated with a dense depth map D, we set the relevance score of l in D to $rel(l, D) = \frac{1}{D[l]+1}$. Given these scores, we can specify the DCG score for a ranking $l_{\pi(1)} \succ l_{\pi(2)} \succ \cdots \succ l_{\pi(n)}$ by

$$DCG(\pi, D) = \sum_{i=1}^{n} \frac{rel(l_{\pi(i)}, D)}{\log_2(i+1)}. \tag{5}$$

For our experiments, we used the normalized DCG (nDCG), which divides (5) by the best DCG possible on D.

For the metric comparison, we assess the root-mean-square error (RMSE) between the dense ground truth and predicted depth maps and the percentage of predictions $\hat{z}$ such that $\max\left(\frac{\hat{z}}{z}, \frac{z}{\hat{z}}\right) = \delta > 1.25$ for the ground truth depth z. To calculate the metrics, we normalized the given ground truth scores by the maximum depth capacity of the corresponding dataset (cf. the dataset characteristics in the supplement) to obtain error values on a similar scale.

4.4. Results

To show the effectiveness of our method proposal, we first compare different losses using the same model architecture and training dataset, followed by a comparison of our method to the baselines. Every reported result is the average of three runs with different randomization seeds.

4.4.1 Loss Comparison

There are many experimental studies in the literature showing improved performance of a method, but not isolating the key factors contributing to the improvement, e.g., the neural network architecture, loss function, training procedure, training data, etc. To assess the influence of a listwise approach to ranking more clearly, we evaluate three methods trained on the same data and with the same neural network architecture, namely (scale-invariant, SI) regression, pairwise, and listwise ranking. It is true that the model, loss, and data may strongly interact with each other (i.e., a loss might work well with a certain architecture on a particular dataset, while the same architecture may harm the performance of a different method). Nevertheless, we found that the ResNet-based architecture as proposed by Xian et al. [33] and subsequently also used in [25] serves as a good basis for a fair comparison.

For our experiments, we re-implemented the SI mean-squared error loss as also used in MiDaS and the pairwise ranking loss as described in [5] and [33]. As training information, we used HR-WSI as a state-of-the-art diverse pseudo-depth dataset. We refer to the supplement for a detailed description of all hyperparameters.

All three methods require different sampling strategies: While the SI-regression uses the complete (masked) image, pair- and listwise methods involve different amounts of sampled points selected per ranking. For a fair comparison, we adopted the number of sampled rankings in the listwise case to the number of drawn pairwise relations, such that one approach does not see much more points than the other during training. In the case of pairwise rankings, we randomly sampled 1k point pairs per image and epoch, resulting in a maximum of 2k seen points per image and epoch. For our listwise approach, we found a size of 5 to achieve a good trade-off between highly informative rankings and effi-

Table 1. Ordinal errors on 50k randomly sampled pairs per loss, using the architecture from [34] trained on HR-WSI (lower is better).

Loss	Ibims	Sintel	DIODE	TUM	Avg. Rank
SI-Regression	0.308	0.311	0.334	0.222	3
Pairwise	0.281	0.299	0.291	**0.192**	1.75
Listwise	**0.273**	**0.289**	**0.285**	0.218	**1.25**

cient training. Hence, we sampled 400 rankings of ranking size 5 per image and epoch. Here, we explicitly stick to random-only sampling to alleviate side effects.

Table 1 presents the results of the method comparison on 50k randomly sampled location pairs per image. As can be seen, the relative models outperform the SI-regression method, suggesting to serve as a better surrogate loss for optimizing the ordinal error. Moreover, our listwise approach seems to perform slightly better than the pairwise approach, although the difference does not appear to be significant.

4.4.2 Ordinal Prediction

After having compared the loss function on a shared model and data level, we now analyze individual depth estimation models with regard to their ordinal error and nDCG performance as trained by the respective authors, who made an attempt at optimizing the interplay between data, network architectures, and training procedures.

For the baseline models, we used the best provided pre-trained models by the authors or, if official implementations were not available, by popular and carefully tested re-implementations. For our PL models, we kept most of the training hyperparameters the same (see supplementary for more details). Within our sampling strategy, we set the factor $N = 5$ (cf. Section 3.5). For MiDaS, we also used our proposed EfficientNet-based architecture, which delivers superior performance compared to the formerly used architecture, for reasons of fairness. Here, as opposed to the version of MiDaS within the loss comparison, where we primary focused on comparing different problem considerations, we employ the trimmed absolute deviation loss providing the best performance among the regarded alternatives (cf. [25]).

Table 2 reports the individual ordinal errors on unequal relations for the four benchmark datasets, again on 50k randomly sampled location pairs per image. As can be seen, our PLDepthEffNet achieves the lowest averaged rank over all datasets, while outperforming the other methods on half of the datasets at the same time, demonstrating the effectiveness of the listwise ranking approach to optimize the ordinal error metric. Supporting the observations made in the previous experiment, the generalization capabilities of MegaDepth

Table 2. Ordinal errors on benchmark datasets with 50k randomly sampled relations for each image (lower is better).

Model	Ibims	Sintel	DIODE	TUM	Avg. Rank
DenseDepth	0.208	0.384	0.317	0.224	5.75
MegaDepth	0.297	0.324	0.316	0.227	7.5
BTS	**0.190**	0.384	0.323	0.251	6.25
MC	0.272	0.387	0.378	0.206	7.25
MiDaS	0.269	0.278	0.263	0.207	3.75
MonoDepth2	0.375	0.425	0.407	0.336	9.75
YouTube3D	0.272	0.292	0.288	0.199	4.75
Xian 2020	0.225	0.278	0.263	**0.184**	2.25
PLDepthResNet	0.245	0.284	0.277	0.213	4.75
PLDepthEffNet	0.213	**0.272**	**0.256**	0.204	**2**

Table 3. nDCG on benchmark datasets with 100 randomly sampled rankings of size 500 for each image (higher is better).

Model	Ibims	Sintel	DIODE	TUM	Avg. Rank
DenseDepth	0.916	0.986	0.821	0.986	4.75
MegaDepth	0.911	0.989	0.815	0.983	7.5
BTS	**0.918**	0.986	0.825	0.983	4.75
MC	0.908	0.986	0.828	0.987	5.5
MiDaS	0.913	0.991	0.806	0.987	6.25
MonoDepth2	0.896	0.981	**0.836**	0.961	7.75
YouTube3D	0.911	0.993	0.816	0.988	4.75
Xian 2020	0.916	0.993	0.817	**0.990**	2.75
PLDepthResNet	0.914	0.993	0.817	0.985	5
PLDepthEffNet	0.916	**0.994**	0.819	0.988	**2.5**

as another scale-invariant regression method, even by having access to over 600k diverse instances, to correctly rank elements are fairly limited. Moreover, in agreement with the previous results, the ranking approaches are consistently among the best models, suggesting ranking losses to be the favorite choice as surrogates for ordinal error minimization.

Additionally, Table 3 reports the results for nDCG as performance metric on 100 randomly sampled rankings of size 500 per image. In accordance with the ordinal errors, ranking methods are well suited to optimize this metric. Here, the top-3 models are all of that kind, with PLDepthEffNet slightly better performing than Xian 2020.

4.4.3 Metric Prediction

As motivated theoretically in Section 3.3, our method provides an interface to recover metric depth information approximated from observed rankings. Here, we compare our model to the baselines with regard to the two metric error measures RMSE and $\delta > 1.25$ using the same models as in Section 4.4.2. As all benchmark datasets have different scales and might be shifted arbitrarily, we rescale and shift the predictions to the resolution of the ground truth as de-

Table 4. Evaluation results on benchmark datasets with regard to metric depth error measures (lower is better in both cases).

| Model	Ibims		Sintel		DIODE		TUM		Avg. Rank	
	RMSE	$\delta > 1.25$	RMSE	$\delta > 1.25$	RMSE	$\delta > 1.25$	RMSE	$\delta > 1.25$	RMSE	$\delta > 1.25$
DenseDepth	**0.016**	20.9	0.128	39.6	0.110	53.5	0.084	69.7	5.25	4.5
MegaDepth	0.020	35.9	0.119	35.5	0.094	55.3	0.082	70.8	6	7
BTS	**0.016**	**18.9**	0.133	41.8	0.112	54.4	0.089	72.4	7	6.25
MC	0.018	31.3	0.128	38.8	0.120	58.7	**0.074**	**67.8**	5.25	5.5
MiDaS	0.019	33.2	**0.091**	**27.7**	**0.081**	53.5	0.085	71.1	4	4.75
MonoDepth2	0.023	42.6	0.143	43.8	0.122	61.1	0.088	72.5	9.75	10
YouTube3D	0.019	31.8	0.101	31.1	0.096	54.5	0.077	68.4	4.75	5.25
Xian 2020	0.018	31.5	0.096	30.5	0.085	**51.4**	0.080	69.4	**3**	**3.25**
PLDepthResNet	0.019	30.9	0.099	30.7	0.092	53.1	0.084	71.9	5	4.75
PLDepthEffNet	0.017	29.1	0.093	29.3	0.085	52.7	0.083	71.6	**3**	3.5
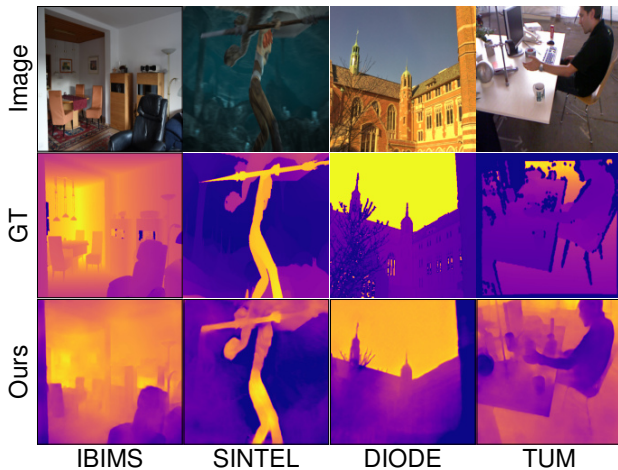
Figure 2. Sample predictions given by the reconstructed metric scores of the PLDepthEffNet model as used in the experiments.

scribed in [25] by optimizing a least-squares criterion.

The results are given in Table 4. As can be seen, although our model was solely trained on rankings, it is capable of recovering the underlying depth structure relatively precisely. Noteworthy, it is superior to all regression baselines and on a par with Xian 2020 for RMSE, although this ranking baseline additionally incorporates a smooth gradient loss term for sharp boundaries, directly accessing the metric depth information at training time. While it delivers the highest $\delta > 1.25$ accuracy, our approach still proves to be very competitive in this regard.

Fig. 2 shows exemplary predictions of our model. Obviously, the model is able to capture tiniest object details, such as tree branches in the image from DIODE, and predicting sharp object boundaries. This shows that, even with simple sampling strategies, listwise ranking is able to reflect and predict such small details, without any need for very complex strategies based on the depth structure of an image.

5. Conclusion

We have proposed to tackle the problem of depth ordering in images as a listwise ranking problem, for which we employed a Plackett-Luce model tailored to the domain of monocular depth estimation. Thus, compared to estimating the exact depth values, we solve an arguably simpler problem, at least if the goal is to minimize an ordinal error metric. Besides, compared to precise numerical data required by regression models for training, a ranking approach allows for leveraging weaker and more diverse training data. Although not directly trained on metric data, our model is capable of providing precise (shift-invariant) depth predictions, essentially by exploiting the relationship between the (latent) distance between image objects and the probability of reversing their order in a ranking.

Through an exhaustive zero-shot cross-dataset evaluation, we showed that our approach, combined with a state-of-the-art neural network as backbone, achieves superior ranking performance compared to previous approaches. In particular, it improves upon existing pairwise ranking methods, in spite of using a much simpler and more efficient sampling technique. Remarkably, our model also performs very competitive on metric error measures.

Motivated by these promising results, we plan to elaborate on further improvements of the listwise ranking approach. This includes an investigation of the effect of varying the ranking size, as well as an extension toward learning from partial rankings and incorporating equality relations. In addition, as we only applied random sampling so far, we plan to develop more sophisticated sampling strategies leading to more informative rankings to learn from.

References

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *CoRR*, abs/1812.11941, 2018.

[2] Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. Learning to rank using gradient descent. In Luc De Raedt and Stefan Wrobel, editors, *Proceedings of the 22nd International Conference on Machine Learning, (ICML), August 7-11, 2005, Bonn, Germany*, volume 119 of *ACM International Conference Proceeding Series*, pages 89–96. ACM, 2005.

[3] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Proceedings of the 12th European Conference on Computer Vision (ECCV), Part VI, October 7-13, 2012, Florence, Italy*, volume 7577 of *Lecture Notes in Computer Science*, pages 611–625. Springer, 2012.

[4] Yuanzhouhan Cao, Tianqi Zhao, Ke Xian, Chunhua Shen, Zhiguo Cao, and Shugong Xu. Monocular depth estimation with augmented ordinal depth relationships. *IEEE Trans. Circuits Syst. Video Technol.*, 30(8):2674–2682, 2020.

[5] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, December 5-10, 2016, Barcelona, Spain*, pages 730–738, 2016.

[6] Weifeng Chen, Shengyi Qian, and Jia Deng. Learning single-image depth from videos using quality assessment networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 16-20, 2019, Long Beach, CA, USA*, pages 5604–5613. Computer Vision Foundation / IEEE, 2019.

[7] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. OASIS: A large-scale dataset for single image 3d in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA*, pages 676–685. IEEE, 2020.

[8] Weiwei Cheng, Krzysztof Dembczynski, and Eyke Hüllermeier. Label ranking methods based on the Plackett-Luce model. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning ICML, June 21-24, 2010, Haifa, Israel*, pages 215–222. Omnipress, 2010.

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems, December 8-13, 2014, Montreal, Quebec, Canada*, pages 2366–2374, 2014.

[10] Ralph Ewerth, Matthias Springstein, Eric Müller, Alexander Balz, Jan Gehlhaar, Tolga Naziyok, Krzysztof Dembczynski, and Eyke Hüllermeier. Estimating relative depth in single images via RankBoost. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), July 10-14, 2017, Hong Kong, China*, pages 919–924. IEEE Computer Society, 2017.

[11] Yoav Freund, Raj D. Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969, 2003.

[12] J. Fürnkranz and E. Hüllermeier, editors. *Preference Learning*. Springer-Verlag, 2011.

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robotics Res.*, 32(11):1231–1237, 2013.

[14] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 27 - November 2, 2019, Seoul, Korea (South)*, pages 3827–3837. IEEE, 2019.

[15] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pages 133–142. ACM, 2002.

[16] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of CNN-based single-image depth estimation methods. In Laura Leal-Taixé and Stefan Roth, editors, *Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Part III, September 8-14, 2018, Munich, Germany*, volume 11131 of *Lecture Notes in Computer Science*, pages 331–348. Springer, 2018.

[17] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *Proceedings of the 4th International Conference on 3D Vision (3DV), October 25-28, 2016, Stanford, CA, USA*, pages 239–248. IEEE Computer Society, 2016.

[18] 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. *CoRR*, abs/1907.10326, 2019.

[19] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Learning the depths of moving people by watching frozen people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 16-20, 2019, Long Beach, CA, USA*, pages 4521–4530. Computer Vision Foundation / IEEE, 2019.

[20] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-

view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 18-22, 2018, Salt Lake City, UT, USA*, pages 2041–2050. IEEE Computer Society, 2018.

[21] Tie-Yan Liu. Learning to rank for information retrieval. In Fabio Crestani, Stéphane Marchand-Maillet, Hsin-Hsi Chen, Efthimis N. Efthimiadis, and Jacques Savoy, editors, *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), July 19-23, 2010, Geneva, Switzerland*, page 904. ACM, 2010.

[22] R. Duncan Luce. *Individual Choice Behavior: A Theoretical analysis*. Wiley, New York, NY, USA, 1959.

[23] D. Mcfadden. Econometric models for probabilistic choice among products. *The Journal of Business*, 53:13–29, 1980.

[24] R. Plackett. The analysis of permutations. *Journal of The Royal Statistical Society Series C-applied Statistics*, 24:193–202, 1975.

[25] 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.

[26] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Proceedings of the 12th European Conference on Computer Vision (ECCV), Part V, October 7-13, 2012, Florence, Italy*, volume 7576 of *Lecture Notes in Computer Science*, pages 746–760. Springer, 2012.

[27] Hossein Azari Soufiani, David C. Parkes, and Lirong Xia. Random utility theory for social choice. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 126–134, 2012.

[28] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 7-12, 2012, Vilamoura, Algarve, Portugal*, pages 573–580. IEEE, 2012.

[29] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning (ICML), 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019.

[30] Igor Vasiljevic, Nicholas I. Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A dense indoor and outdoor depth dataset. *CoRR*, abs/1908.00463, 2019.

[31] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *Proceedings of the 7th International Conference on 3D Vision (3DV), September 16-19, 2019, Québec City, QC, Canada*, pages 348–357. IEEE, 2019.

[32] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: Theory and algorithm. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Proceedings of the 25th International Conference on Machine Learning ICML, Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 1192–1199. ACM, 2008.

[33] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), June 18-22, 2018, Salt Lake City, UT, USA*, pages 311–320. IEEE Computer Society, 2018.

[34] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA*, pages 608–617. IEEE, 2020.

[35] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T. Freeman. Learning ordinal relationships for mid-level vision. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile*, pages 388–396. IEEE Computer Society, 2015.