Structure-Guided Ranking Loss for Single Image Depth Prediction

Ke Xian[1], Jianming Zhang[2], Oliver Wang[2], Long Mai[2], Zhe Lin[2], and Zhiguo Cao[1]*
[1]National Key Laboratory of Science and Technology on MultiSpectral Information Processing,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China
[2]Adobe Research
{kexian, zgcao}@hust.edu.cn, {jianmzha, owang, malong, zlin}@adobe.com
https://github.com/KexianHust/Structure-Guided-Ranking-Loss

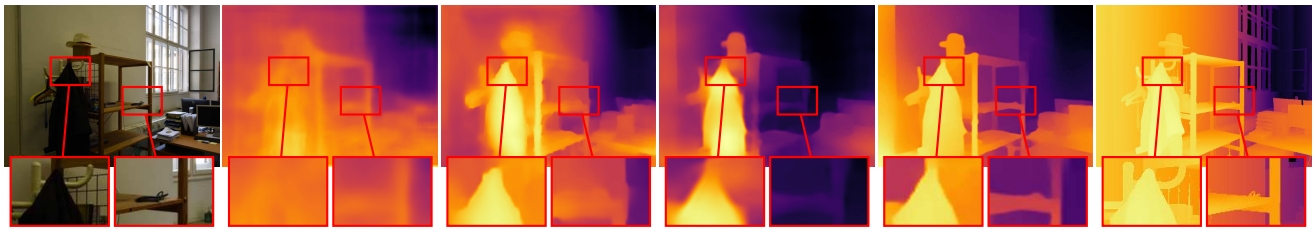


Figure 1. Qualitative results from state-of-the-art models Y3D [4] and MC [21], our baseline model trained using the affine-invariant loss proposed by [19] and the same model trained using our proposed structure-guided ranking loss. The model trained using our loss provides more details of local depth structure and higher accuracy at depth boundaries. The test image is from Ibims [16], which is not used in the training for any of the above models.

Abstract

Single image depth prediction is a challenging task due to its ill-posed nature and challenges with capturing ground truth for supervision. Large-scale disparity data generated from stereo photos and 3D videos is a promising source of supervision, however, such disparity data can only approximate the inverse ground truth depth up to an affine transformation. To more effectively learn from such pseudo-depth data, we propose to use a simple pair-wise ranking loss with a novel sampling strategy. Instead of randomly sampling point pairs, we guide the sampling to better characterize structure of important regions based on the low-level edge maps and high-level object instance masks. We show that the pair-wise ranking loss, combined with our structure-guided sampling strategies, can significantly improve the quality of depth map prediction. In addition, we introduce a new relative depth dataset of about 21K diverse high-resolution web stereo photos to enhance the generalization ability of our model. In experiments, we conduct cross-dataset evaluation on six benchmark datasets and show that our method consistently improves over the baselines, leading to superior quantitative and qualitative results.

1. Introduction

Monocular depth prediction (monodepth) is a fundamental task in computer vision, and can be used in many real-world applications. Due to its ill-posed nature, monodepth critically relies on scene semantics and thus requires a diverse set of training data to ensure its generalization ability to unseen content.

Traditional depth supervision datasets are captured by either active or passive depth sensors, and as such are generally restricted to a single domain or scene type, e.g., road [11] or indoor [29]. To improve data diversity, recent monodepth works [36] have used large-scale disparity data generated from web stereo photos and 3D movies [19, 33]. However, both stereo photos and 3D movies may have been post-edited to optimize for viewing experience. For example, a common technique for changing the stereo viewing experience is through the positioning the stereo window by adjusting the virtual baseline and minimum disparity[1]. Therefore, their disparity data only approximate the inverse ground truth depth up to an affine transformation. Although affine-invariant losses have been proposed to address this issue [10, 19, 33], we find that using such loss function often leads to sub-optimal results with blurry depth boundaries

*Corresponding author.

[1]http://www.shortcourses.com/stereo/stereo3-11.html

and missing details (see Fig. 1).

To effectively learn from such pseudo-depth data, we revisit the pair-wise ranking loss used in [3, 36]. The ranking loss only depends on the depth ordinal relationship (*e.g.* point A is in front of point B, or vice versa), and thus is applicable to pseudo-depth data from various sources [3, 10, 21, 22, 36]. Compared to pixel-wise regression losses, ranking losses penalize the wrong pairwise ordinal prediction between a sparse set of pixel pairs. We observe that *how* point pairs are sampled can have a big impact on model's performance.

The sampling space of the point pairs is large, but only a small set of point pairs contains important constraints to characterize the salient structure of the depth map, *e.g.* the location of depth boundaries. Therefore, the random sampling scheme, employed in prior work [3, 36], spends a lot of training computation on uninformative point pairs. In addition, depending on the application, accuracy in certain regions can be of significantly higher importance. For example, one prominent source of errors in depth maps, is inconsistent depth prediction in salient object instances like human. When part of a human is "cut-off" in the depth map, striking visual artifacts will appear in downstream applications, such as shallow DoF rendering and view synthesis.

Motivated by these observations, we propose structure-guided ranking loss which employs two carefully crafted sampling strategies: Edge-Guided Sampling and Instance-Guided Sampling. Edge-guided sampling focuses on point pairs that characterize the location of the depth boundaries and suppress false depth boundaries caused by strong image edges. Instance-guided sampling is intended for improving depth structural accuracy regarding salient object instances. We show that this structure-guided ranking loss can produce higher quality depth maps than baseline losses. Based on the proposed ranking loss, we train our model on a newly collected large-scale web stereo photo dataset of about 21K diverse photos. Our model achieves superior cross-dataset generalization performance on six benchmark datasets compared to state-of-the-art methods and baselines, showing the effectiveness of the structure-guided ranking loss on stereo depth data.

Our main contributions are summarized as follows.

- We propose a structure-guided ranking loss formulation with two novel sampling strategies for monocular depth prediction.
- We introduce a large relative depth dataset of about 21K high resolution web stereo photos.
- With our proposed loss and the new dataset, our model achieves state-of-the-art cross-dataset generalization performance.

2. Related Work

Monodepth methods Traditional monodepth methods rely on direct supervision [15, 23, 27] mainly through hand-crafted features, to learn 3D priors from images. In recent years, supervised deep learning models [2, 5, 6, 7, 18, 20, 24, 25, 37, 38, 39] have achieved state-of-the-art performance in the task of monocular metric depth prediction. These methods, trained on RGB-D datasets, learn a mapping function from RGB to depth. Despite the fact that these models can predict accurate depth when testing on the same or similar datasets, they cannot be easily generalized to novel scenes. In addition to these supervised methods, unsupervised or semi-supervised algorithms have also been studied. The key idea behind these methods [9, 12, 17, 35, 40] is the image reconstruction loss for view synthesis, requiring calibrated stereo pairs or video sequences for training. However, these models share the same issue with supervised deep learning based methods. In other words, the model cannot be generalized to new datasets. To address this issue, multiple in-the-wild (*-i.e.*, contains both indoor and outdoor scenes) RGB-D datasets [3, 4, 19, 21, 34, 22, 33, 36] have been proposed.

Monodepth losses Since the depth of these datasets is ambiguous in scale axis, directly using ℓ_1 or ℓ_2 loss functions cannot train a model correctly. There are two alternatives for addressing that problem. One solution is to design a scale-invariant loss [6, 19, 21, 22, 33]. The other way is to design a suitable ranking loss [3, 4, 36], which can be used for training regardless of depth scale. The scale-invariant loss, paying equal attention to each pixel in an image, is prone to generate blurry predictions with details missing. By contrast, the ranking loss computes losses on the selected point pairs, which has potential to generate consistent predictions with sharp depth discontinuities. However, previous ranking losses only computes their loss on pre-defined or randomly sampled pairs of points. This leads to many irrelevant pairs being selected for ranking, which may not be useful for training procedures. We propose a structure-guided ranking loss for training a monodepth model. Rather than pre-defining or random sampling candidate pairs for ranking, we instead perform online sampling guided by the overall structure of the scnee including edges and object instance masks.

Depth datasets Existing RGB-D datasets mainly come from three sources: depth sensors, synthetic data, and Internet images. Depth sensors (*e.g.*, Kinect and laser scanner) are commonly used to acquire accurate metric depth. However, are limited to indoor scenes [29, 30], or sparse reconstructions [27, 31]. Recently, other active sensors have been used to capture ground truth depth [16, 28, 32], however due

to capture restrictions, these datasets consist of mostly rigid objects. All of the above datasets are limited in terms of diversity, and do not generalize to images in-the-wild as one might capture with their mobile phone.

Another source to acquire RGB and depth pairs is synthetic data, *e.g.* [1, 8, 26]. These datasets are free of noise, and can have accurate metric depth with sharp discontinuities, however, there exists a domain gap between synthetic and real data, which necessitates domain adaptation for real world applications.

In order to explore the diversity of the visual world, more and more attention has been paid to the source of Internet images/videos, *e.g.* by; annotated monocular images [3], image collections or video for multi-view reconstructions [4, 21, 22], stereo images [36], or stereo videos [19, 33]. Most similar to ours, `ReDWeb` generates dense relative depth maps from stereo images via computing stereo disparity, but only has 3,600 low-resolution images for training. In contrast to the above datasets, ours consists of a large quantity (20K) of high-resolution, diverse, training images.

3. Structure-Guided Ranking Loss for Monocular Depth Prediction

Given an RGB image I, we hope to learn a function $P = \mathcal{F}(I)$, which generates a single channel depth map P. We propose to train a model from web stereo images, where only derived disparity maps are present for supervision. These disparity maps have an unknown shift and scale factor to relate to (inverse) depth. For the sake of convenience, in the following, we use depth to refer to inverse depth unless mentioned otherwise.

DIW [3] proposes a pair-wise ranking loss to train models on this type of pseudo-depth data. The loss is defined on a sparse set of point pairs with ordinal annotations. Specifically, for a pair of points with predicted depth values $[p_0, p_1]$, the pair-wise ranking loss is

$$\phi(p_0 - p_1) = \begin{cases} \log(1 + \exp(-\ell(p_0 - p_1))), & \ell \neq 0 \\ (p_0 - p_1)^2, & \ell = 0, \end{cases} \quad (1)$$

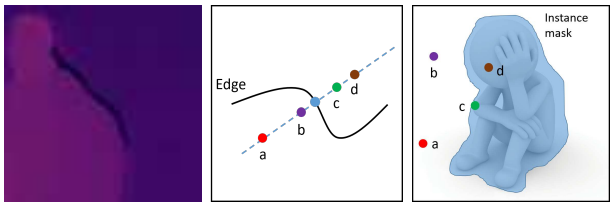
where ℓ is the ground truth ordinal label, which can be induced by a ground truth pseudo-depth map:

$$\ell = \begin{cases} +1, & p_0^*/p_1^* >= 1 + \tau, \\ -1, & p_0^*/p_1^* <= \frac{1}{1+\tau}, \\ 0, & otherwise. \end{cases} \quad (2)$$

Here τ is a tolerance threshold, which is set to 0.03 in our experiments, and p_i^* denotes the ground-truth pseudo-depth value. When the pair of point are close in the depth space, *i.e.*, $\ell_i = 0$, the loss encourages the predicted p_0 and p_1 to be the same; otherwise, the difference between p_0 and p_1 must be large to minimize the loss. Given a set of sampled point pairs $\mathcal{P} = \{[p_{i,0}, p_{i,1}], i = 1, \ldots, N\}$, the overall ranking loss can be defined as follows,

$$\mathcal{L}_{rank}(\mathcal{P}) = \frac{1}{N} \sum_i \phi(p_{i,0} - p_{i,1}). \quad (3)$$

This type of pair-wise ranking losses is very general and thus can be applied to various types of depth and pseudo-depth data. However, how the specific point pairs are sampled can have a big impact on the reconstruction quality. Instead of using random sampling [3, 36], we propose a segment-guided sampling strategy based on the combination of 1) image edges and 2) instance segmentation masks. The goal is to focus the network's attention on the regions that we specifically care about, *i.e.* the salient depth structures of the scene.



(a) Halo artifacts (b) Edge-guided (c) Instance-guided

Figure 2. (a) shows the halo effect along depth boundary region generated by only sampling two points across an image edge; (b) and (c) show the procedure of edge- and instance-guided sampling.

3.1. Edge-Guided Sampling

In general, depth maps follow a strong piece-wise smooth prior. In other words, the depth values change smoothly in most of the regions, except at sharp depth discontinuity that occur in a small portion of the image. Ultimately, getting the correct depth at these discontinuities is critical for most downstream applications. As a result, randomly sampled point pairs waste most of their supervision on unimportant relationships, and depth prediction computed with this strategy often looks blurry and lacks detail.

How can we predict where such depth discontinuities lie? One solution is to concentrate on regions where there are images edges, as most object boundaries exhibit image edges as well. Equally important to successfully predicting depth boundaries at image edges, is *not* predicting depth boundaries at *texture* edges, *e.g.* strong image edges that have no depth change. Again, randomly sampled point pairs are often not helpful in that regard. Therefore, we propose to handle both of these cases by simply sampling points around image edges as much as possible.

One way to do this is to sample local point pairs that reside across an image edge (see Fig. 2 (a)). We find that such sampling has a side effect of over-sharpening the depth

Algorithm 1 The procedure for edge-guided sampling

Require: Edge masks E, gradient maps G_x, G_y and gradient magnitude G, number of edge pixels L to be sampled
Initial: Sampled points $\mathcal{S} = \emptyset$
1: **for** $i = 1, 2, ..., L$ **do**
2: Sample an edge point $\mathbf{e}$
3: Sample 4 points $[(x_k, y_k), k = a, b, c, d]$ according Eqn. 5
4: Add (a, b), (b, c) and (c, d) to $\mathcal{S}$
5: **end for**
6: Return point pair set $\mathcal{S}$

boundaries, leading to halo artifacts along the depth boundaries. Therefore, we propose a 4-point sampling scheme to enforce the smoothness on each side of a depth boundary (see Fig. 2 (b)). The 4 points lie on an orthogonal line crossing a sampled edge point. Within a small distance range of the edge point, we random sample two points on each side, resulting in three pairs of points ($[(a, b), (b, c), (c, d)]$ in Fig. 2 (b)) for our ranking loss.

Given an image, we convert it to an gray-scale image and use the Sobel operator to get the gradient maps G_x and G_y, and the gradient magnitude map G. Then we compute an edge map E by thresholding the gradient magnitude map.

$$E = \mathbb{I}[G \geq \alpha \cdot max(G)], \tag{4}$$

where α is a threshold to control the density of E. For each edge point $\mathbf{e} = (x, y)$ sampled from E, we sample 4 points $[(x_k, y_k), k = a, b, c, d]$ by

$$\begin{cases} x_k = x + \delta_k G_x(\mathbf{e})/G(\mathbf{e}) \\ y_k = y + \delta_k G_y(\mathbf{e})/G(\mathbf{e}). \end{cases} \tag{5}$$

We have $\delta_a < \delta_b < 0 < \delta_c < \delta_d$, and they are sampled within a small distance range β from the edge point $\mathbf{e}$. In experiments, the α and β are set to 0.1 and 30, respectively. To avoid sampling points too near to the edge point $\mathbf{e}$, where the ground truth depth value can be hard to define, we also set a two-pixel margin on each side of the edge. The whole sampling process is summarized in Alg. 1.

3.2. Instance-Guided Sampling

The above approach improves depth predictions around image edges, however, often times such low level cues miss important boundaries, and end up bisecting salient objects, e.g, humans. This almost always leads to strong visual artifacts (*e.g.* a person's head being "cut off") in downstream applications such as view synthesis and shallow DoF rendering. A pure low-level edge-based sampling will still undersample these critical regions. Therefore, we propose an instance-guided sampling strategy to make the ranking loss more sensitive to such salient depth structures.

Rather than leveraging an edge map, we rely on instance segmentation masks as predicted by a network trained on manual segmentation annotations. Similar to edges, we use a 4-point scheme to sample three pairs of points to characterize the depth structure of the object (see Fig. 2 (c)). Specifically, we randomly sample a pair of points outside and inside of mask respectively ((a, b) and (c, d)), and use one point of each pair to form a cross-boundary pair ((b, c)).

We use segmentations generated by Mask R-CNN [13] for this sampling strategy. A common issue of such instance masks is that they also often miss small parts of the object, such as human's head, arms, *etc*. Once those small object parts are not included in the mask, they have much smaller chance to be sampled. Thus, we add an additional dilation operation to expand the instance mask from Mask R-CNN.

3.3. Model Training

Point pair sampling Edge- and instance-guided sampling can greatly enhance the local details of the depth prediction, but we also find that they are not very effective in preserving global structures, such as ground planes, walls, *etc*. Therefore, we combine edge-guided sampling, instance-guided sampling and random sampling to produce the final point pairs for the ranking loss. For edge-guided sampling, we enumerate through all N edge pixels and use the 4-point scheme to sample three pairs, resulting in $3N$ point pairs. Then we sample N random pairs to augment the sample set. Note that, N is image dependent constant because different images have different numbers of intensity edges. For instance-guided sampling, the number of sampled pairs per instance is proportional to the area M of the mask, namely $3M$ based on the 4-point scheme.

Losses To enforce smooth gradients and sharp discontinuities, we follow prior work [22] and add a multi-scale scale-invariant gradient matching loss in inverse depth space. We denote $R_i = p_i - p_i^*$ and define the loss as:

$$\mathcal{L}_{grad} = \frac{1}{M} \sum_s \sum_i (|\nabla_x {R_i}^s| + |\nabla_y {R_i}^s|), \tag{6}$$

where M is the number of pixels with valid ground truth, and R^s represents the difference of disparity maps at scale s. Following [22], we use four scales in our experiments.

To obtain consistent depth with sharp discontinuities, we combine $\mathcal{L}_{rank}$ and $\mathcal{L}_{grad}$ together to supervise the training. Thus, our final loss can be given as:

$$\mathcal{L} = \mathcal{L}_{rank} + \lambda \mathcal{L}_{grad} \tag{7}$$

where λ is a balancing factor, which is set to 0.2 in our experiments.

Model For the model, we use the ResNet50-based network architecture [36] as our backbone model. The network is trained with synchronized stochastic gradient descent (SGD) using default parameters over 4 GPUs. The

batch size is set to 40 (10 images/GPU), and models are trained with ranking loss for 40 epochs with an initial learning rate of 0.02, which is then multiplied by .1 after 20 epochs. During training, images are horizontally flipped with a 50% chance, and resized to 448×448.

4. Dataset

In this section, we introduce the "High-Resolution Web Stereo Image" (`HR-WSI`) dataset (Fig 3), a diverse collection of high-resolution stereo images collected from the web. Similar to prior work [33, 36], we use FlowNet2.0 [14] to generate disparity maps as our ground truth. However, these generated disparity maps contain errors where flow prediction fails, *i.e*. textureless regions and disocclusion parts. To improve the data quality, we use a forward-backwards flow consistency check to mask out outliers which are ignored for training. Furthermore, as sky regions pose a specific challenge, we also compute high-quality sky segmentation masks via a pretrained network (see the supplementary material for details), and set the disparity of sky regions to be the minimum observed value. After manually rejecting bad ground truth data, we are left with 20378 images for training, and 400 images for validation.

Compared to `ReDWeb` [36], our dataset has three advantages: more training samples (20378 vs. 3600), higher resolution (833×1251 vs. 408×465) and better handling of regions of incorrect disparities, *i.e*. sky. For instance segmentation masks, we use Mask-R-CNN that pre-trained on the COCO dataset, and retain all bounding boxes with confidence score ≥ 0.7. More specifically, we keep 15 labels including living body (*e.g*., human, dog, *etc*.) and non-living body (*e.g*., car, chair, and dining table). Our statistical analysis indicates that roughly 45.8% of images of our dataset contain humans, which is starkly different from most depth supervision datasets, and about 72% images contain object instances that we use to drive our sampling.

5. Experiments

In this section, we first evaluate the performance of our model through zero-shot cross-dataset evaluation on six RGB-D benchmarks that were unseen during training. After that, we present ablation studies of our method to demonstrate the benefit of our structure-guided ranking loss.

5.1. Test data

`Ibims` [16] consists of 100 RGB-D pairs from indoor scenes for evaluation. In order to provide a detailed analysis on specific characteristics (*e.g*., depth discontinuities) of depth maps, the dataset provides masks for distinct depth transitions and planar regions.

`TUM` [30] consists of RGB-D images of indoor scenes of people performing different actions. For fair comparisons, we use the same data as [21] for evaluation. In particular, there are 11 image sequences with 1815 images. This dataset provides human masks, so we can also evaluate depth edges around human instances.

`Sintel` [1] is derived from the open source 3D animated short film consisting of 1064 images with accurate ground truth depth maps.

`NYUDv2` [29] is an indoor dataset with depth captured by a Kinect depth sensor. We use the official test split (654 images) for evaluation. The original resolution of each image is 480×640, we follow the previous method [6] to report scores on a pre-defined center cropping.

`KITTI` [31] consists of over 93K outdoor images collected from a car with stereo cameras and Lidar ground truth. In our experiments, we use the commonly used test split (697 images) provided by Eigen *et al*. [6] and only evaluate on pixels with valid ground truth depth.

`DIODE` [32] contains both indoor and outdoor static scenes with accurate ground truth depth. We use the official test set (771 images).

5.2. Metrics

For zero-shot cross-dataset evaluation, we use an ordinal error [3], analogous to Weighted Human Disagreement Rate (WHDR) [41]. The ordinal error can be defined as

$$Ord = \frac{\sum_i \omega_i \mathbb{I}(\ell_i \neq \ell^*_{i,\tau}(p))}{\sum_i \omega_i}, \tag{8}$$

where ω_i is set to 1, and the ordinal relationships ℓ_i and $\ell^*_{i,\tau}(p)$ are computed using Eqn. 2. For each image, we randomly sample 50000 point pairs to compute the ordinal error. This ordinal error is a general metric for evaluating the ordinal accuracy of a depth map, and it can be directly used with difference sources of depth ground truth.

Although our model is mainly trained with the ordinal ranking loss, we also report various metric depth error scores for completeness. Our model gives competitive results under these metrics. Due to limited space, please refer to the supplementary material for details.

5.3. Zero-shot Cross-dataset Evaluation

We report the ordinal error of different models in Table 1 compared with state-of-the-art methods. Note that these models were trained using different datasets and different losses. The ordinal error may not fairly reflect all the aspects of their quality, but it is a meaningful metric to demonstrate their cross-dataset generalization performance.

Compared methods For DIW [3], we use their released model that trained on `DIW` using a ranking loss. RW [36] was trained with a ranking loss on the `RW` dataset which is derived from web stereo images. DL [34] was trained with a

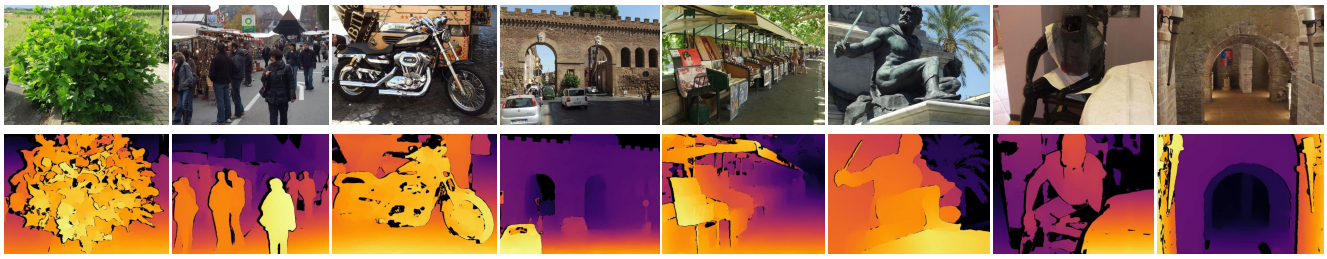


Figure 3. Example images from our (HR-WSI) dataset, consisting of high-resolution stereo images in the wild, and derived disparity maps with consistency checking.

Methods	Training Datasets	Loss	Ibims	TUM	Sintel	NYUDv2	KITTI	DIODE	Avg. Ranking
DIW [3]	DIW	PR	46.97	39.62	43.50	37.33	29.92	45.40	11.00
DL [34]	ID	L1	40.92	31.62	36.63	31.67	25.40	43.77	9.50
RW [36]	RW	PR	33.13	30.07	31.12	26.76	16.40	39.42	6.50
MD [22]	MD	SI+MGM+ROD	36.82	31.88	38.07	27.84	17.50	39.07	8.83
YT3D [4]	RW+DIW+YT3D	PR	31.73	30.37	33.88	26.39	15.08	35.57	5.33
MC [21]	MC	SI+MGM+MES	31.30	26.22	37.49	25.48	22.46	40.85	6.17
MiDaS [19]	RW+MD+MV	AI+MGM	30.09	27.20	**28.35**	25.25	14.73	35.27	2.50
Ours_AI	HRWSI	AI+MGM	32.59	27.82	34.06	27.57	15.99	37.92	6.17
Ours†	RW	SR+MGM	32.29	29.07	30.98	26.93	16.69	38.91	5.83
Ours_R	HRWSI	PR	27.91	27.44	31.89	23.31	14.92	33.24	3.00
Ours	HRWSI	SR+MGM	**27.23**	**25.67**	30.70	**23.21**	**14.01**	**33.11**	**1.17**

Table 1. **Ordinal error (%) of zero-shot cross-dataset evaluation.** Existing monodepth methods were trained on different sources of data: DIW, ReDWeb (RW), MegaDepth (MD), 3D Movies (MV), iPhone Depth (ID), YouTube3D (YT3D), and MannequinChallenge (MC); with different losses: pair-wise ranking loss (PR), affine-invariant MSE loss (AI), multi-scale gradient matching loss (MGM), L1 loss (L1), scale-invariant loss (SI), robust ordinal depth loss (ROD) and multi-scale edge-aware smoothness loss (MES). Our structure-guided ranking loss is denoted as SR. To disentangle the effect of datasets from that of losses, we also evaluate three baseline models: 1) Ours_AI: using the same losses as MiDaS; 2) Ours†: using our final loss on the RW dataset; 3) Ours_R: using the pair-wise ranking loss [36]. To evaluate the robustness of trained models, we compare our models with the state-of-the-art methods on six RGBD datasets that were unseen during training. The lowest error is boldfaced and the second lowest is underlined.

ℓ_1 loss on the iPhone Depth (ID) dataset of about 2k images which was collected by an iPhone camera. MD [22] was trained with a scale-invariant loss and a multi-scale gradient matching loss on a large scale dataset MD which focuses on famous outdoor landmarks. YT3D [4] also combined different sources of data (*i.e.*, RW, DIW, and YT3D) to improve the robustness of the model. They used the original ranking loss for training. MC [21] targets on learning the depth of people, so the model was trained on the MC dataset in which each image contains humans. Since only a single image can be used, we use the single-view model of MC for comparisons. Similar to MD, they used a scale-invariant loss and a a multi-scale gradient matching loss for training. MiDaS [19] was trained using an affine-invariant loss combined with a multi-scale gradient matching term. The model also used much more data by mixing RW, MD, and MV.

As shown in Table 1, MiDaS achieves very competitive generalization performance, but their model was trained on a collection of large-scale datasets. To better compare with their loss functions, we train a baseline model Ours_AI on our proposed dataset using the same loss functions and settings proposed by [19].

Despite the fact that both MiDaS and YT3D mixed different sources of data for training, our model still achieves the best performance in this setting. The only exception is that MiDaS performs slightly better than ours on Sintel dataset. This is likely due to the fact the characteristics of Sintel and MV are similar, since both of them were derived from movie data. MD and MC were trained on datasets of outdoor landmarks and humans, respectively. As a result, MD performs well in outdoor scenes (*e.g.*, KITTI), but falls short in indoor scenes (*e.g.*, Ibims, TUM, and NYUDv2). Similarly, MC generalizes worse on datasets that contain outdoor scenes (*e.g.*, KITTI and DIODE). Compared to these state-of-the-art methods, our final model shows stronger robustness in unconstrained scenes, showing the advantage of our web stereo dataset.

Compared with the baseline models (*i.e.* Ours_AI, Ours†, and Ours_R), our model consistently achieves lower ranking errors. In addition to quantitative comparisons, we also

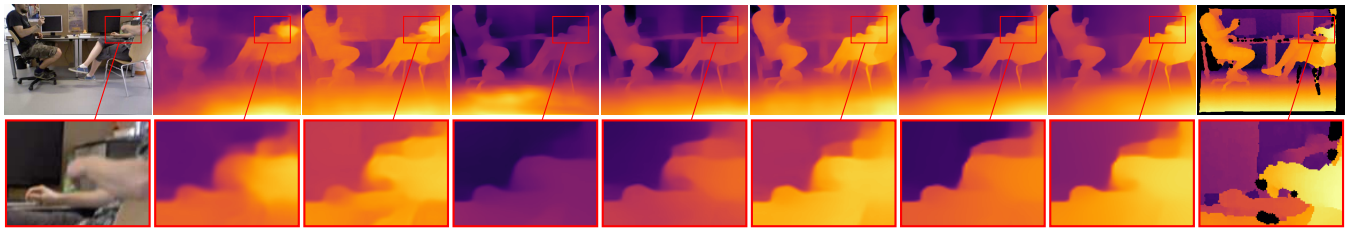


Figure 4. Qualitative evaluation of different sampling strategies and the affine-invariant loss. Best viewed zoomed in on-screen. Our full model trained with a combination of the structure-guide ranking loss and the multi-scale gradient matching loss generates a globally consistent depth map with sharp depth boundaries and detailed depth structures.

demonstrate qualitative results in Fig. 5. One can see that our model can get more consistent depth with sharper depth discontinuities. Benefited from our accurate sky masks, Ours_AI is also able to generate more accurate depth at sky regions. However, the affine-invariant loss does not seem sensitive to local depth structures and the depth boundaries, and thus the predictions are more blurry and lack details.

5.4. Analysis of structure-guided sampling

In order to analyse the effectiveness of structure-guided sampling, we conduct ablation studies on two RGB-D benchmarks. In particular, we analyze the effect of loss functions on the accuracy of depth boundary localization. We use boundary errors (ε_{acc} and ε_{comp}) on `Ibims` according to their definitions [16]. In addition, we follow [21] to measure human-related scale-invariant RMSE (si-hum, si-intra, and si-inter) on `TUM`. We refer the reader to the corresponding papers for the definition of these metrics. Note that, all the compared models are only able to generate relative depth. In order to evaluate these with respect to metric depth ground truth, we align the scale and shift of all predictions to the ground truth before evaluation [19].

Our full model uses a combination of sampling strategies such as random sampling (R), edge-guided sampling (E), instance-guide sampling (I) as well as the multi-scale gradient matching loss term (M). The results of different combinations of these components are shown in Table. 2. Some qualitative results are shown in Fig. 4.

From Table. 2 and Fig. 4, one can observe that our ranking-based models perform better at depth boundaries as well as depth consistency of people. In general, Ours_AI tends to be blurry and lack local depth details comparing with other baselines, but it provide pretty good depth consistency. Similarly, since Ours_R is trained on random point pairs, the prediction also tends to be less accurate on local structures. If we only use edge-guided sampling that samples point pairs around edges (*i.e.*, Ours_E), depth boundaries become sharper at the cost of depth consistency both qualitatively and quantitatively. Therefore, global and local information are both important in our task, and the combination of the two (Ours_ER) strikes a good balance. To further improve the performance of depth consistency on object instances, we incorporate instance-guided sampling into our sampling strategy (*i.e.*, Ours_ERI). One can observe that Ours_ERI's performances of depth consistency and boundary accuracy are both improved over Ours_ER. The importance of instance-guided sampling is also reflected in the improvements of Ours_ERIM over Ours_ERM, as the only difference between the two is whether using instance guidance or not. Overall, our full model achieves the best performance.

Methods	Ibims		TUM		
	ε_{acc}	ε_{comp}	si-human	si-intra	si-inter
DIW [3]	8.083	82.549	0.437	0.345	0.474
DL [34]	2.391	41.456	0.319	0.268	0.339
RW [36]	3.029	67.725	0.304	0.238	0.330
MD [22]	3.439	75.719	0.349	0.266	0.379
YT3D [4]	7.542	85.921	0.347	0.288	0.369
MC [21]	3.588	64.495	0.294	0.227	0.319
MiDaS [19]	2.766	66.290	0.288	0.228	0.309
Ours_AI	2.829	73.197	0.287	0.230	0.308
Ours_R	2.345	50.507	0.296	0.227	0.320
Ours_E	2.029	61.380	0.322	0.240	0.350
Ours_ER	2.203	47.585	0.301	0.226	0.326
Ours_ERI	1.936	53.029	0.291	0.225	0.315
Ours_ERM	2.093	**34.962**	0.296	0.228	0.319
Ours_ERIM	**1.835**	41.294	**0.280**	**0.212**	**0.303**

Table 2. Quantitative evaluation, and an ablation of variants on our loss function, including: Ours_AI: the baseline model trained on our data with affine invariant and multi-scale gradient losses as in [19]; Ours_R: random sampling ranking loss; Ours_E: edge-guided sampling; Ours_ER: edge-guided sampling + Ours_R; Ours_ERI: instance-guided sampling + Ours_ER; Ours_ERM: multi-scale gradient matching term + Ours_ER; Ours_ERIM: our model trained with our final loss functions. For all metrics, lower is better.

6. Conclusion

Disparity data generated from stereo images and videos is a promising source of supervision for depth prediction methods, but it can only approximate the true inverse depth

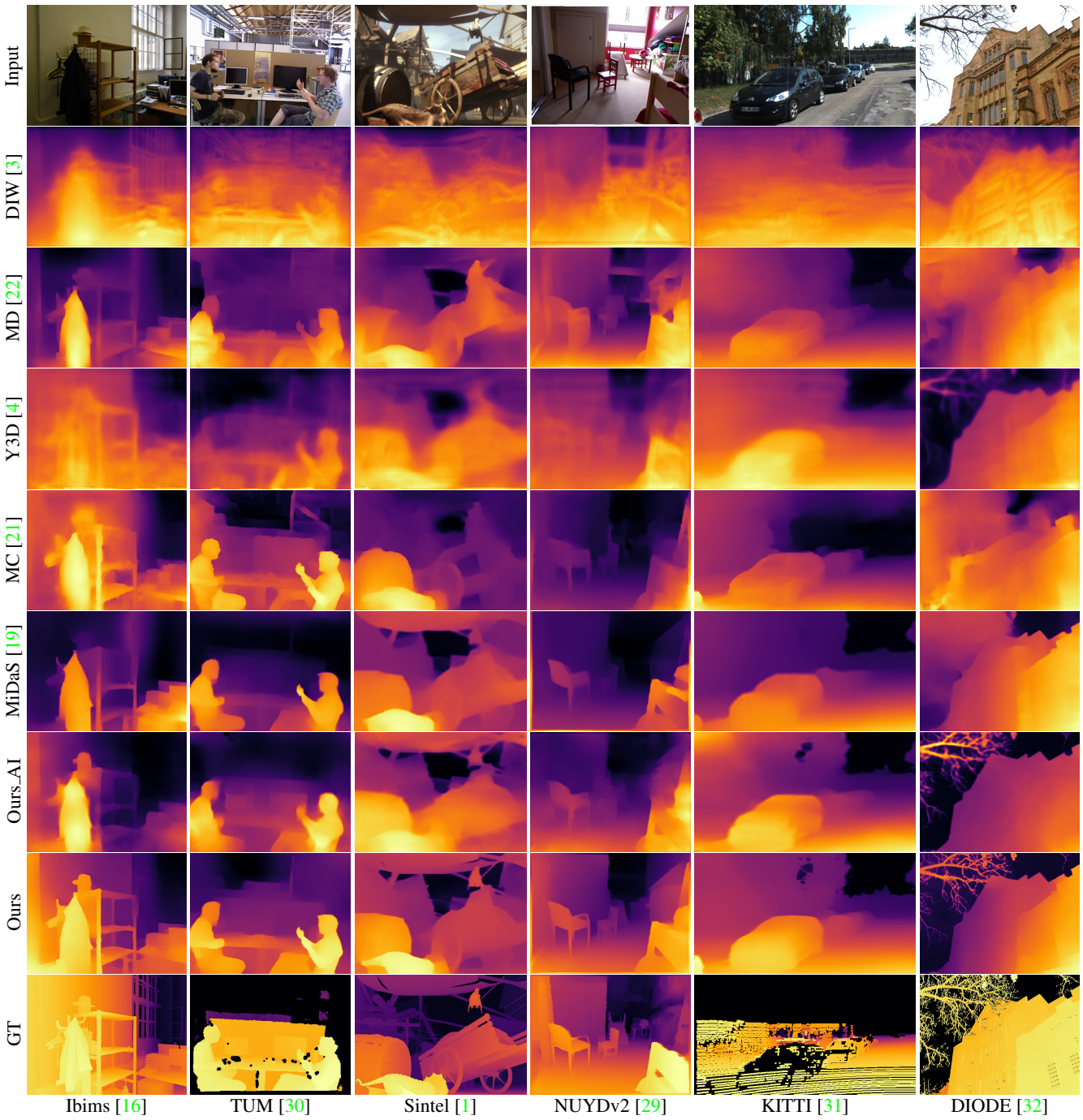


Figure 5. Qualitative results of single image depth prediction methods applied to different datasets.

up to an affine transformation. As a result, we have introduced a structure-guided ranking loss to guide the network towards the hardest, and most critical, components of depth reconstruction: depth discontinuities. In addition, we introduced a high-resolution web stereo image dataset that covers diverse scenes with dense ground truth, and showed that our proposed loss can learn consistent predictions with sharp depth discontinuities. One feature of our loss is that it is easily *guide-able*, meaning for any new task and dataset, the sampling strategy could be tweaked to address specific attributes necessary for downstream applications.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (Grant No. 61876211 and U1913602) and in part by the Adobe Gift. Part of the work was done when KX was an intern at Adobe Research.

References

[1] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 611–625, 2012. 3, 5, 8

[2] Ayan Chakrabarti, Jingyu Shao, and Gregory Shakhnarovich. Depth from a single image by harmonizing overcomplete local network predictions. In *Advances in Neural Information Processing Systems*, 2016. 2

[3] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems*, pages 730–738. 2016. 2, 3, 5, 6, 7, 8

[4] Weifeng Chen, Shengyi Qian, and Jia Deng. Learning single-image depth from videos using quality assessment networks. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 5604–5613, 2019. 1, 2, 3, 6, 7, 8

[5] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proc. Int. Conf. on Computer Vision (ICCV)*, 2015. 2

[6] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, 2014. 2, 5

[7] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep Ordinal Regression Network for Monocular Depth Estimation. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[8] A Gaidon, Q Wang, Y Cabon, and E Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[9] Ravi Garg and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Proc. European Conf. on Computer Vision (ECCV)*, 2016. 2

[10] Rahul Garg, Neal Wadhwa, Sameer Ansari, and Jonathan T Barron. Learning single camera depth estimation using dual-pixels. *arXiv preprint arXiv:1904.05822*, 2019. 1, 2

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[12] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 2961–2969, 2017. 4

[14] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 5

[15] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depthtransfer: Depth extraction from video using non-parametric sampling. *Trans. Pattern Analysis and Machine Intelligence*, 2014. 2

[16] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. In *Proc. European Conf. on Computer Vision Workshop (ECCV-WS)*, pages 331–348, 2018. 1, 2, 5, 7, 8

[17] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[18] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *Proc. IEEE Int. Conf. 3D Vision (3DV)*, 2016. 2

[19] Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv:1907.01341*, 2019. 1, 2, 3, 6, 7, 8

[20] Jae-Han Lee, Minhyeok Heo, Kyung-Rae Kim, and Chang-Su Kim. Single-image depth estimation based on fourier domain analysis. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[21] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 3, 5, 6, 7, 8

[22] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 4, 6, 7, 8

[23] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2010. 2

[24] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *Trans. Pattern Analysis and Machine Intelligence*, 2015. 2

[25] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[26] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[27] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Learning 3d scene structure from a single still image. *Trans. Pattern Analysis and Machine Intelligence*, 2009. 2

[28] Thomas Schöps, Johannes L. Schönberger, 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. Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[29] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. European Conf. on Computer Vision (ECCV)*, 2012. 1, 2, 5, 8

[30] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. Int. Conf. on Intelligent Robot Systems (IROS)*, 2012. 2, 5, 8

[31] Jonas Uhrig, Nick Schneider, Lucas Schneider, Uwe Franke, Thomas Brox, and Andereas Geiger. Sparsity invariant cnns. In *Proc. IEEE Int. Conf. on 3D Vision (3DV)*, 2017. 2, 5, 8

[32] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *arxiv: 1908.00463*, 2019. 2, 5, 8

[33] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *Proc. IEEE Int. Conf. on 3D Vision (3DV)*, 2019. 1, 2, 3, 5

[34] Lijun Wang, Xiaohui Shen, Jianming Zhang, Oliver Wang, Zhe Lin, Chih-Yao Hsieh, Sarah Kong, and Huchuan Lu. Deeplens: Shallow depth of field from a single image. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 37(6):6:1–6:11, 2018. 2, 5, 6, 7

[35] Jamie Watson, Michael Firman, Gabriel J. Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proc. Int. Conf. on Computer Vision (ICCV)*, 2019. 2

[36] 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. Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 3, 4, 5, 6, 7

[37] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[38] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[39] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proc. Int. Conf. on Computer Vision (ICCV)*, 2019. 2

[40] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[41] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T Freeman. Learning ordinal relationships for mid-level vision. In *Proc. Int. Conf. on Computer Vision (ICCV)*, 2015. 5