Hierarchical Normalization for Robust Monocular Depth Estimation

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we address monocular depth estimation with deep neural networks. To enable training of deep monocular estimation models with various sources of datasets, state-of-the-art methods adopt image-level normalization strategies to generate affine-invariant depth representations. However, learning with the image-level normalization mainly emphasizes the relations of pixel representations with the global statistic in the images, such as the structure of the scene, while the fine-grained depth difference may be overlooked. In this paper, we propose a novel multi-scale depth normalization method that hierarchically normalizes the depth representations based on spatial information and depth distributions. Compared with previous normalization strategies applied only at the holistic image level, the proposed hierarchical normalization can effectively preserve the fine-grained details and improve accuracy. We present two strategies that define the hierarchical normalization contexts in the depth domain and the spatial domain, respectively. Our extensive experiments show that the proposed normalization strategy remarkably outperforms previous normalization methods, and we set new state-of-the-art on five zero-shot transfer benchmark datasets.

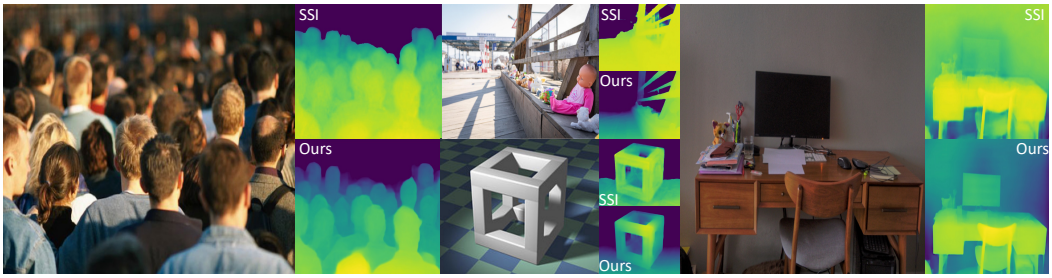

Figure 1: We propose a hierarchical depth normalization strategy to improve the training of monocular depth estimation models. Compared with previous normalization strategy SSI [22], our design effectively improves the details and smoothness of predictions. We visualize the predictions of close regions to observe details.

1 Introduction

Data-driven deep learning based monocular depth estimation has gained wide interest in recent years, due to its low requirements of sensing devices and impressive progress. Among various learning objectives of deep monocular estimation, zero-shot transfer carries the promise of learning a generic

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

depth predictor that can generalize well across a variety of scenes. Rather than training and evaluation on the subsets of individual benchmarks that usually share similar characteristics and biases, zero-shot transfer expects the models to be deployed for predictions of any in-the-wild images.

To achieve this goal, large-scale datasets with equally high diversity for training are necessary to enable good generalization. However, collecting data with high-quality depth annotations is expensive, and existing benchmark datasets often show limitations in scales or diversities. Many recent works [22, 33] seek mix-dataset training, where datasets captured by various sensing modalities and in diverse environments can be jointly utilized for model training, which largely alleviates the difficulty of obtaining diverse annotated depth data at scale. Nevertheless, the mix-data training also comes with its challenges, as different datasets may demonstrate inconsistency in depth representations, which causes incompatibility between datasets. For example, the disparity map generated from web stereo images [30] or 3D movies [22] can only provide depth annotations up to a scale and shift, due to varied and unknown camera models.

To solve this problem, state-of-the-art methods [22, 33] seek training objectives invariant to the scale-and-shift changes in the depth representations by normalizing the predictions or depth annotations based on statistics of the image instance, which largely facilitates the mix-data learning of depth predictors. However, as the depth is represented by the magnitude of values, normalization based on the instance inevitably squeezes the fine-grained depth difference, particularly in close regions. Suppose that an object-centric dataset with depth annotations is available for training, where the images sometimes include backgrounds with high depth values. Normalizing the depth representations with global statistics can distinctly separate the foreground and background, but it may meanwhile lead to an overlook of the depth difference in objects, which may be our real interest. As the result, the learned depth predictor often excels at predicting the relative depth representations of each pixel location with respect to the entire scene in the image, such as the overall scene structure, but struggles to capture the fine-grained details.

Motivated by the bias issue in existing depth normalization approaches, we aim to design a training objective that should have the flexibility to optimize both the overall scene structure and fine-grained depth difference. Since depth estimation is a dense prediction task, we take inspiration from classic local normalization approaches, such as local response normalization in AlexNet [15] and local histogram normalization in SIFT [18] and HOG features [6], which rely on normalized local statistics to enhance local contrast or generate discriminative local descriptors. By varying the size of the local window, we can control how much context is involved to generate a normalized representation. With such insights, we present a hierarchical depth normalization strategy that normalizes depth representations with different scopes of contexts for learning. Intuitively, a large context for normalization emphasizes the depth difference globally, while a small context focuses on the minor difference locally. We present two implementations that define the multi-scale contexts in the spatial domain and the depth domain, respectively. For the strategy in the spatial domain, we divide the image into several sub-regions and the context is defined as the pixels in individual cells. In this way, the fine-grained difference between spatially close locations is emphasized. By varying the grid size, we obtain multiple depth representations of each pixel that rely on different contexts. For the strategy in the depth domain, we group pixels based on the ground truth depth values to construct contexts, such that pixels with similar depth values can be differentiated. Similarly, we can change the number of groups to control the context size. By combining the normalized depth representations under various contexts for optimization, the learner emphasizes both the fine-grained details and the global scene structure, as shown in Fig. 1.

To validate the effectiveness of our design, we conduct extensive experiments on various benchmark datasets. Our empirical results show that the proposed hierarchical depth normalization remarkably outperforms the existing instance-level normalization qualitatively and quantitatively. Our contributions are summarized as follows:

- We propose a hierarchical depth normalization strategy to improve the learning of deep monocular estimation models.
- We present two implementations to define the multi-scale normalization contexts of HDN based on spatial information and depth distributions, respectively.
- Experiments on five popular benchmark datasets show that our method significantly outperforms the baselines and sets new state-of-the-art results.

2 Related Work

Deep monocular depth estimation. As opposed to early works on monocular depth estimation based on hand-crafted features, recent studies advocate end-to-end learning based on deep neural networks. Since the pioneer work [7] first adopts deep neural networks to undertake monocular depth estimation, significant progress has been made from many aspects, such as network architectures [16, 20], large-scale and diverse training datasets [32, 34], loss functions [22, 33], multi-task learning [35], synthesized dataset [8], geometry constraint [32, 20] and various sources of supervisions [31, 33].

For supervised training, collecting high-diversity data with ground truth depth annotations at scale is expensive. Recent works based on ranking loss [2, 31] and scale-and-shift invariant losses [22, 33] enable network training with other forms of annotations, such as ordinal depth annotations [2, 30] , or relative inverse depth map [31, 32, 22] generated by uncalibrated stereo images using optical flow algorithms [26]. In particular, scale-and-shift invariant (SSI) loss [22] and image-level normalization loss [33] allow data from multiple sources to be learned in a fully supervised depth regression manner, which largely facilitates large-scale training and improves the generation ability of learning based depth estimators. The SSI loss removes the major incompatibility between various datasets, *i.e.*, the scale and shift changes, by transforming the depth representation into a canonical space through normalization. With such advances, zero-shot transfer is made possible, where the network learned on a large-scale database with high diversity can be directly evaluated on various benchmarks without seeing their training samples, which is the focus of this paper.

Furthermore, some literature [1, 38, 10, 24] proposes to solve the monocular depth estimation problem without sensor-captured ground truth but leverages the training signal from consecutive temporal frames or stereo videos. However, most of these methods need the camera intrinsic parameters for supervision.

Normalization in CNNs. Normalization is widely adopted in deep neural networks, while different normalization strategies are employed for different purposes. For instance, batch normalization (BN) [13] normalizes the feature representations along the batch dimension to stabilize training and accelerate convergence. BN usually prefers large normalization contexts to obtain robust feature representation. On the other hand, another line of normalization methods relies on local statistics. For example, instance normalization [27] and its variants [12, 19] based on instance-level statistics dominate the style transfer task, as they emphasize the unique styles of individual images. A collection of literature seeks normalization in local regions. For example, the well-known SIFT feature [18] and HOG features [6] are based on the normalized local statistics to generate discriminative local features. Since the local details and the overall scene structure are both important for a depth estimator, we incorporate the ideas of both global normalization and local normalization in the monocular depth estimation models.

3 Method

In this section, we first briefly summarize the preliminaries of the task. Then we define a unified form of depth normalization and show that the normalization strategy in scale-and-shift invariant loss[22] is a special case. Finally, we present two implementations of our proposed hierarchical depth normalization approaches based on the spatial domain and the depth domain, respectively.

3.1 Preliminaries

We aim to boost the performance of zero-shot monocular depth estimation with diverse training data. In our pipeline, we input a single RGB image $I \in \mathbb{R}^{H \times W \times 3}$ to the depth prediction network $\mathcal{F}(\cdot)$ to generate a depth map $D \in \mathbb{R}^{H \times W \times 3}$. Instead of directly regressing the output map with the ground-truth depth supervision, state-of-the-art methods [32, 22, 33] normalize the depth representations before computing the regression loss. In our work, we focus on the depth normalization part, which means our methods can be easily combined with state-of-the-art network structures or loss functions.

3.2 Unified Form of Depth Normalization

Let $\mathbf{d} \in \mathbb{R}^M$ and $\mathbf{d}^* \in \mathbb{R}^M$ denote the vectorized predicted depth map and the ground truth depth map, respectively, where M is the number of pixels with valid annotations. Our goal is to generate

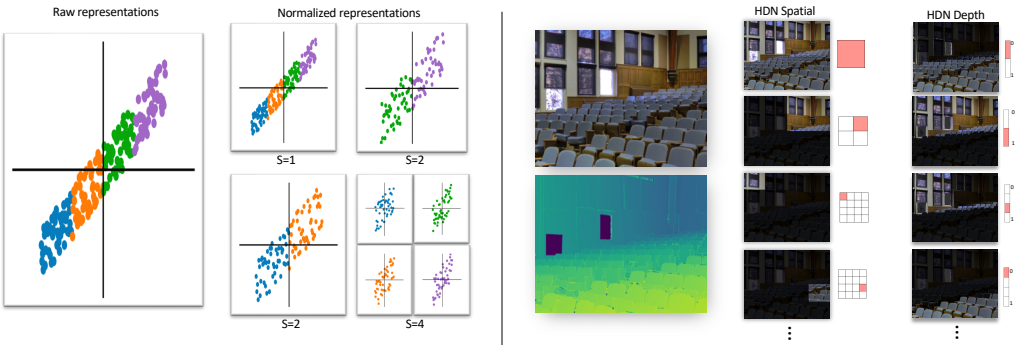

Figure 2: Left: An illustration of our proposed hierarchical depth normalization (HDN) strategy. We divide representations into S groups, denoted by different colors, and apply normalization to individual groups to generate representations for learning. We select different S to obtain multiple representations for learning. By doing so, we can emphasize the fine-grained difference in representations during training. Right: Our two normalization strategies for the depth estimation task that group pixels in the spatial domain (HDN-Spatial) and the depth domain (HDN-Depth), respectively. The HDN-Spatial groups pixels in sub regions, while the HDN-Depth groups pixels based on depth distributions. The red cells denote the normalization contexts.

the normalized representations, $\mathcal{N}_{u_i}(d_i)$ and $\mathcal{N}_{u_i}(d_i^*)$ for computing the regression loss of each location i, where u_i is the set of location indexes that constitute the context of location i, and $\mathcal{N}_{u_i}$ is the normalization function based on the context u_i. For example, the normalization employed in scale-and-shift invariant loss (SSI) [22] is to explicitly remove the estimated scale and shift from the raw depth representations:

$$\mathcal{N}_{u_i}(d_i) = \frac{d_i - \text{median}_{u_i}(\mathbf{d})}{\frac{1}{|u_i|}\sum_{j=1}^{|u_i|}|d_i - \text{median}u_i(\mathbf{d})|}, \quad \mathcal{N}_{u_i}(d_i^*) = \frac{d_i^* - \text{median}_{u_i}(\mathbf{d}^*)}{\frac{1}{|u_i|}\sum_{j=1}^{|u_i|}|d_i^* - \text{median}u_i(\mathbf{d}^*)|} \tag{1}$$

, where median_{u_i} operator computes the median depth of locations that belong to u_i. The instance-level normalization in SSI assigns a global context u_{glo} for each pixel that involves all pixel locations with valid depth annotations, *i.e.*, $u_j = u_{\text{glo}}, \forall j \in u_{\text{glo}}$. Finally, the SSI loss $\mathcal{L}_{u_i}^{\text{SSI}}$ computes the mean absolute error between the normalized prediction and the ground truth, and losses over all locations are averaged to obtain the final loss $\mathcal{L}^{\text{SSI}}$:

$$\mathcal{L}^{\text{SSI}} = \frac{1}{M}\sum_{i=1}^{M}\mathcal{L}_i^{\text{SSI}}, \text{where } \mathcal{L}_i^{\text{SSI}} = |\mathcal{N}_{u_i}(d_i) - \mathcal{N}_{u_i}(d_i^*)| \tag{2}$$

Intuitively, a large normalization context that covers a large scope of locations emphasizes the relative depth representations with the global statistics, while a small context emphasizes the depth difference more locally. Based on this observation, we propose to assign multiple contexts at different scales for each location.

3.3 Hierarchical Depth Normalization

We next present two strategies that define the multi-scale contexts in the spatial domain and the depth domain, respectively, which are illustrated in Fig. 2

In spatial domain (HDN-S). Since images have 2D grid-like representations, we can define the context based on spatial positions. In this way, the context corresponds to a group of pixels in the sub region. Defining context in spatial domain based on grids was also seen in segmentation models, where poolings are used to explicitly enlarge the effective receptive fields of features[37]. Here we adopt a similar strategy that evenly divides the image plane into several sub-regions based on a $S \times S$ grid. Then the pixels belonging to the same region share the same context for normalization. By varying the grid size, we can obtain multiple contexts for each pixel location. In our design, we select the grid size S_{spatial} from $\{2^0, 2^1, 2^2, ...\}$ and apply normalization at each level. For example, at the highest level, the normalization is essentially the instance-level normalization that covers all pixels as the context, while at the lowest level, the normalization is limited to an $\frac{H}{S} \times \frac{W}{S}$ image patch.

In depth domain (HDN-D). The other way to group close pixels is based on their ground-truth depth distributions, such that pixels with similar depth values share the same context. Here we present two strategies. The first one, denoted by **HDN-DP**, is to sort all the pixels based on their ground-truth depth values and evenly divide them into S bins, where the numbers of pixels in each bin are $\frac{M}{S}$. The second strategy, denoted by HDN-DR, is to evenly divide the depth range presented in an image into S bins, and classify pixels into different bins. Pixels belonging to the same bin share the same context. Since pixels with similar depth values may not be spatially close, long-range spatial dependency can be established by the shared contexts to promote global coherency. The bin number S_{depth} is also selected from $\{2^0, 2^1, 2^2, ...\}$.

Suppose that the multi-scale contexts of location i constitutes a set U_i, we compute the SSI losses generated by each context from U_i and average them to obtain the final loss $\mathcal{L}^{\text{HDN}}$:

$$\mathcal{L}^{\text{HDN}} = \frac{1}{M}\sum_{i=1}^{M}\mathcal{L}_i^{\text{HDN}}, \text{where } \mathcal{L}_i^{\text{HDN}} = \frac{1}{|U_i|}\sum_{u\in U_i}|\mathcal{N}_u(d_i) - \mathcal{N}_u(d_i^*)|. \tag{3}$$

As we can see, by combing the contexts at different scales, the proposed learning objective emphasizes the depth difference at different levels, which can preserve the local fine-grained details while maintaining structural consistency. In contrast, the SSI loss is a special case that only focuses on the global structure.

Our proposed hierarchical normalization strategies are lightweight and easy to implement. Since the SSI loss functions are usually implemented as a function of ground truth depth maps, predicted depth maps, and valid-pixel maps, our approaches can be easily implemented by only modifying the valid-pixel maps to obtain batched computations, where the pixels out of the context are masked out.

4 Experiments

Datasets We follow Leres [33] to construct a mix-data training set, which includes 114K images from Taskonomy dataset, 121K images from DIML dataset, 48K images from Holopix50K, and 20K images from HRWSI [31]. We withhold 1K images from all datasets for validation during training. The details of each dataset are as follows.

Taskonomy. Taskonomy [34] is a high-quality and large-scale dataset, which contains over 4 million images of indoor scenes from about 600 buildings. It includes annotations for over 20 tasks. In our experiments, we sampled around 114k RGB-D pairs for training.

DIML. DIML [14] contains synchronized RGB-D frames from Kinect v2 or Zed stereo camera. For the outdoor split, they are mainly captured by the calibrated stereo cameras. It contains various outdoor places, *e.g.*, offices, rooms, dormitory, exhibition center, street, road and so on. We used GANet [36] to recompute the disparity and depth maps for training.

HRWSI & Holopix50k. HRWSI [31] and Holopix50k [11] are both diverse relative depth datasets. Although they contain diverse scenes and various camera settings, their provided stereo images are uncalibrated. Therefore, the stereo matching methods are inapplicable to recover the metric depth information. We use RAFT [26] to recover their relative depth representations for training.

Implementation details. We use the state-of-the-art monocular depth estimation network DPT-Hybrid [21] in our experiments. The input image size is 480×480 at both training and testing time. For evaluation on benchmarks, we re-size the outputs to the raw image resolutions with the bilinear interpolation. Random horizontal flip and random crop are employed for data augmentation. The default random crop size in all experiments is sampled in $[0.5, 1]$ of the raw image size, with the aspect ratio restricted in $[3/4, 4/3]$, and the random cropped patches are re-sized to the input resolution for training. We use the Adam optimizer with a learning rate of 10^{-5}. The model is trained on 8 V100 GPU with the batch size of 32. In each mini-batch, we sample the equal number of images from different training data sources. For HDN in the spatial domain, we select the grid size S_{spatial} from $\{2^0, 2^1, 2^2, 2^3\}$ to construct the hierarchical contexts. For HDN in depth domain, the group number S_{depth} is chosen from $\{2^0, 2^1, 2^2\}$.

Evaluation Metrics. We include 5 popular benchmarks that are unseen during training, including, DIODE [28], ETH3D [23], KITTI [9], NYU [25], and ScanNet [5]. We follow the previous works to evaluate our model on zero-shot cross-dataset transfer. We use the mean absolute value of the relative

Norm. Method	DIODE	ETH3D	KITTI AbsRel ↓	NYU	ScanNet	Mean Improve.
Instance	47.1	19.4	13.2	9.6	10.22	-
Batch	45.0 (-4%)	25.2 (+30%)	14.9 (+12%)	11.5 (+19%)	12.2 (+19%)	(+15%)
Local-S	33.1 (-30%)	14.5 (-25%)	19.3 (+45%)	11.8 (+22%)	11.5 (+12%)	(+5%)
Local-DR	25.7 (-45%)	12.7 (-35%)	14.7 (+11%)	9.4 (-2%)	9.9 (-3%)	(15%)
Local-DP	28.8 (-39%)	14.4 (-26%)	23.3 (+76%)	10.4 (+8%)	10.95 (+7%)	(+5%)
HDN-S	36.4 (-22%)	14.5 (-25%)	12.5 (-6%)	8.8 (-8%)	9.5 (-6%)	(-14%)
HDN-DR	24.9 (-47%)	12.7 (-35%)	11.9 (-10%)	8.7 (-10%)	9.4 (-8%)	(-22%)
HDN-DP	25.1 (-47%)	12.8 (-34%)	13.2 (0%)	8.7 (-9%)	9.3 (-8%)	(-19%)

Table 1: Comparison of different depth normalization strategies. Our proposed HDN outperforms the instance-level normalization baseline remarkably. Relying only on local normalization produces unstable results across benchmarks.

error (AbsRel): $(1/M)\sum_{i=1}^{M}|d_i - d_i^*|/d_i^*$ and the percentage of pixels with $\delta_1 = \max(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}) < 1.25$. Following Midas [22] and Leres [33], we align the predictions and ground truth in scale and shift before evaluation. Please refer to our supplementary material for more experiment results and analysis.

4.1 Analysis

For ablation study and analysis, we sample a subset of 16K images evenly from our four training sets, and the input image size is 384×384.

Comparison of normalization strategies. At the beginning, we compare different normalization strategies. The first baseline is the instance-level normalization, *i.e.*, scale-and-shift invariant loss. We also include the **batch**-level depth normalization for a comprehensive study. For our proposed HDN, we report the results of models based on the spatial domain (**HDN-S**) and the depth domain (**HDN-DP** and **HDN-DR**). We also take the finest normalization level of the proposed HDN to validate whether local contexts alone help the training, which is denoted by **Local**. A detailed record of the experiment results is presented in Table 1.

As we can see, normalization along the batch dimension yields the worst results, which means that the scale-and-shift changes between data can not be addressed by utilizing batch-level statistics. Our proposed HDNs in both the depth domain and the spatial domain remarkably outperform the instance-level normalization baseline, which validates the effectiveness of our design. The HDNs in the depth domain are more effective. In particular, HDN-DR reduces the error by an average of 22% on the five benchmarks. We find that the instance normalization baseline performs poorly on DIODE and ETH3D datasets, particularly in their outdoor splits. This may be caused by dataset bias or the shortage of diversity in the training datasets. In contrast, the proposed HDNs perform well on all benchmark datasets and even *reach state-of-the-art with only 16K training data, which are dozens of times fewer than the training sets used in recent methods.* Relying solely on local normalization, *i.e.*, the finest normalization level, can improve the baseline on some benchmarks, *e.g.*, DIODE and ETH3D, but it may meanwhile cause significant performance drops on other benchmarks, *e.g.*, KITTI. This suggests that the combination of local normalization and global normalization is the optimal strategy for robust training and good generalization.

Delve into HDN. We next study the characteristics of the proposed HDN. As the key difference between HDNs and instance-level normalization is the extra local normalization contexts, we first validate whether our methods still work well when different strength of random cropping is applied for data augmentation. We select the lower bound of cropping size from $\{1/8, 1/4, 1/2\}$ of the raw size to observe the improvements. We then gradually add more fine-grained normalization context levels and observe the performance changes. For example, if three levels are adopted for HDN-Depth, the group number S_{depth} is chosen from $\{2^0, 2^1, 2^2\}$. When only the top level is adopted, the HDN degrades to the instance-level normalization baseline. We also report performance changes of our method relying on the finest levels alone, which is also the **Local** model variants in Table 1. For all experiments, we report the mean relative error drop rate (AbsRel) over the instance-level normalization baseline in Table 1, whose lower bound of random crop is 0.5.

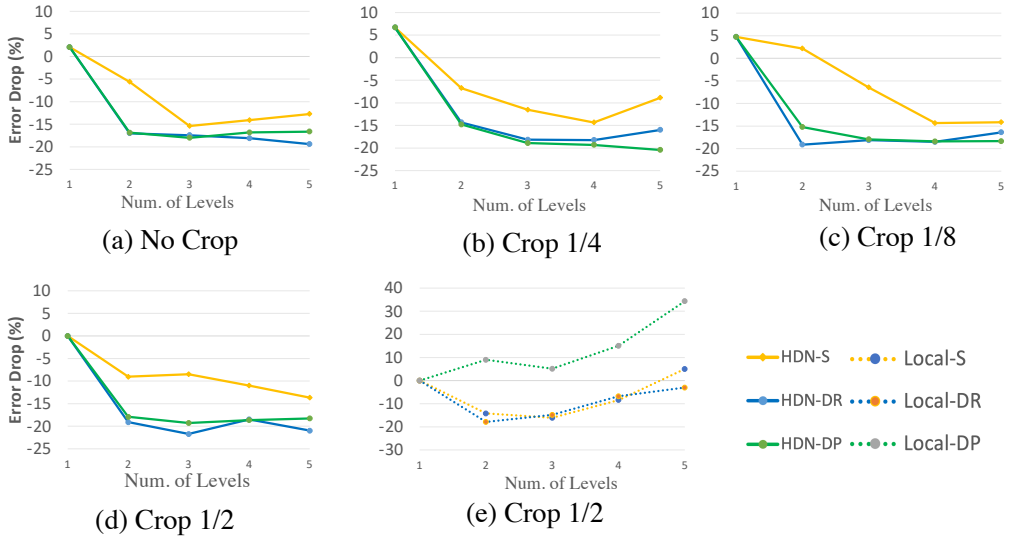

Figure 3: Different numbers of context levels for the proposed Hierarchical Depth Normalization and the variants based solely on the lowest level (Local). Lower means better. The proposed normalization approaches consistently outperform the baselines (number of level $= 0$) under various data augmentation strengths. Relying only on local normalization with very small contexts may harm the training.

The results are presented in Fig 3. Removing the random crop (**No Crop**) harms the overall performance, which can be seen from the starting point of different charts that denote the instance-level normalization baselines. Although instance-level normalization with aggressive random crop (**Crop 1/8** and **Crop 1/4**) can also enforce local spatial normalization contexts, it may meanwhile lose useful contexts for feature encoding and causes performance drops. Our proposed HDNs can consistently and significantly outperform the baselines under various augmentation strengths, which indicates that data augmentation with random cropping and our HDNs are complementary.

For our HDNs, adding more fine-grained normalization context levels can not continually improve the performance. By observing the results of normalization relying on the local contexts alone, we find that they can still outperform the instance-level normalization baseline when the appropriate level is assigned. However, the performance is likely to get worse significantly when the contexts become too local, as can be seen from the result of **Local-DP**.

Qualitative results. We present some qualitative comparisons in Figure 4. We mainly compare with the instance-level normalization baseline, *i.e.*, the scale-shift-invariant (SSI) loss [22]. All comparison cases are sampled from the zero-shot testing datasets. We can observe that our hierarchical normalization methods make better predictions in the smoothness of object surface and the sharpness of edges. Therefore, compared to the global instance-level normalization, our extra local normalization in the depth domain or the depth domain can force the network to master better knowledge of local geometries. Furthermore, we randomly sample 2K pixels from each image and plot the predicted depth value and ground truth depth value in the last two columns. Ideally, the predicted affine-invariant depth should be strictly linear to the ground truth, *i.e.*, the red diagonal line. We can observe that the linearity of our method is much better than the baseline. Note that even if the baseline could achieve similar prediction accuracy sometimes (such as the second example), our methods generate more find-grained details.

4.2 Comparison with State-of-the-Art

Finally, we compare our depth estimator (HDN-DR) with the state-of-the-art methods on five benchmark datasets, including, DIODE [28], ETH3D [23], KITTI [9], NYU [25], and ScanNet [5], which are unseen during training. As the result shows in Table 2, our method outperforms previous

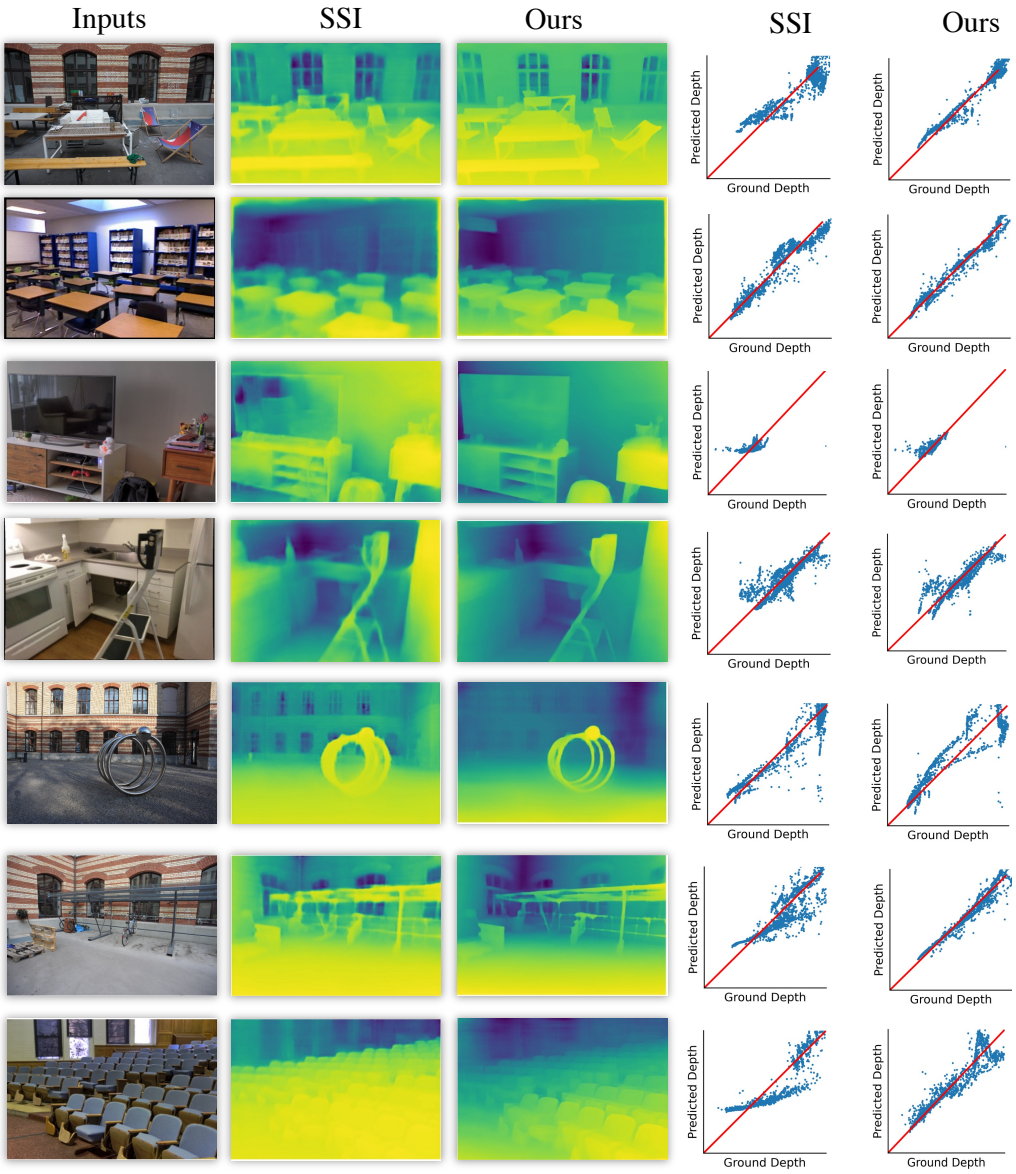

Figure 4: Qualitative comparison of our method and the baseline replying on instance-level normalization (SSI [22]). The red diagonal lines denote the ground-truth, and closer to it means better. Our method generates smoother surfaces, sharper edges and more details.

methods by a large margin on multiple benchmarks with fewer training data than recent state-of-the-art methods, such as Leres [33] and Midas [22].

5 Discussion

5.1 Limitation

We find that for datasets with very high image resolutions, such as ETH3D with 4032×6048 raw resolutions, a relatively small inference resolution, *e.g.*, 384×384, is worse than a relatively large resolution, *e.g.*, 800×800, but further increasing the resolution can not consistently improve the performance. Therefore, there is a lack of principle for setting the optimal input resolutions at test

Method	Training Data	NYU		KITTI		DIODE		ScanNet		ETH3D	
		AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑
OASIS [4]	$\sim$ 140K	21.9	66.8	31.7	43.7	48.4	53.4	19.8	69.7	29.2	59.5
MegaDepth [17]	$\sim$ 150K	19.4	71.4	20.1	66.3	39.1	71.2	26.0	64.3	39.8	52.7
Xian *et al.*[31]	$\sim$ 30K	16.6	77.2	27.0	52.9	42.5	61.8	17.4	75.9	27.3	63.0
WSVD [29]	$\sim$ 1.5M	22.6	65.0	24.4	60.2	35.8	63.8	18.9	71.4	26.1	61.9
Chen *et al.*[3]	$\sim$ 795K	16.6	77.3	32.7	51.2	37.9	66.0	16.5	76.7	23.7	67.2
DiverseDepth [32]	$\sim$ 300K	11.7	87.5	19.0	70.4	37.6	63.1	10.8	88.2	22.8	69.4
MiDaS [22]	$\sim$ 2M	11.1	88.5	23.6	63.0	33.2	71.5	11.1	88.6	18.4	75.2
Leres [33]	$\sim$ 360K	9.0	91.6	14.9	78.4	27.1	76.6	9.5	91.2	17.1	77.7
Ours	$\sim 300K$	**6.9**	**94.8**	**11.5**	**86.7**	**24.6**	**78.0**	**8.0**	**93.9**	**12.1**	**83.3**

Table 2: Comparison with state-of-the-art methods on six zero-shot transfer benchmark datasets. Our model significantly outperforms previous methods and sets new state-of-the-art in many benchmarks. We use the MiDaS model reported in their original paper for comparison.

time. We believe research on adaptive input resolutions and learnable test-time augmentations would be promising solutions in the future.

5.2 Conclusion

In this paper, we have presented a novel hierarchical depth normalization strategy for monocular depth estimation tasks. Compared with the existing instance-level normalization strategy that mainly focuses on the global structure of the scene in the image, our normalization and loss preserve both the fine-grained details and overall structure. Extensive experiments validate the effectiveness of our two implementations in the spatial domain and the depth domain, and new state-of-the-art performance is set on multiple benchmarks.

References

[1] Jia-Wang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Proc. Advances in Neural Inf. Process. Syst.*, 2019.

[2] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 730–738, 2016.

[3] Weifeng Chen, Shengyi Qian, and Jia Deng. Learning single-image depth from videos using quality assessment networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5604–5613, 2019.

[4] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 679–688, 2020.

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5828–5839, 2017.

[6] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Ieee, 2005.

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 2366–2374, 2014.

[8] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4340–4349, 2016.

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3354–3361. IEEE, 2012.

[10] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 3828–3838, 2019.

[11] Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. Holopix50k: A large-scale in-the-wild stereo image dataset. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, June 2020.

[12] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1501–1510, 2017.

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. Int. Conf. Mach. Learn.*, 2015.

[14] Youngjung Kim, Hyungjoo Jung, Dongbo Min, and Kwanghoon Sohn. Deep monocular depth estimation via integration of global and local predictions. *IEEE Trans. Image Process.*, 27(8):4131–4144, 2018.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Advances in Neural Inf. Process. Syst.*, 2012.

[16] 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. *arXiv: Comp. Res. Repository*, page 1907.10326, 2019.

[17] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2041–2050, 2018.

[18] David G Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[19] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2337–2346, 2019.

[20] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 283–291, 2018.

[21] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 12179–12188, 2021.

[22] 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 Trans. Pattern Anal. Mach. Intell.*, 2020.

[23] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3260–3269, 2017.

[24] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *Proc. Eur. Conf. Comp. Vis.*, pages 572–588, 2020.

[25] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. Eur. Conf. Comp. Vis.*, pages 746–760. Springer, 2012.

[26] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proc. Eur. Conf. Comp. Vis.*, pages 402–419. Springer, 2020.

[27] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[28] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv: Comp. Res. Repository*, page 1908.00463, 2019.

[29] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *Int. Conf. 3D. Vis.*, pages 348–357. IEEE, 2019.

[30] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 311–320, 2018.

[31] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 611–620, 2020.

[32] Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraintsfor accurate and robust depth prediction. *arXiv: Comp. Res. Repository*, page 2103.04216, 2021.

[33] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021.

[34] Amir Zamir, Alexander Sax, , William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* IEEE, 2018.

[35] Amir R. Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J. Guibas. Robust learning through cross-task consistency. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2020.

[36] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 185–194, 2019.

[37] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2881–2890, 2017.

[38] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1851–1858, 2017.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [No]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We have clearly present all experiments settings and data in our experiments sections and supplementary materials.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We have repeated experiments with different random seeds, and all conclusions in our experiments still hold.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See section 6.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [No] The data we employed are open-source. Their licenses have been clarified on their websites.

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]