Single-Image Depth Perception in the Wild

Weifeng Chen **Zhao Fu** **Dawei Yang** **Jia Deng**
University of Michigan, Ann Arbor
{wfchen,zhaofu,ydawei,jiadeng}@umich.edu

Abstract

This paper studies single-image depth perception in the wild, i.e., recovering depth from a single image taken in unconstrained settings. We introduce a new dataset "Depth in the Wild" consisting of images in the wild annotated with relative depth between pairs of random points. We also propose a new algorithm that learns to estimate metric depth using annotations of relative depth. Compared to the state of the art, our algorithm is simpler and performs better. Experiments show that our algorithm, combined with existing RGB-D data and our new relative depth annotations, significantly improves single-image depth perception in the wild.

提出相对深度数据，并利用相对深度估计

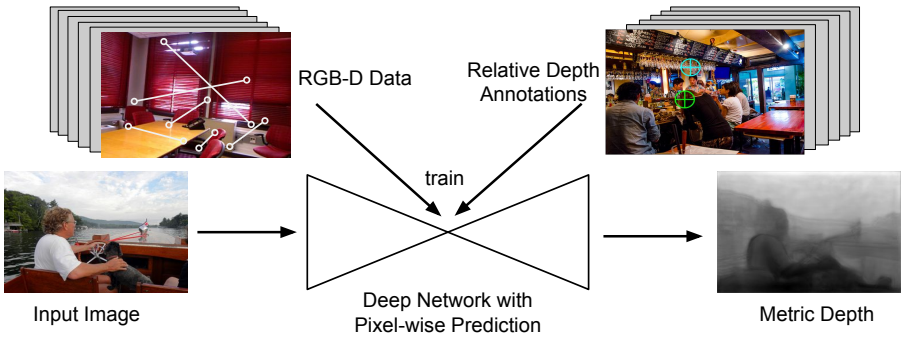

Figure 1: We crowdsource annotations of relative depth and train a deep network to recover depth from a single image taken in unconstrained settings ("in the wild").

1 Introduction

Depth from a single RGB image is a fundamental problem in vision. Recent years have seen rapid progress thanks to data-driven methods [1, 2, 3], in particular, deep neural networks trained on large RGB-D datasets [4, 5, 6, 7, 8, 9, 10]. But such advances have yet to broadly impact higher-level tasks. One reason is that many higher-level tasks must operate on images "in the wild"—images taken with no constraints on cameras, locations, scenes, and objects—but the RGB-D datasets used to train and evaluate image-to-depth systems are constrained in one way or another.

Current RGB-D datasets were collected by depth sensors [4, 5], which are limited in range and resolution, and often fail on specular or transparent objects [11]. In addition, because there is no Flickr for RGB-D images, researchers have to manually capture the images. As a result, current RGB-D datasets are limited in the diversity of scenes. For example, NYU depth [4] consists mostly of indoor scenes with no human presence; KITTI [5] consists mostly of road scenes captured from a car; Make3D [3, 12] consists mostly of outdoor scenes of the Stanford campus (Figure. 2). While these datasets are pivotal in driving research, it is unclear whether systems trained on them can generalize to images in the wild.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Is it possible to collect ground-truth depth for images in the wild? Using depth sensors in unconstrained settings is not yet feasible. Crowdsourcing seems viable, but humans are not good at estimating metric depth, or 3D metric structure in general [13]. In fact, metric depth from a single image is fundamentally ambiguous: a tree behind a house can be slightly bigger but further away, or slightly smaller but closer—the absolute depth difference between the house and the tree cannot be uniquely determined. Furthermore, even in cases where humans can estimate metric depth, it is unclear how to elicit the values from them.

But humans are better at judging relative depth [13]: "Is point A closer than point B?" is often a much easier question for humans. Recent work by Zoran et al. [14] shows that it is possible to learn to estimate metric depth using only annotations of relative depth. Although such metric depth estimates are only accurate up to monotonic transformations, they may well be sufficiently useful for high-level tasks, especially for occlusion reasoning. The seminal results by Zoran et al. point to two fronts for further progress: (1) collecting a large amount of relative depth annotations for images in the wild and (2) improving the algorithms that learn from annotations of relative depth.

In this paper, we make contributions on both fronts. Our first contribution is a new dataset called "Depth in the Wild" (DIW). It consists of 495K diverse images, each annotated with randomly sampled points and their relative depth. We sample one pair of points per image to minimize the redundancy of annotation [1]. To the best of our knowledge this is the first large-scale dataset consisting of images in the wild with relative depth annotations. We demonstrate that this dataset can be used as an evaluation benchmark as well as a training resource [2].

Our second contribution is a new algorithm for learning to estimate metric depth using only annotations of relative depth. Our algorithm not only significantly outperforms that of Zoran et al. [14], but is also simpler. The algorithm of Zoran et al. [14] first learns a classifier to predict the ordinal relation between two points in an image. Given a new image, this classifier is repeatedly applied to predict the ordinal relations between a sparse set of point pairs (mostly between the centers of neighboring superpixels). The algorithm then reconstructs depth from the predicted ordinal relations by solving a constrained quadratic optimization that enforces additional smoothness constraints and reconciles potentially inconsistent ordinal relations. Finally, the algorithm estimates depth for all pixels assuming a constant depth within each superpixel.

In contrast, our algorithm consists of a single deep network that directly predicts pixel-wise depth (Fig. 1). The network takes an entire image as input, consists of off-the-shelf components, and can be trained entirely with annotations of relative depth. The novelty of our approach lies in the combination of two ingredients: (1) a multi-scale deep network that produces pixel-wise prediction of metric depth and (2) a loss function using relative depth. Experiments show that our method produces pixel-wise depth that is more accurately ordered, outperforming not only the method by Zoran et al. [14] but also the state-of-the-art image-to-depth system by Eigen et al. [8] trained with ground-truth metric depth. Furthermore, combing our new algorithm, our new dataset, and existing RGB-D data significantly improves single-image depth estimation in the wild.

2 Related work

RGB-D Datasets: Prior work on constructing RGB-D datasets has relied on either Kinect [4, 15, 16, 17] or LIDAR [3, 5]. Existing Kinect-based datasets are limited to indoor scenes; existing LIDAR-based datasets are biased towards scenes of man-made structures [3, 5]. In contrast, our dataset covers a much wider variety of scenes; it can be easily expanded with large-scale crowdsourcing and the virually umlimited Internet images.

Intrinsic Images in the Wild: Our work draws inspiration from Intrinsic Images in the Wild [18], a seminal work that crowdsources annotations of relative reflectance on unconstrained images. Our work differs in goals as well as in several design decisions. First, we sample random points instead of centers of superpixels, because unlike reflectance, it is unreasonable to assume a constant depth within a superpixel. Second, we sample only one pair of points per image instead of many to maximize the value of human annotations.

Depth from a Single Image: Image-to-depth is a long-standing problem with a large body of literature [1, 6, 7, 8, 9, 10, 12, 19, 19, 20, 21, 22, 23, 24, 25, 26]. The recent convergence of deep

[1] A small percentage of images have duplicates and thus have multiple pairs.

[2] Project website: `http://www-personal.umich.edu/~wfchen/depth-in-the-wild`.

Figure 2: Example images from current RGB-D datasets and our Depth in the Wild (DIW) dataset.

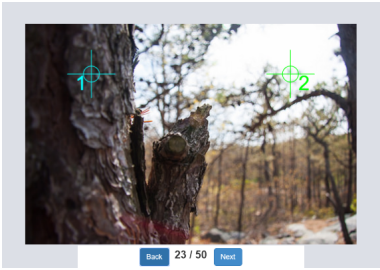

Figure 3: Annotation UI. The user presses '1' or '2' to pick the closer point.

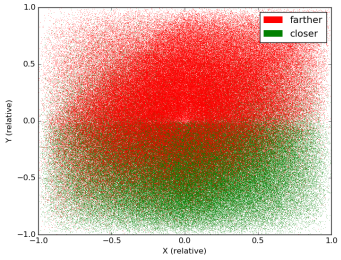

Figure 4: Relative image location (normalized to [-1,1]) and relative depth of two random points.

neural networks and RGB-D datasets [4, 5] has led to major advances [6, 8, 10, 14, 27, 28]. But the networks in these previous works, with the exception of [14], were trained exclusively using ground-truth metric depth, whereas our approach uses relative depth.

Our work is inspired by that of Zoran et al. [14], which proposes to use a deep network to repeatedly classify pairs of points sampled based on superpixel segmentation, and to reconstruct per-pixel metric depth by solving an additional optimization problem. Our approach is different: it consists of a single deep network trained end-to-end that directly predicts per-pixel metric depth; there is no intermediate classification of ordinal relations and as a result no optimization needed to resolve inconsistencies.

Learning with Ordinal Relations: Several recent works [29, 30] have used the ordinal relations from the Intrinsic Images in the Wild dataset [18] to estimate surface refletance. Similar to Zoran et al. [14], Zhou et al. [29] first learn a deep network to classify the ordinal relations between pairs of points and then make them globally consistent through energy minimization.

Narihira et al. [30] learn a "lightness potential" network that takes an image patch and predicts the metric reflectance of the center pixel. But this network is applied to only a sparse set of pixels. Although in principle this lightness potential network can be applied to every pixel to produce pixel-wise reflectance, doing so would be quite expensive. Making it fully convolutional (as the authors mentioned in [30]) only solves it partially: as long as the lightness potential network has downsampling layers, which is the case in [30], the final output will be downsampled accordingly. Additional resolution augmentation (such as the "shift and stitch" approach [31]) is thus needed. In contrast, our approach completely avoids such issues and directly outputs pixel-wise estimates.

Beyond intrinsic images, ordinal relations have been used widely in computer vision and machine learning, including object recognition [32] and learning to rank [33, 34].

3 Dataset construction

We gather images from Flickr. We use random query keywords sampled from an English dictionary and exclude artificial images such as drawings and clip arts. To collect annotations of relative depth, we present a crowd worker an image and two highlighted points (Fig. 3), and ask "which point is closer, point 1, point 2, or hard to tell?" The worker presses a key to respond.

How Many Pairs? How many pairs of points should we query per image? We sample just one per image because this maximizes the amount of information from human annotators. Consider the other extreme—querying all possible pairs of points in the same image. This is wasteful because pairs of points in close proximity are likely to have the same relative depth. In other words, querying one

从Flickr上爬去照片，然后根据字典采样并排除人工图像，人工对图像中两个高亮点进行远近标注
每张图像只标注一对点，因为标注所有点太多余，太接近的点对很可能有相同的相对深度

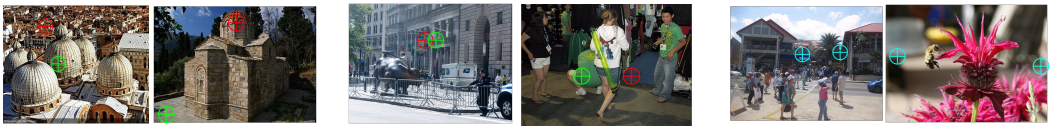

Figure 5: Example images and annotations. Green points are those annotated as closer in depth.

more pair from the same image may add less information than querying one more pair from a new image. Thus querying only one pair per image is more cost-effective.

标注一对点就够了

Which Pairs? Which two points should we query given an image? The simplest way would be to sample two random points from the 2D plane. But this results in a severe bias that can be easily exploited: if an algorithm simply classifies the lower point in the image to be closer in depth, it will agree with humans 85.8% of the time (Fig. 4). Although this bias is natural, it makes the dataset less useful as a benchmark.

An alternative is to sample two points uniformly from a random horizontal line, which makes it impossible to use the y image coordinate as a cue. But we find yet another bias: if an algorithm simply classifies the point closer to the center of the image to be closer in depth, it will agree with humans 71.4% of the time. This leads to a third approach: uniformly sample two *symmetric* points with respect to the center from a random horizontal line (the middle column of Fig. 5). With the symmetry enforced, we are not able to find a simple yet effective rule based purely on image coordinates: the left point is almost equally likely (50.03%) to be closer than the right one.

Our final dataset consists of a roughly 50-50 combination of unconstrained pairs and symmetric pairs, which strikes a balance between the need for representing natural scene statistics and the need for performance differentiation.

Protocol and Results: We crowdsource the annotations using Amazon Mechanical Turk (AMT). To remove spammers, we insert into all tasks gold-standard images verified by ourselves, and reject workers whose accumulative accuracy on the gold-standard images is below 85%. We assign each query (an image and a point pair) to two workers, and add the query to our dataset if both workers can tell the relative depth and agree with each other; otherwise the query is discarded. Under this protocol, the chance of adding a wrong answer to our dataset is less than 1% as measured on the gold-standard images.

We processed 1.24M images on AMT and obtained 0.5M valid answers (both workers can tell the relative depth and agree with each other). Among the valid answers, 261K are for unconstrained pairs and 240K are for symmetric pairs. For unconstrained pairs, It takes a median of 3.4 seconds for a worker to decide, and two workers agree on the relative depth 52% of the time; for symmetric pairs, the numbers are 3.8s and 32%. These numbers suggest that the symmetric pairs are indeed harder. Fig. 5 presents examples of different kinds of queries.

4 Learning with relative depth

学习超像素中心的序关系
然后调和使得能量最小化
然后插值每个超像素得到
密集的逐像素深度

How do we learn to predict metric depth given only annotations of relative depth? Zoran et al. [14] first learn a classifier to predict ordinal relations between centers of superpixels, and then reconcile the relations to recover depth using energy minimization, and then interpolate within each superpixel to produce per-pixel depth.

网络代替映射：
1 输出与输入大小一致
2 用相对深度训练

We take a simpler approach. The idea is that any image-to-depth algorithm would have to compute a function that maps an image to pixel-wise depth. Why not represent this function as a neural network and learn it from end to end? We just need two ingredients: (1) a network design that outputs the same resolution as the input, and (2) a way to train the network with annotations of relative depth.

Network Design: Networks that output the same resolution as the input are aplenty, including the recent designs for depth estimation [8, 35] and those for semantic segmentation [36] and edge detection [37]. A common element is processing and passing information across multiple scales.

多尺度信息

In this work, we use a variant of the recently introduced "hourglass" network (Fig. 6), which has been used to achieve state-of-the-art results on human pose estimation [38]. It consists of a series

采用沙漏结构网络，在人体姿态估计中表现最好，包含一系列卷积操作，或者inception模块

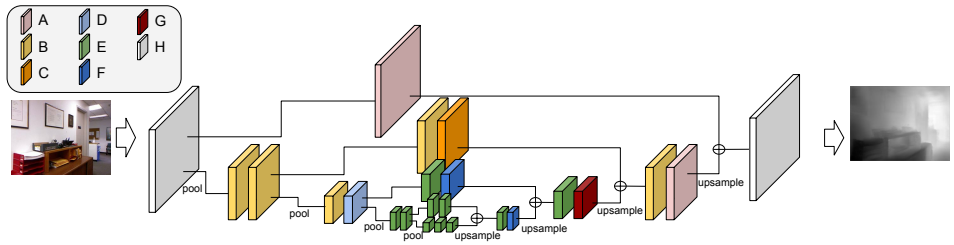

Figure 6: Network design. Each block represents a layer. Blocks sharing the same color are identical. The $\oplus$ sign denotes the element-wise addition. Block H is a convolution with 3x3 filter. All other blocks denote the Inception module shown in Figure 7. Their parameters are detailed in Tab. 1

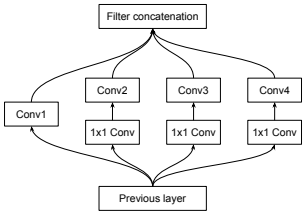

Figure 7: Variant of Inception Module [39] used by us.

Table 1: Parameters for each type of layer in our network. *Conv1* to *Conv4* are sizes of the filters used in the components of Inception module shown in Figure.7. Conv2 to 4 share the same number of input and is specified in *Inter Dim*.

Block Id	A	B	C	D	E	F	G
#In/#Out	128/64	128/128	128/128	128/256	256/256	256/256	256/128
Inter Dim	64	32	64	32	32	64	32
Conv1	1x1	1x1	1x1	1x1	1x1	1x1	1x1
Conv2	3x3	3x3	3x3	3x3	3x3	3x3	3x3
Conv3	7x7	5x5	7x7	5x5	5x5	7x7	5x5
Conv4	11x11	7x7	11x11	7x7	7x7	11x11	7x7

of convolutions (using a variant of the inception [39] module) and downsampling, followed by a series of convolutions and upsampling, interleaved with skip connections that add back features from high resolutions. The symmetric shape of the network resembles a "hourglass", hence the name. We refer the reader to [38] for comparing the design to related work. For our purpose, this particular choice is not essential, as the various designs mainly differ in how information from different scales is dispersed and aggregated, and it is possible that all of them can work equally well for our task.

采用这种结构就是为了多尺度信息的扩散与聚集

Loss Function: How do we train the network using only ordinal annotations? All we need is a loss function that encourages the predicted depth map to agree with the ground-truth ordinal relations. Specifically, consider a training image I and its K queries $R = \{(i_k, j_k, r_k)\}, k = 1, \ldots, K$, where i_k is the location of the first point in the k-th query, j_k is the location of the second point in the k-th query, and $r_k \in \{+1, -1, 0\}$ is the ground-truth depth relation between i_k and j_k: closer $(+1)$, further (-1), and equal (0). Let z be the predicted depth map and z_{i_k}, z_{j_k} be the depths at point i_k and j_k. We define a loss function

i，j是两个点，r表示远近1近，-1远，0相等

$$L(I, R, z) = \sum_{k=1}^{K} \psi_k(I, i_k, j_k, r, z), \tag{1}$$

where $\psi_k(I, i_k, j_k, z)$ is the loss for the k-th query loss函数的远近定义反了

$$\psi_k(I, i_k, j_k, z) = \begin{cases} \log\left(1 + \exp(-z_{i_k} + z_{j_k})\right), & r_k = +1 \\ \log\left(1 + \exp(z_{i_k} - z_{j_k})\right), & r_k = -1 \\ (z_{i_k} - z_{j_k})^2, & r_k = 0. \end{cases} \tag{2}$$

This is essentially a ranking loss: it encourages a small difference between depths if the ground-truth relation is equality; otherwise it encourages a large difference.

gt相等时差别很小，gt相差很大是差别也很大

Novelty of Our Approach: Our novelty lies in the combination of a deep network that does pixel-wise prediction and a ranking loss placed on the pixel-wise prediction. A deep network that does pixel-wise prediction is not new, nor is a ranking loss. But to the best of our knowledge, such a combination has not been proposed before, and in particular not for estimating depth.

第一次将深度估计像素级预测与ranking损失结合起来

5 Experiments on NYU Depth

We evaluate our method using NYU Depth [4], which consists of indoor scenes with ground-truth Kinect depth. We use the same setup as that of Zoran et al. [14]: point pairs are sampled from the

Figure 8: Qualitative results on NYU Depth by our method, the method of Eigen et al. [8], and the method of Zoran et al. [14]. All depth maps except ours are directly from [14]. More results are in the Appendix.

training images (the subset of NYU Depth consisting of 795 images with semantic labels) using superpixel segmentation and their ground-truth ordinal relations are generated by comparing the ground-truth Kinect depth; the same procedure is applied to the test set to generate the point pairs for evaluation (around 3K pairs per image). We use the same training and test data as Zoran et al. [14].

Table 2: Left table: ordinal error measures (disagreement rate with ground-truth depth ordering) on NYU Depth. Right able: metric error measures on NYU Depth. Details for each metric can be found in [8]. There are two versions of results by Eigen et al. [8], one using AlexNet (Eigen(A)) and one using VGGNet (Eigen(V)). Lower is better for all error measures.

Method	WKDR	$\text{WKDR}^{=}$	$\text{WKDR}^{\neq}$
Ours	**35.6%**	**36.1%**	**36.5%**
Zoran [14]	43.5%	44.2%	41.4%
rand_12K	**34.9%**	**32.4%**	**37.6%**
rand_6K	36.1%	32.2%	39.9%
rand_3K	35.8%	28.7%	41.3%
Ours_Full	**28.3%**	**30.6%**	**28.6%**
Eigen(A) [8]	37.5%	46.9%	32.7%
Eigen(V) [8]	34.0%	43.3%	29.6%

Method	RMSE	RMSE (log)	RMSE [a] (s.inv)	absrel	sqrrel
Ours	1.13	0.39	0.26	0.36	0.46
Ours_Full	1.10	0.38	0.24	0.34	0.42
Zoran [14]	1.20	0.42	-	0.40	0.54
Eigen(A) [8]	0.75	0.26	0.20	0.21	0.19
Eigen(V) [8]	**0.64**	**0.21**	**0.17**	**0.16**	**0.12**
Wang [28]	0.75	-	-	0.22	-
Liu [6]	0.82	-	-	0.23	-
Li [10]	0.82	-	-	0.23	-
Karsch [1]	1.20	-	-	0.35	-
Baig [40]	1.0	-	-	0.3	-

As the system by Zoran et al. [14], our network predicts one of the three ordinal relations on the test pairs: equal ($=$), closer ($<$), or farther ($>$). We report WKDR, the weighted disagreement rate between the predicted ordinal relations and ground-truth ordinal relations [3]. We also report $\text{WKDR}^{=}$ (disagreement rate on pairs whose ground-truth relations are $=$) and $\text{WKDR}^{\neq}$ (disagreement rate on pairs whose ground-truth relations are $<$ or $>$).

Since two ground-truth depths are almost never exactly the same, there needs to be a relaxed definition of equality. Zoran et al. [14] define two points to have equal depths if the ratio between their ground-truth depths is within a pre-determined range. Our network predicts an equality relation if the depth difference is smaller than a threshold τ. The choice of this threshold will result in different values for the error metrics (WKDR, $\text{WKDR}^{=}$, $\text{WKDR}^{\neq}$): if τ is too small, most pairs will be predicted to be unequal and the error metric on equality relations ($\text{WKDR}^{=}$) will be large; if τ is too big, most pairs will be predicted to be equal and the error metric on inequality relations ($\text{WKDR}^{\neq}$) will be large. We choose the threshold τ that minimizes the maximum of the three error metrics on a validation set held out from the training set. Tab. 2 compares our network (*ours*) versus that of Zoran et al. [14]. Our network is trained with the same data [4] but outperforms [14] on all three metrics.

Following [14], we also compare with the state-of-art image-to-depth system by Eigen et al. [8], which is trained on pixel-wise ground-truth metric depth from the full NYU Depth training set (220K images). To compare fairly, we give our network access to the full NYU Depth training set. In addition, we remove the limit of 800 point pairs per training image placed by Zoran et al and use all available pairs. The results in Tab. 2 show that our network (*ours_full*) achieves superior performance in estimating depth ordering. Granted, this comparison is not entirely fair because [8] is not optimized for predicting ordinal relations. But this comparison is still significant in that it shows

[a] Computed using our own implementation based on the definition given in [35].

[3] WKDR stands for "Weighted Kinect Disagreement Rate"; the weight is set to 1 as in [14]

[4] The code released by Zoran et al. [14] indicates that they train with a random subset of 800 pairs per image instead of all the pairs. We follow the same procedure and only use a random subset of 800 pairs per image.

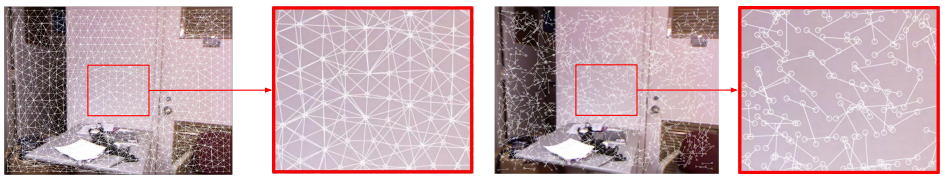

Figure 9: Point pairs generated through superpixel segmentation [14] (left) versus point pairs generated through random sampling with distance constraints (right).

that we can train on only relative depth and rival the state-of-the-art system in estimating depth up to monotonic transformations.

In Figure. 8 we show qualitative results on the same example images used by Zoran et al. [14]. We see that although imperfect, the recovered metric depth by our method is overall reasonable and qualitatively similar to that by the state-of-art system [8] trained on ground-truth metric depth.

Metric Error Measures. Our network is trained with relative depth, so it is unsurprising that it does well in estimating depth up to ordering. But how good is the estimated depth in terms of metric error? We thus evaluate conventional error measures such as RMSE (the root mean squared error), which compares the absolute depth values to the ground truths. Because our network is trained only on relative depth and does not know the range of the ground-truth depth values, to make these error measures meaningful we normalize the depth predicted by our network such that the mean and standard deviation are the same as those of the mean depth map of the training set. Tab. 2 reports the results. We see that under these metric error measures our network still outperforms the method of Zoran et al. [14]. In addition, while our metric error is worse than the current state-of-the-art, it is comparable to some of the earlier methods (e.g. [1]) that have access to ground-truth metric depth.

Superpixel Sampling versus Random Sampling. To compare with the method by Zoran et al. [14], we train our network using the same point pairs, which are pairs of centers of superpixels (Fig. 9). But is superpixel segmentation necessary? That is, can we simply train with randomly sampled points?

To answer this question, we train our network with randomly sampled points. We constrain the distance between the two points to be between 13 and 19 pixels (out of a 320×240 image) such that the distance is similar to that between the centers of neighboring superpixels. The results are included in Tab. 2. We see that using 3.3k pairs per image (***rand_3K***) already achieves comparable performance to the method by Zoran et al. [14]. Using twice or four times as many pairs (***rand_6K***, ***rand_12K***) further improves performance and significantly outperforms [14].

It is worth noting that in all these experiments the test pairs are still from superpixels, so training on random pairs incurs a mismatch between training and testing distributions. Yet we can still achieve comparable performance despite this mismatch. This shows that our method can indeed operate without superpixel segmentation. 不用超像素分割也可以工作

6 Experiments on Depth in the Wild

In this section we experiment on our new Depth in the Wild (DIW) dataset. We split the dataset into 421K training images and 74K test images [5].

We report the WHDR (Weighted Human Disagreement Rate) [6] of 5 methods in Tab. 3: (1) the state-of-the-art system by Eigen et al. [8] trained on full NYU Depth; (2) our network trained on full NYU Depth (Ours_Full); (3) our network pre-trained on full NYU Depth and fine-tuned on DIW (Ours_NYU_DIW); (4) our network trained from scratch on DIW (Ours_DIW); (5) a baseline method that uses only the location of the query points: classify the lower point to be closer or guess randomly if the two points are at the same height (Query_Location_Only).

We see that the best result is achieved by pre-training on NYU Depth and fine-tuning on DIW. Training only on NYU Depth (Ours_NYU and Eigen) does not work as well, which is expected because NYU Depth only has indoor scenes. Training from scratch on DIW achieves slightly better performance

[5] 4.38% of images are duplicates downloaded using different query keywords and have more than one pairs of points. We have removed test images that have duplicates in the training set.

[6] All weights are 1. A pair of points can only have two possible ordinal relations (farther or closer) for DIW.

Figure 10: Qualitative results on our Depth in the Wild (DIW) dataset by our method and the method of Eigen et al. [8]. More results are in the Appendix.

Table 3: Weighted Human Disagreement Rate (WHDR) of various methods on our DIW dataset, including Eigen(V), the method of Eigen et al. [8] (VGGNet [41] version)

Method	Eigen(V) [8]	Ours_Full	Ours_NYU_DIW	Ours_DIW	Query_Location_Only
WHDR	25.70%	31.31%	**14.39%**	22.14%	31.37%

than those trained on only NYU Depth despite using much less supervision. Pre-training on NYU Depth and fine-tuning on DIW leaverages all available data and achieves the best performance. As shown in Fig. 10, the quality of predicted depth is notably better with fine-tuning on DIW, especially for outdoor scenes. These results suggest that it is promising to combine existing RGB-D data and crowdsourced annotations to advance the state-of-the art in single-image depth estimation.

7 Conclusions

We have studied single-image depth perception in the wild, recovering depth from a single image taken in unconstrained settings. We have introduced a new dataset consisting of images in the wild annotated with relative depth and proposed a new algorithm that learns to estimate metric depth supervised by relative depth. We have shown that our algorithm outperforms prior art and our algorithm, combined with existing RGB-D data and our new relative depth annotations, significantly improves single-image depth perception in the wild.

Acknowledgments

This work is partially supported by the National Science Foundation under Grant No. 1617767.

References

[1] K. Karsch, C. Liu, and S. B. Kang, "Depthtransfer: Depth extraction from video using non-parametric sampling," *TPAMI*, 2014.

[2] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," *TOG*, 2005.

[3] A. Saxena, M. Sun, and A. Ng, "Make3d: Learning 3d scene structure from a single still image," *TPAMI*, 2009.

[4] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, Springer, 2012.

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, p. 0278364913491297, 2013.

[6] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *CVPR*, 2015.

[7] L. Ladicky, J. Shi, and M. Pollefeys, "Pulling things out of perspective," in *CVPR*, IEEE, 2014.

[8] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *ICCV*, 2015.

[9] M. H. Baig and L. Torresani, "Coupled depth learning," *arXiv preprint arXiv:1501.04537*, 2015.

[10] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs," in *CVPR*, 2015.

[11] W. W.-C. Chiu, U. Blanke, and M. Fritz, "Improving the kinect by cross-modal stereo.," in *BMVC*, 2011.

[12] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *NIPS*, 2005.

[13] J. T. Todd and J. F. Norman, "The visual perception of 3-d shape from multiple cues: Are observers capable of perceiving metric structure?," *Perception & Psychophysics*, pp. 31–47, 2003.

[14] D. Zoran, P. Isola, D. Krishnan, and W. T. Freeman, "Learning ordinal relationships for mid-level vision," in *ICCV*, 2015.

[15] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell, "A category-level 3d object dataset: Putting the kinect to work," in *Consumer Depth Cameras for Computer Vision*, Springer, 2013.

[16] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *CVPR*, 2015.

[17] S. Choi, Q.-Y. Zhou, S. Miller, and V. Koltun, "A large dataset of object scans," *arXiv preprint arXiv:1602.02481*, 2016.

[18] S. Bell, K. Bala, and N. Snavely, "Intrinsic images in the wild," *TOG*, 2014.

[19] J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading," *TPAMI*, 2015.

[20] A. Saxena, S. H. Chung, and A. Y. Ng, "3-d depth reconstruction from a single still image," *IJCV*, 2008.

[21] Y. Xiong, A. Chakrabarti, R. Basri, S. J. Gortler, D. W. Jacobs, and T. Zickler, "From shading to local shape," *TPAMI*, 2015.

[22] C. Hane, L. Ladicky, and M. Pollefeys, "Direction matters: Depth estimation with a surface normal classifier," in *CVPR*, 2015.

[23] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *CVPR*, 2010.

[24] E. Shelhamer, J. Barron, and T. Darrell, "Scene intrinsics and depth from a single image," in *ICCV Workshops*, 2015.

[25] J. Shi, X. Tao, L. Xu, and J. Jia, "Break ames room illusion: depth from general single images," *TOG*, 2015.

[26] W. Zhuo, M. Salzmann, X. He, and M. Liu, "Indoor scene structure analysis for single image depth estimation," in *CVPR*, 2015.

[27] Z. Zhang, A. G. Schwing, S. Fidler, and R. Urtasun, "Monocular object instance segmentation and depth ordering with cnns," in *ICCV*, 2015.

[28] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," in *CVPR*, 2015.

[29] T. Zhou, P. Krahenbuhl, and A. A. Efros, "Learning data-driven reflectance priors for intrinsic image decomposition," in *ICCV*, 2015.

[30] T. Narihira, M. Maire, and S. X. Yu, "Learning lightness from human judgement on relative reflectance," in *CVPR*, IEEE, 2015.

[31] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[32] D. Parikh and K. Grauman, "Relative attributes," in *ICCV*, IEEE, 2011.

[33] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *ICML*, ACM, 2007.

[34] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2002.

[35] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *NIPS*, 2014.

[36] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[37] S. Xie and Z. Tu, "Holistically-nested edge detection," *CoRR*, vol. abs/1504.06375, 2015.

[38] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *arXiv preprint arXiv:1603.06937*, 2016.

[39] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[40] M. H. Baig, V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, and N. Sundaresan, "Im2depth: Scalable exemplar based depth transfer," in *WACV*, IEEE, 2014.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

Appendix

Figure 11: Additional example images from current RGB-D datasets and our Depth in the Wild (DIW) dataset.

Figure 12: Additional qualitative results on NYU Depth by our method, the method of Eigen et al. [8], and the method of Zoran et al. [14]. All depth maps except ours are directly from [14].

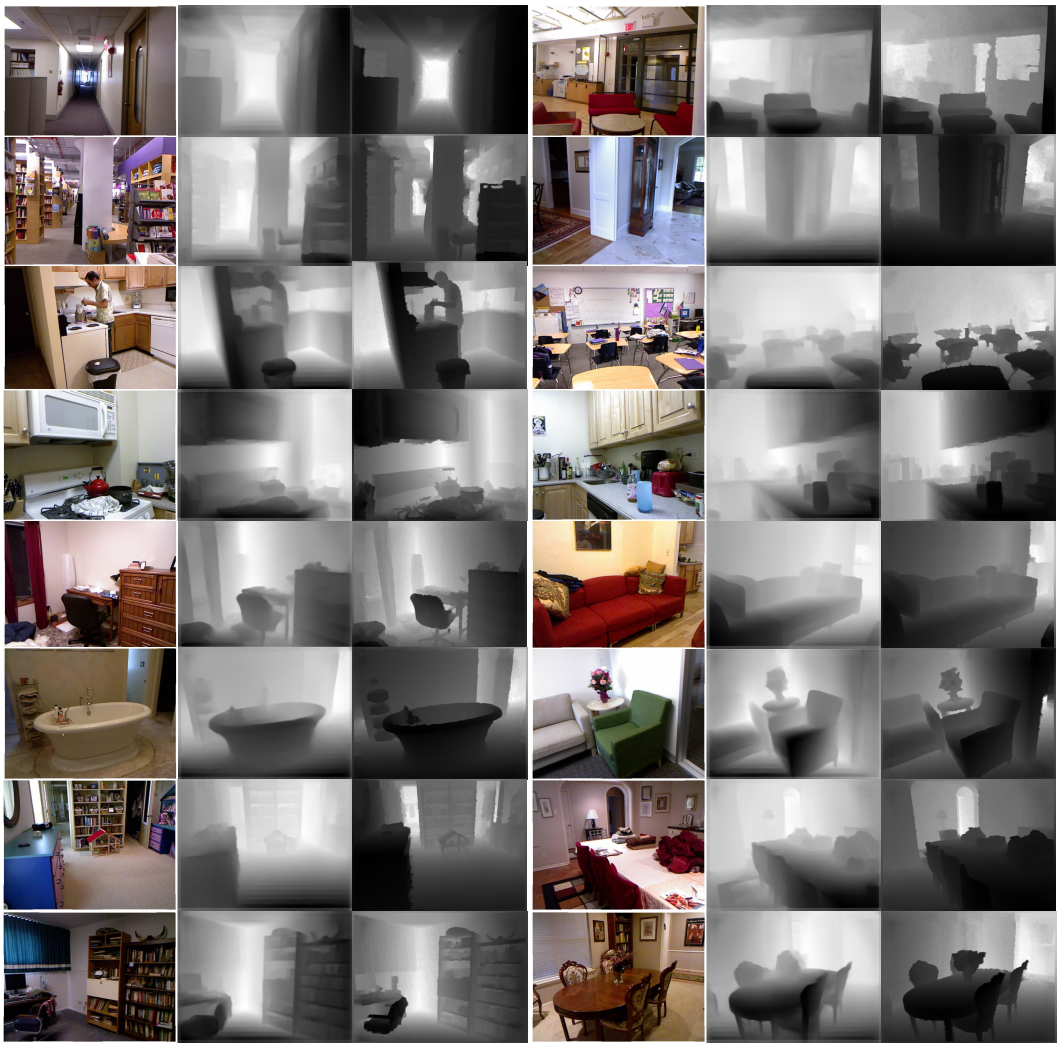

Figure 13: Additional qualitative results on NYU Depth test set by our method. Here we show the original input images and the depth maps by our method, as well as the ground truth.

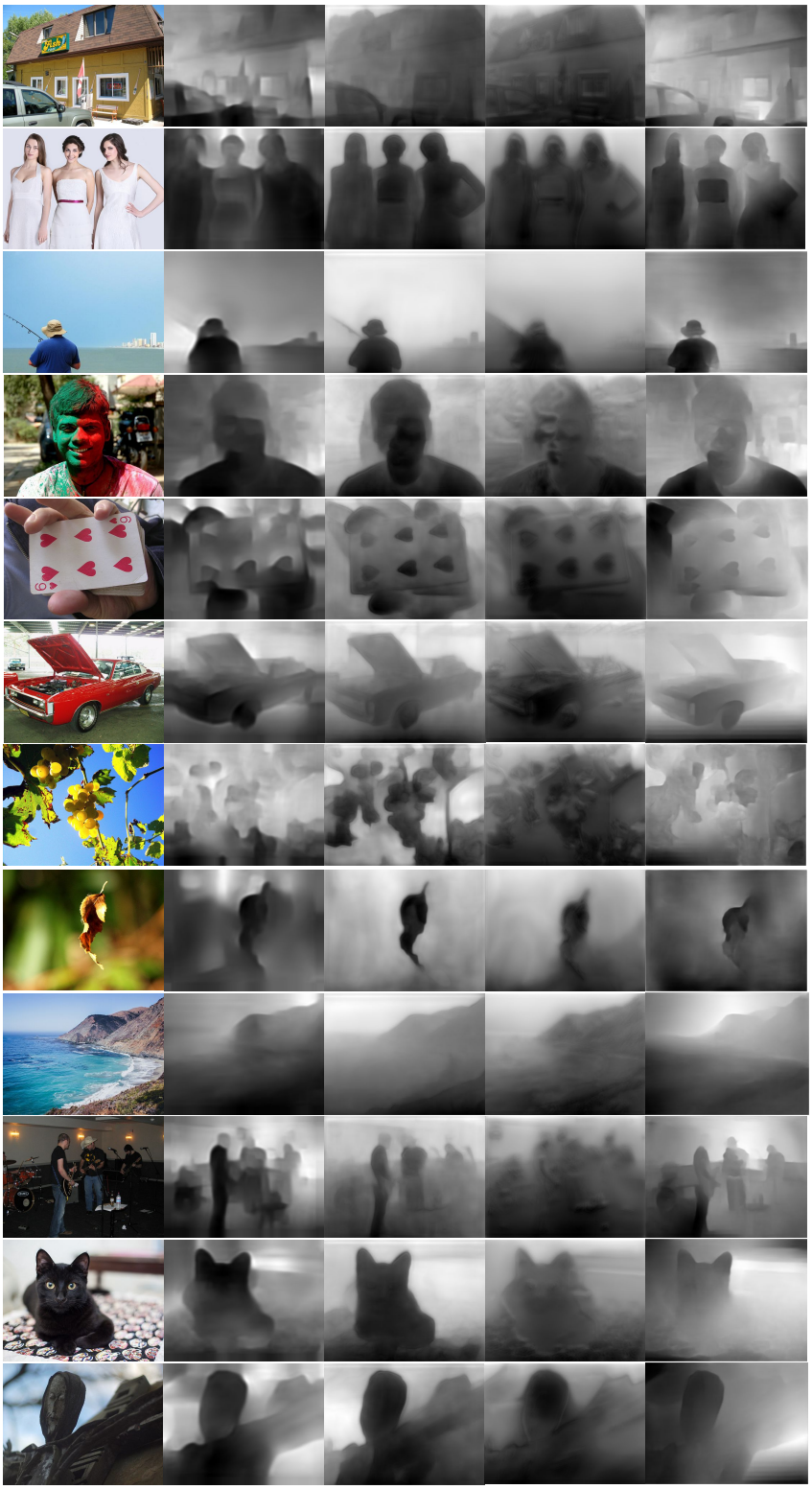

Figure 14: Additional qualitative results on our Depth in the Wild (DIW) dataset by our method and the method of Eigen et al. [8].

Table 4: Table 2 of the main paper reports the metric error of our network trained on relative depth pairs. Here we provide additional results by training our network on the full depth map. The network **Ours** is our network trained on the 795 NYU Depth training subset, and **Ours_Full** is our network trained on the full NYU Depth training set.

Method	RMSE	RMSE (log)	RMSE (s.inv)	absrel	sqrrel
Ours	0.89	0.32	0.25	0.27	0.29
Ours_Full	0.74	0.26	0.21	0.21	0.19
Eigen(V) [8]	**0.64**	**0.21**	**0.17**	**0.16**	**0.12**

Table 5: Table 2 of the main paper reports the performance of our network versus the number of randomly sampled non-superpixel point pairs on NYU Depth. Here we report additional results by sampling more pairs. $\mathbf{rand\_N}$ denotes a network trained with N pairs per image.

Method	WKDR	$\text{WKDR}^{=}$	$\text{WKDR}^{\neq}$
rand_48K	**34.3%**	**31.7%**	**37.1%**
rand_24K	34.5%	32.6%	36.9%
rand_12K	34.9%	32.4%	37.6%
rand_6K	36.1%	32.2%	39.9%
rand_3K	35.8%	28.7%	41.3%

Table 6: Table 2 of the main paper reports the performance of our network trained on 800 superpixel point pairs. Here we report additional results by decreasing the number of point pairs.

#Depth Pairs	WKDR	$\text{WKDR}^{=}$	$\text{WKDR}^{\neq}$
800	**35.6%**	**36.1%**	**36.5%**
500	37.2%	37.7%	38.2%
250	38.0%	37.4%	39.7%
100	42.3%	41.1%	44.0%