DME: Unveiling the Bias for Better Generalized Monocular Depth Estimation

Songsong Yu, Yifan Wang, Yunzhi Zhuge, Lijun Wang*, Huchuan Lu

Dalian University of Technology
22209083@mail.dlut.edu.cn, {wyfan, ljwang, lhchuan}@dlut.edu.cn, zgyzzgyz@gmail.com

Abstract

This paper aims to design monocular depth estimation models with better generalization abilities. To this end, we have conducted quantitative analysis and discovered two important insights. First, the Simulation Correlation phenomenon, commonly seen in long-tailed classification problems, also exists in monocular depth estimation, indicating that the imbalanced depth distribution in training data may be the cause of limited generalization ability. Second, the imbalanced and long-tail distribution of depth values extends beyond the dataset scale, and also manifests within each individual image, further exacerbating the challenge of monocular depth estimation. Motivated by the above findings, we propose the Distance-aware Multi-Expert (DME) depth estimation model. Unlike prior methods that handle different depth range indiscriminately, DME adopts a divide-and-conquer philosophy where each expert is responsible for depth estimation of regions within a specific depth range. As such, the depth distribution seen by each expert is more uniform and can be more easily predicted. A pixel-level routing module is further designed and learned to stitch the prediction of all experts into the final depth map. Experiments show that DME achieves state-of-the-art performance on both NYU-Depth v2 and KITTI, and also delivers favorable zero-shot generalization capability on unseen datasets.

Introduction

Despite remarkable progress being achieved in recent years (Bhat, Alhashim, and Wonka 2021; Ranftl, Bochkovskiy, and Koltun 2021; Wang et al. 2021; Ren et al. 2022), monocular depth estimation still suffers from unsatisfactory generalization ability, which hinders their applicability in complex real-world scenarios. The lack of generalization issue is mostly attributed to either significant scene diversities or scale variations, and addressed by methods of two main kinds. First, scene-aware models incorporate multi-branch architectures to estimate depth tailored to different scenes (Ren, El-Khamy, and Lee 2019; Bhat et al. 2023). However, they require scene priors. Second, relative depth models train in a scale-invariant manner on diverse, large-scale data to enable metric depth prediction (Ranftl et al. 2020; Ranftl, Bochkovskiy, and Koltun 2021).

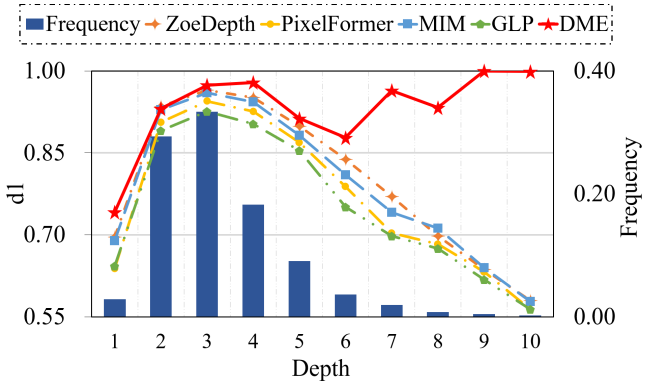

Figure 1: Simulation Correlation. The bar chart displays the frequency distribution of depth values in the NYUD v2 dataset. The line graph represents the model's d1 accuracy, with the values corresponding to the primary axis. It can be observed that the depth values exhibit a unimodal distribution, while the model's performance shows a positive correlation with the frequency variation.

Though achieving state-of-the-art performance, they cannot restore absolute depth values. While these approaches have advanced the field, their limitations highlight the need for methods that improve generalization without sacrificing metric accuracy or requiring scene priors.

In the literature of image classification, the generalization issue has been intensively explored and largely attributed to the *Simulation Correlation* effect (Hong et al. 2021) found in long-tail data distributions, *i.e.*, the classification model tends to overly focus on high-frequent categories with degraded performance on low frequent ones. In contrast, although the imbalanced distribution of monocular depth has been identified by prior works (Jiao et al. 2018), its connection to the generalization issue has been rarely explored. We thus make one of the first attempts by asking the questions of whether monocular depth also exhibits long-tailed distribution and whether the generalization issue can be explained by the *Simulation Correlation* effect.

To answer the aforementioned questions, we have conducted quantitative studies on depth distribution and its impact on depth estimation models. Two important findings has

*Corresponding author.

been made. First, we discover that monocular depth values of both indoor and outdoor scenes follow long-tail distributions, and identify the prevalence of the *Simulation Correlation* phenomenon in most of state-of-the-art depth estimation methods (*c.f.* Figure 1). Second, unlike the image classification task, the long-tail depth distribution is not only present at the dataset level, but can also be found in individual images.

Motivated by the above findings, we design a Distance-aware Multi-Expert model to enhance the generalization ability of depth estimation. We propose to divide the depth range into multi-intervals. Each expert is responsible for estimating depth within its corresponding interval. As the depth distribution in each interval is more balanced, the long-tailed issue is effectively alleviated, and the specifically learned experts are more focused, giving rise to improved depth precision. To achieve the final depth estimation results, a pixel-level routing module is designed which can automatically aggregate the outputs of the multi-experts without requiring any prior information, resulting in a comprehensive and accurate depth map.

In summary, the contribution of this paper is threefold:

- We conduct an in-depth analysis of widely-adopted depth estimation datasets, which indicates that the generalization issue may be caused by long-tailed depth distribution as well as the *Simulation Correlation* effect.
- We propose the Distance-aware Multi-Expert paradigm to enhance depth estimation, which can effectively address the long-tailed depth distribution issue, providing better generalization ability.
- Our method sets new state-of-the-art performance on NYUD v2 and KITTI datasets, and has shown superior zero-shot generalization performance on five unseen datasets of diverse scenarios.

Our work provides a new perspective to address the generalization issue in depth estimation. The code can be obtained at https://github.com/YUsong360/DME-Unveiling-the-bias.

Related Works

Scene-Aware Depth Estimation Model

Researchers (Ranftl et al. 2020) find that models face challenges in optimization when simultaneously training on indoor and outdoor datasets. It is currently understood that these challenges arise due to differences in camera parameters and dataset scales, making it difficult for models to learn effectively. Consequently, previous depth estimation models (Xie et al. 2023; Wang et al. 2020b; Bhat, Alhashim, and Wonka 2021; Lee et al. 2019a; He et al. 2023) are being trained and tested on a single dataset, resulting in limited generalization capabilities. ZoeDepth and DS-SIDENet (Ren, El-Khamy, and Lee 2019) introduce scene understanding modules to enable models to learn scene differentiation, achieving a divide-and-conquer effect (Wang et al. 2020a). Specifically, ZoeDepth employs indoor and outdoor heads to predict the NYUD v2 (Silberman et al. 2012) and KITTI (Geiger et al. 2013) datasets respectively, and additionally designs a scene discriminator to distinguish between indoor and outdoor images. DS-SIDENet proposes a two-stage model, where the first stage incorporates two different scene understanding modules based on scene classification and coarse depth estimation, while the second stage utilizes the DS-SIDENet trained on specific depth range images to obtain accurate depth maps. Our model takes ZoeDepth as the baseline, requiring similar routing mechanisms, but with a key distinction of employing pixel-level routing rather than simple image-based scene classification.

Long-Tail Phenomenon in Classification

Extensive and in-depth research is being conducted to address the issue of long-tail distribution in classification. When confronted with data imbalance, the most direct solution is resampling. SimCal (Wang et al. 2020c) proposes a two-level sampling strategy that combines image-level and instance-level resampling to alleviate class imbalance in instance segmentation. DCL (Wang et al. 2019) develops a novel curriculum strategy where the probability of subsequent sampling from a class decreases as the number of instances sampled from that class increases, dynamically rebalancing the class distribution. meta-softmax (Ren et al. 2020) introduces a meta-learning-based sampling approach that optimizes the model's classification performance on a balanced meta-validation set to learn the optimal sampling rates for different categories in long-tail learning.

Recently, numerous ensemble learning and decoupling methods have achieved significant progress in long-tailed recognition. MiSLAS (Zhong et al. 2021) enhances representation learning through data mixing and proposes a label-aware smoothing strategy to improve model generalization. BBN (Zhou et al. 2020) proposes the use of two network branches, namely the conventional learning branch and the re-balancing branch, for handling long-tail recognition tasks. RIDE (Wang et al. 2020d) trains independent softmax losses for each expert and introduces a diversity-promoting loss based on KL divergence to increase the diversity among different experts. TADE (Zhang et al. 2021) develops a new multi-expert framework and innovates expert training schemes by introducing domain-specific knowledge-guided losses to promote diversity in handling different class distributions.

Our approach aligns more with ensemble learning methods such as RIDE and TADE, but we focus on reasonable grouping based on continuous depth space, distinguishing our work from these discrete classification problems.

Long-Tail Distribution in Depth Estimation

Depth estimation requires obtaining specific depth values for each pixel. Since it is a task with continuous values and no concept of categories, there is limited discussion and research on the phenomenon of data imbalance in depth estimation tasks over the years. Attention-Driven Loss (Jiao et al. 2018) conducts a statistical analysis of the frequency distribution of depth in the NYUD v2 and KITTI datasets. They design a reweighted loss function targeting the depth range of distant regions to enhance the model's focus on

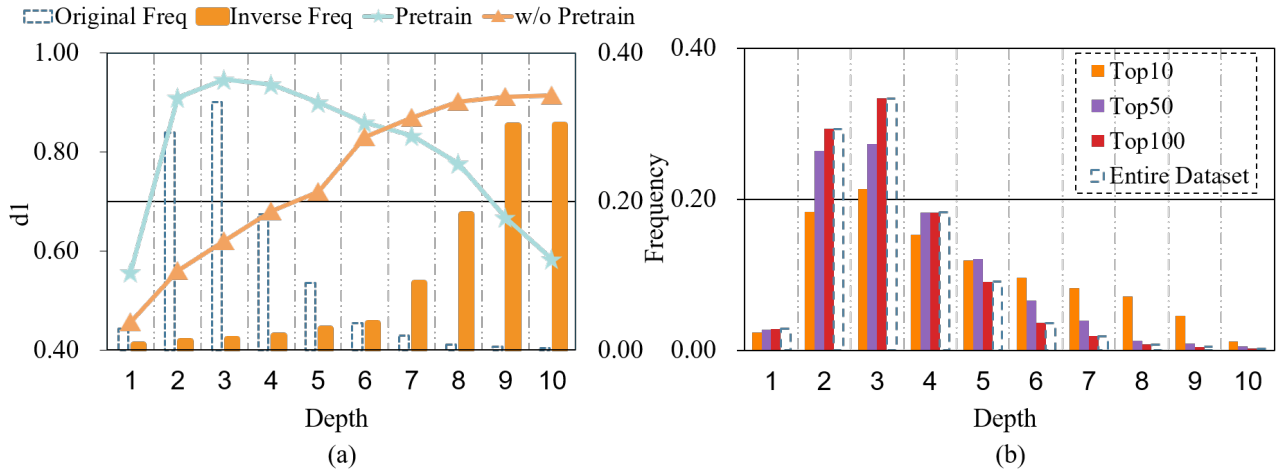

Figure 2: (a) Inverse frequency training. Both initialization methods were trained using inverse frequency as the supervision signal. (b) Depth distribution. The rarity scores were calculated and the samples were sorted in descending order. 'Total' represents the depth frequency distribution of the entire dataset. The top 10, top 50, and top 100 samples are selected for statistical analysis, revealing a unimodal frequency distribution.

depth values of underrepresented classes and improve performance during training. However, their research primarily focuses on distant depth values and does not provide a detailed and comprehensive analysis of the frequency distribution in depth estimation tasks.

Taking the frequency distribution into account, we conduct a thorough investigation, demonstrating that the long-tail distribution of depth values is a pixel-level problem. We also validate the specific impact of this imbalanced distribution on the model's performance. Additionally, we conduct experiments to test common approaches for addressing data imbalance and propose a more effective solution.

Method

Analysis on Monocular Depth Distribution

The *Simulation Correlation* effect (Hong et al. 2021) in long-tail classification tasks implies that data skewness can introduce undesirable bias to the model, thereby affecting its generalization capability. To investigate whether a similar phenomenon also exists in monocular depth estimation, we conduct quantitative analysis on both indoor (NYUD v2 (Silberman et al. 2012)) and outdoor (KITTI (Geiger et al. 2013)) datasets to study the depth distribution and its impact on depth estimation performance. Since the major findings are mostly consistent, we report the detailed results on NYUD v2 in the following.

Interval-Wise Evaluation: Unlike classification, depth estimation as a regression problem has continuous output space. To leverage a similar analysis approach, we evenly divide the depth values of the NYUD v2 dataset into 10 intervals and calculate their frequencies. Interval-wise evaluations are then conducted on a series of recent state-of-the-art methods, including GLP (Kim et al. 2022), MIM (Xie et al. 2023), ZoeDepth (Bhat et al. 2023), and PixelFormer (Agarwal and Arora 2023). Among them, GLP and MIM are single-stage regression models, while ZoeDepth and PixelFormer are two-stage classification-regression models. The depth distribution and evaluation results are shown in Figure 1. It confirms that the depth distribution of entire dataset indeed follows a unimodal long-tail distribution, with most of the depth values in the range of 2-5 meters. Meanwhile, we can also observe a strong positive correlation between the depth frequency and the performance of all the compared methods. This may be caused by the *Simulation Correlation* effect, but may also be simply attributed to the fact that depth estimation at long distances is inherently more challenging (Jiao et al. 2018).

Inverse-Frequency Training: To further confirm the *Simulation Correlation* effect, we analyze the impact of inverse-frequency training on ZoeDepth. To this end, we first invert the frequency of training depth values through resampling, transforming the dominant depth intervals into minority ones and vice versa. The inverse as well as the original training depth frequency are shown in Figure 2 (a). We then train two variants of ZoeDepth on the inverse-frequency data: 'Pretrain' denotes the one initialized from large-scale pre-trained parameters (Bhat et al. 2023), and 'w/o Pretrain' is trained from scratch. By comparing Figure 1 and 2 (a), we can conclude that the depth estimation performance is mostly determined by the training data distribution. The performance of the 'Pretrain' variant further suggests that large-scale pretraining on relative depth can improve generalization across depth intervals to a certain extent. However, the improvement is still limited in remote regions due to the long-tailed distribution of pretraining data. The above results and analysis confirms that the weak generalization issue of depth estimation can also be explained by the *Simulation Correlation* effect.

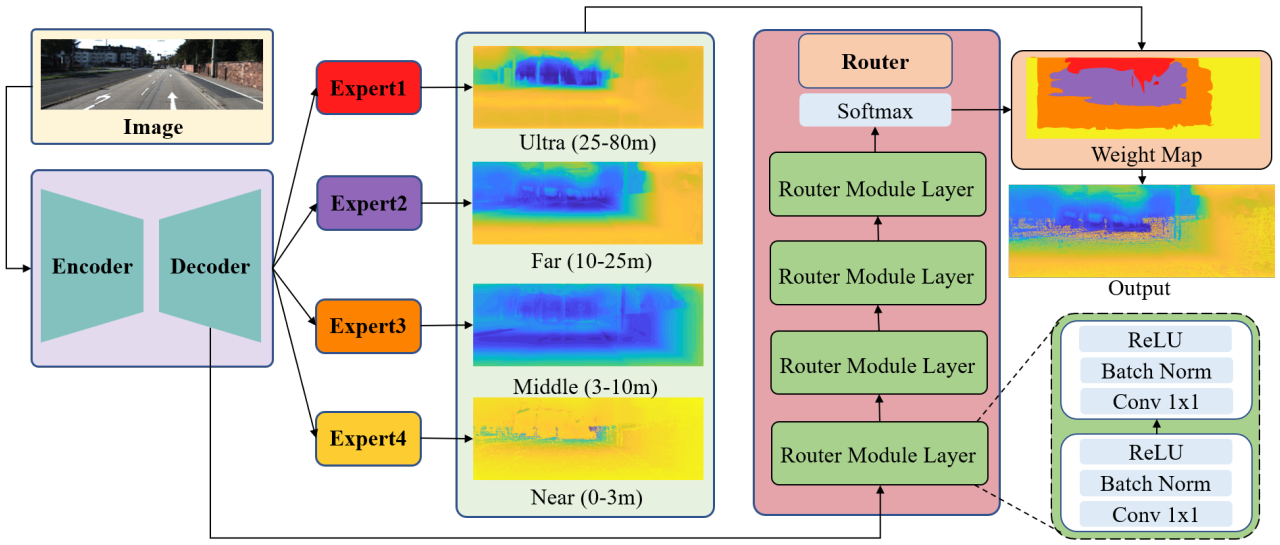

Figure 3: DME architecture. The RGB image is processed through an encoder-decoder framework, yielding depth estimation results for four distance ranges. Each expert is responsible for performing depth estimation within a different distance range: near, middle, far, and ultra distances. The decoder's features are connected to a routing module, where confidence scores for each distance range are computed. Finally, the depth estimation results at these distances are linearly combined to obtain the ultimate depth estimation.

Analysis on Image-level Depth Distribution: As depth estimation involves dense prediction, an intriguing question arises: Does the long-tailed depth distribution persist at the image level as well? To investigate this, our basic idea is to identify the most uncommon image samples in the dataset, whose depth distributions significantly diverge from the overall depth distribution of the entire dataset. If these rare samples also display a similar long-tailed and unimodal pattern in their depth distribution, it suggests that the long-tailed property is a phenomenon observed at the image level. To this purpose, we define the rarity score S of an image as its KL divergence to dataset:

$$S = \sum_i p_i \log \frac{p_i}{q_i}, \tag{1}$$

where p_i and q_i denotes the frequency of the i-th depth interval calculated from the entire dataset and the image, respectively. We then sort all the images based on their rarity scores in descending order and select the top-K images to recalculate their depth distributions. As shown in Figure 2 (b), the depth distributions for $K = 10, 50, 100$ exhibit striking similarities and successively converge towards that of the entire dataset. This observation suggests that the long-tailed depth distribution is a prevalent characteristic at the image level. This is actually easy to comprehend when considering the nature of perspective projection, where closer regions tend to dominate in the resulting images, leading to unbalanced depth distribution.

Distance-Aware Multi-Expert Model

In light of the above findings, we conjecture that the generalization issue in depth estimation may be caused by multiple factors, including the imbalanced depth distribution present in both training datasets and individual monocular images, as well as the influence of the *Simulation Correlation* effect. To verify this, we propose the Distance-aware Multi-Expert (DME) depth estimation paradigm. Our major insight is to divide the overall depth range into multiple intervals handled by specific experts. The long-tailed pattern of depth distribution in each interval will significantly diminish. As such, the difficulty of depth estimation for each depth interval will be effectively alleviated. Figure 3 present an overview of the architecture. Given an input image, we adopt an encoder-decoder structure (Touvron et al. 2021) to extract multi-scale features. The experts are convolutional sub-networks built upon the extracted features and responsible to predict depth within specific intervals. A pixel-wise routing module is further designed, which aggregates the output of all experts into the final depth estimation results. Though conceptually simple, our method motivated by quantitative analysis has shown superior performance in our experiments.

As opposed to the uniform depth partition, we empirically divide the depth range based on observations. We consider NYUD v2 and KITTI datasets as they contain both indoor and outdoor scenes and are thus more representative. To ensure depth distribution of each interval to be equalization, we divide the depth range into 4 intervals, including 1-3 meters, 3-10 meters, 10-25 meters, and 25-80 meters. As shown in our experiments, this empirical partition approach can well generalize to even unseen datasets.

Pixel-Level Routing Module We further design a lightweight routing module to automatically combine the output of all the experts into the final depth estimation results. Different from the image-level routing method in (Bhat et al. 2023), our proposed routing module operates

at the pixel-level. As shown in Figure 3, The input features from the multi-scale hierarchy are firstly projected into C channels via 1×1 convolutions and then upsampled to the original image resolution $H \times W$ through consecutive bilinear interpolations. The concatenation of the upsampled features are passed through a series of router layers, which consist of the two 1×1 convolution layers interleaved by batch normalization and ReLU non-linearity layers. The router finally generates a weight map of size $H \times W \times N$ with N denoting the number of experts. The weight map are normalized using a Softmax layer along the channel dimension. The final depth map is obtained through a summation of the depth maps predicted all the experts weighted by the weight map.

Network Training

Based on our preliminary experiments, we develop a two-stage training strategy, which can facilitate better model learning, allowing experts to acquire more diverse skills (Zhang et al. 2021). In the first stage, we train the encoder-decoder and the four experts. Specifically, we utilize ground truth as routing information, *i.e.*, which expert model should predict each pixel. For a particular expert, its loss solely stems from the depth range it is responsible for, without considering the estimation performance of other ranges. Consequently, each expert can achieve favorable estimation performance within its designated range, endowing the entire model with the capability to handle multiple depth ranges. In the second stage, we freeze all the network parameters obtained from the first stage and only train the router. In both stages, we train the model using the scale-invariant loss (Lee et al. 2019b) as follows:

$$L = \alpha \sqrt{\frac{1}{T}\sum_i g_i^2 - \left(\frac{1}{T}\sum_i g_i\right)^2 + (1-\lambda)\left(\frac{1}{T}\sum_i g_i\right)^2}, \qquad (2)$$

where $g_i = log(d_i) - log(\hat{d}_i)$ with d_i and $\hat{d}_i$ being the ground truth and predicted depth, respectively; T denotes the number of valid pixels; λ and α are set to 0.85 and 10, respectively.

Experiments

In this section, we conduct comprehensive comparisons and ablations to verify the motivation and effectiveness of the proposed method.

Setup

Implementation Details Our training data consists of NYUD v2 (Silberman et al. 2012) and KITTI (Geiger et al. 2013) training datasets. Data augmentation including random horizontal flipping, random changes in brightness, contrast, and random rotation is adopted following ZoeDepth (Bhat et al. 2023). For parameter initialization, the encoder-decoder is initialized using the pre-trained weights of MiDas (Ranftl et al. 2020). During training, the Adam optimizer (Kingma and Ba 2014) is employed with a batch size of 2 and a weight decay of $1e$-2. The initial learning rate is set to $3e$-4. The training process is performed on one NVIDIA GeForce RTX 3090Ti GPU, taking about 20 hours in total.

Our method is evaluated on two standard benchmark datasets: the NYUD v2 test set and the KITTI test set. In addition, to validate the generalization ability of our approach, we conduct further evaluation on four additional datasets that have never been seen during training: DIODE (Vasiljevic et al. 2019), iBims benchmark (Koch et al. 2018), DIML (Kim et al. 2018) and Virtual KITTI 2 (Cabon, Murray, and Humenberger 2020).

Comparisons on NYUD v2 and KITTI

Table 1 reports the quantitative results on NYUD v2 and KITTI test sets. Note that the NYUD v2 and KITTI datasets are collected from indoor and outdoor scenes, respectively, exhibiting significant differences in terms of camera parameters and depth distributions. Consequently, simply merging the two training datasets for network learning without employing specific designs would lead to a performance degradation. To this end, the compared methods BTS (Lee et al. 2019a), Adabins (Bhat, Alhashim, and Wonka 2021), LocalBins (Bhat, Alhashim, and Wonka 2022), PixelFormer (Agarwal and Arora 2023), and NeWCRFs (Yuan et al. 2022) require specific model design and respective network training for each dataset. Neverthless, they still show inferior performance on both datasets with limited generalization capabilities. The recent leading method ZoeDepth (Bhat et al. 2023) employs a scene discriminator and two prediction heads (*i.e.,* indoor and outdoor heads) and does well for the two datsets with one set of model parameters. In contrast, our method DME achieves the best performance with single prediction head upon our distance-aware mechanism. In addition, we also compare with one of our oracle methods DME-GT, which utilizes the ground truth depth instead of the expert predictions for the final Router. It shows that DME obtains comparable and even slightly better performance compared with DME-GT.

Generalization to Unseen Datasets

To validate the generalization performance of our design, we further conduct zero-shot testing on four unseen datasets, including two indoor and three outdoor scenarios. Apart from ZoeDepth (Bhat et al. 2023) and our DME, other compared methods use different sets of trained parameters that are respectively trained using NYUD v2 and KITTI datasets for indoor and outdoor evaluation.

The results of outdoor and indoor scenarios are presented in Table 2 and 3, respectively. For the DIODE Indoor dataset (Vasiljevic et al. 2019), our DME delivers significantly better performance than previous state-of-the-art models. Compared to the prior best method ZoeDepth on the iBims-1 benchmark (Koch et al. 2018), we achieve a reduction of 0.148 in terms of RMSE. For the unseen outdoor datasets Virtual KITTI 2 (Cabon, Murray, and Humenberger 2020), DIML Outdoor (Kim et al. 2018), and DIODE Outdoor (Vasiljevic et al. 2019), DME with the same trained network parameters also show remarkable performance.

Ablation Study

To investigate the main contributions and key designs of our method, a series of ablation experiments are conducted.

Dataset	Method	d1↑	d2↑	d3↑	RMSE↓	log10↓	Abs.Rel↓	RMSE log↓
NYUD v2	BTS	0.885	0.978	0.995	0.392	0.047	0.110	∼
	AdaBins	0.903	0.984	0.997	0.364	0.044	0.103	∼
	LocalBins	0.907	0.987	0.998	0.357	0.042	0.099	∼
	NeWCRFs	0.922	0.992	0.998	0.344	0.041	0.095	∼
	PixelFormer	0.929	0.991	0.998	0.322	0.039	0.090	∼
	ZoeDepth	0.953	0.995	0.999	0.277	0.033	0.077	∼
	DME	0.956	0.995	0.999	0.268	0.032	0.074	0.094
	DME-GT	**0.964**	**0.997**	**1.0**	**0.244**	**0.03**	**0.069**	**0.087**
KITTI	AdaBins	0.964	0.995	0.999	2.360	∼	0.058	0.088
	NeWCRFs	0.974	0.997	0.999	2.129	∼	0.052	0.079
	PixelFormer	0.976	0.997	0.999	2.081	0.077	0.051	∼
	MIM	0.977	0.998	**1.0**	1.966	∼	0.05	0.075
	ZoeDepth	0.967	0.995	0.999	2.290	∼	0.057	0.091
	DME	0.980	**0.999**	**1.0**	**1.905**	0.023	0.05	0.075
	DME-GT	**0.982**	0.998	**1.0**	1.907	**0.021**	**0.048**	**0.073**

Table 1: Quantitative comparison on NYUD v2 and KITTI. The best results are in bold, and the second best is underlined.

	Virtual KITTI 2			DIML Outdoor			DIODE Outdoor		
Method	d1↑	REL↓	RMSE↓	d1↑	REL↓	RMSE↓	d1↑	REL↓	RMSE↓
BTS	0.831	0.115	5.368	0.016	1.785	5.908	0.171	0.837	10.48
AdaBins	0.826	0.122	5.420	0.013	1.941	6.272	0.161	0.863	10.35
LocalBins	0.810	0.127	5.981	0.016	1.820	6.706	0.170	0.821	10.27
NeWCRFs	0.829	0.117	5.691	0.010	1.918	6.283	0.176	0.854	9.228
ZoeDepth	0.850	0.105	5.095	0.292	0.641	3.610	0.208	0.757	7.569
DME	0.840	0.118	4.433	0.199	0.835	3.793	0.251	0.777	9.570
DME-GT	**0.881**	**0.097**	**3.943**	**0.296**	**0.472**	**2.12**	**0.508**	**0.360**	**5.713**

Table 2: Results of zero-shot transfer on three outdoor datasets not seen during training. The results of the prior works are from the original paper of ZoeDepth (Bhat et al. 2023). The best results are in bold, and the second best is underlined.

	DIODE Indoor			iBims-1 Benchmark		
Method	d1	REL	RMSE	d1	REL	RMSE
BTS	0.210	0.418	1.905	0.538	0.231	0.919
AdaBins	0.174	0.443	1.963	0.555	0.212	0.901
LocalBins	0.229	0.412	1.853	0.558	0.211	0.880
NeWCRFs	0.187	0.404	1.867	0.548	0.206	0.861
ZoeDepth	0.386	0.331	1.598	**0.615**	**0.186**	0.777
DME	0.479	0.744	0.862	0.585	0.316	0.635
DME-GT	**0.654**	**0.219**	**0.822**	0.589	0.315	**0.629**

Table 3: Results of zero-shot transfer on two indoor datasets not seen during training. The results of the prior works are from the original paper of ZoeDepth (Bhat et al. 2023). The best results are in bold, and the second best is underlined.

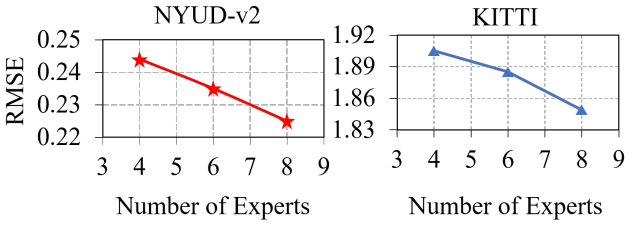

Figure 4: Impact of the Number of Experts on Performance. The two images on the left and right display the test results of the NYUD v2 and KITTI datasets, respectively.

Ablation on Distance Grouping To verify our distance-aware strategy, two basic variants are designed. Considering that ZoeDepth (Bhat et al. 2023) has the comparable parameters as ours but treats all the distance indiscriminately, we retrain it as our 'Baseline' method under the same experimental settings as ours. Besides, we make our four experts learn to be responsible for equal interval distances by dividing the 0-80 meter range into four distance ranges: 0-20 meters, 20-40 meters, 40-60 meters, and 60-80 meters. We term this variant as 'Equally spaced', which does not take into account the frequency distribution of depth values. As shown in Table 4, the results indicate that the equal interval grouping slightly improves the performance compared to the baseline model. On the other hand, our Distance-aware grouping approach outperforms the equal interval grouping. These results indicate that grouping based on the similarity of frequency distributions helps alleviate the issue of data imbalance within each interval, thereby contributing to an overall performance boost.

Furthermore, we compare the number of experts and find that as the number of experts increases, the error decreases further. We still perform depth value grouping based on frequency, and when we use 8 experts, the RMSE decreases to 0.225 for NYUD v2, as shown in Figure 4.

Evaluating Class Imbalance Techniques To evaluate the effectiveness of our method, we compare two commonly used approaches that address the long-tail distribution issue:

Dataset	Method	d1↑	d2↑	d3↑	RMSE↓	log10↓	Abs.Rel ↓
NYUD v2	Baseline	0.953	0.995	0.999	0.277	0.033	0.077
	Equally spaced	0.955	0.995	0.999	0.270	0.032	0.075
	Distance-aware	**0.964**	**0.997**	**1.0**	**0.244**	**0.030**	**0.069**
KITTI	baseline	0.977	**0.998**	**1.0**	2.103	0.024	0.051
	Equally spaced	0.978	**0.998**	**1.0**	2.049	0.022	0.050
	Distance-aware	**0.982**	**0.998**	**1.0**	**1.907**	**0.021**	**0.048**

Table 4: The Impact of Different Grouping Methods. The baseline refers to the results obtained by training ZoeDepth under the same training and testing conditions as ours. 'Equally spaced' corresponds to the method of equally dividing depth values into intervals, while 'Distance-aware' corresponds to the method of grouping depth values based on frequency similarity.

Method	d1↑	d2↑	d3↑	RMSE↓	log10↓	Abs.Rel↓
Baseline	0.953	0.995	0.999	0.277	0.033	0.077
Reweighting-freq	0.945	0.993	0.999	0.334	0.034	0.080
Reweighting-depth	0.923	0.989	0.997	0.337	0.039	0.093
Resampling	0.955	0.995	0.999	0.269	0.032	0.075
MDE-GT-8	**0.966**	**0.998**	**1.0**	**0.225**	**0.030**	**0.069**

Table 5: Comparison of Different Methods for Addressing Imbalanced Data. The resampling approach involves discarding supervision signals from the majority class based on a specified ratio, while the reweighting approach adjusts the weights in the loss function based on the frequency or depth values. All methods are trained on NYUD v2 dataset.

resampling and reweighting. Resampling is a simple and effective technique for handling data imbalance, where we categorize depth values into dominant, common, and minority classes based on frequencies. Specifically, we employ a posterior analysis approach and set the dropout frequencies for the three categories to 0.95, 0.75, and 0.00, respectively. By reducing the sampling frequency of the dominant and common classes, we aim to achieve roughly equal frequencies for the three categories of supervised signals. Reweighting is also considered as a method to address data imbalance, where the loss is adjusted based on the class to increase attention on the minority class. We explore two weighting approaches. The first approach, as described in (Jiao et al. 2018), assigns higher weights to depth values that are farther away from the camera based on their magnitude. The second approach leverages our observations on simulated correlations, assigning larger weights to depth values with longer periods to guide the model's focus toward infrequently occurring depth values.

The experimental results in Table 5 demonstrate that DME-GT-8 which employs 8 experts is the most effective. Our approach can not only address the issue of data skewness but also improves the overall performance. This may be attributed to suboptimal sampling strategies or non-optimal hyperparameters for both methods. However, we are well aware that finding appropriate hyperparameters is a time-consuming process that can render the model fragile. In comparison, our design offers stability and effectiveness.

Limitation

Our work showcases outstanding generalization capabilities on diverse datasets, providing novel insights and methodologies for the research and application of depth estimation. However, a limitation of the current approach lies in the insufficient accuracy of the routing mechanism, which restricts the overall model performance. Therefore, one future research direction is to design a more accurate and elegant routing approach to further enhance the model's performance and streamline the operational workflow.

Conclusion

In this study, we conduct a comprehensive investigation into the phenomenon of distribution skewness in the task of depth estimation and empirically demonstrate the adverse bias it introduces to models. Based on this observation, we propose a Distance-aware Multi-Expert regression model to enhance the performance of depth estimation. The model is designed with a two-stage architecture, where the first stage accurately estimates depth values in different depth ranges, and the second stage utilizes a routing mechanism to achieve a reasonable combination of depth values, yielding the final depth estimation results.

Through this study, we aim to not only contribute to a better understanding of long-tail learning in continuous space for researchers and the academic community but also drive advancements in this field. We hope that this work serves as a reference for improving depth estimation tasks and inspires further research and exploration in the realm of long-tail learning in continuous space.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (U23A20386, 62276045, 62293540,

62293542, 62006036), Dalian Science and Technology Talent Innovation Support Plan (2022RY17), OPPO Research Fund, and Fundamental Research Funds for Central Universities (DUT22LAB124, DUT22QN228)

References

Agarwal, A.; and Arora, C. 2023. Attention Attention Everywhere: Monocular Depth Prediction with Skip Attention. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 5861–5870.

Bhat, S. F.; Alhashim, I.; and Wonka, P. 2021. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4009–4018.

Bhat, S. F.; Alhashim, I.; and Wonka, P. 2022. Localbins: Improving depth estimation by learning local distributions. In *European Conference on Computer Vision*, 480–496.

Bhat, S. F.; Birkl, R.; Wofk, D.; Wonka, P.; and Müller, M. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*.

Cabon, Y.; Murray, N.; and Humenberger, M. 2020. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237.

He, J.; Wang, Y.; Wang, L.; Lu, H.; Luo, B.; He, J.-Y.; Lan, J.-P.; Geng, Y.; and Xie, X. 2023. Towards Deeply Unified Depth-aware Panoptic Segmentation with Bi-directional Guidance Learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 4111–4121.

Hong, Y.; Han, S.; Choi, K.; Seo, S.; Kim, B.; and Chang, B. 2021. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6626–6636.

Jiao, J.; Cao, Y.; Song, Y.; and Lau, R. 2018. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the European Conference on Computer vision*, 53–69.

Kim, D.; Ka, W.; Ahn, P.; Joo, D.; Chun, S.; and Kim, J. 2022. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*.

Kim, Y.; Jung, H.; Min, D.; and Sohn, K. 2018. Deep monocular depth estimation via integration of global and local predictions. *IEEE Transactions on Image Processing*, 27(8): 4131–4144.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koch, T.; Liebel, L.; Fraundorfer, F.; and Korner, M. 2018. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0–0.

Lee, J. H.; Han, M.-K.; Ko, D. W.; and Suh, I. H. 2019a. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*.

Lee, J. H.; Han, M.-K.; Ko, D. W.; and Suh, I. H. 2019b. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 12179–12188.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3): 1623–1637.

Ren, H.; El-Khamy, M.; and Lee, J. 2019. Deep Robust Single Image Depth Estimation Neural Network Using Scene Understanding. In *CVPR Workshops*, volume 2.

Ren, J.; Yu, C.; Ma, X.; Zhao, H.; Yi, S.; et al. 2020. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems*, 33: 4175–4186.

Ren, W.; Wang, L.; Piao, Y.; Zhang, M.; Lu, H.; and Liu, T. 2022. Adaptive co-teaching for unsupervised monocular depth estimation. In *European Conference on Computer Vision*, 89–105.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. *Proceedings of the European Conference on Computer vision*, 7576: 746–760.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357.

Vasiljevic, I.; Kolkin, N.; Zhang, S.; Luo, R.; Wang, H.; Dai, F. Z.; Daniele, A. F.; Mostajabi, M.; Basart, S.; Walter, M. R.; et al. 2019. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*.

Wang, L.; Wang, Y.; Wang, L.; Zhan, Y.; Wang, Y.; and Lu, H. 2021. Can scale-consistent monocular depth be learned in a self-supervised scale-invariant manner? In *Proceedings of the IEEE International Conference on Computer Vision*, 12727–12736.

Wang, L.; Zhang, J.; Wang, O.; Lin, Z.; and Lu, H. 2020a. SDC-Depth: Semantic divide-and-conquer network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 541–550.

Wang, L.; Zhang, J.; Wang, Y.; Lu, H.; and Ruan, X. 2020b. Cliffnet for monocular depth estimation with hierarchical embedding loss. In *European Conference on Computer Vision*, 316–331. Springer.

Wang, T.; Li, Y.; Kang, B.; Li, J.; Liew, J.; Tang, S.; Hoi, S.; and Feng, J. 2020c. The devil is in classification: A simple framework for long-tail instance segmentation. In *Proceedings of the European Conference on Computer vision*, 728–744. Springer.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. X. 2020d. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*.

Wang, Y.; Gan, W.; Yang, J.; Wu, W.; and Yan, J. 2019. Dynamic curriculum learning for imbalanced data classification. In *Proceedings of the IEEE International Conference on Computer Vision*, 5017–5026.

Xie, Z.; Geng, Z.; Hu, J.; Zhang, Z.; Hu, H.; and Cao, Y. 2023. Revealing the dark secrets of masked image modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 14475–14485.

Yuan, W.; Gu, X.; Dai, Z.; Zhu, S.; and Tan, P. 2022. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3916–3925.

Zhang, Y.; Hooi, B.; Hong, L.; and Feng, J. 2021. Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. *arXiv e-prints*, arXiv–2107.

Zhong, Z.; Cui, J.; Liu, S.; and Jia, J. 2021. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16489–16498.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9719–9728.