Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer

Katrin Lasinger* ETH Zurich
René Ranftl* Intel Labs
Konrad Schindler ETH Zurich
Vladlen Koltun Intel Labs

Figure 1. We show how to leverage multiple, complementary training sets for single-view depth estimation, in spite of varying and unknown depth range and scale. Our approach enables strong generalization across datasets. From left to right: input image, estimated inverse depth, textured 3D model, untextured 3D model. Input images from the Microsoft COCO dataset, which was not seen during training.

Abstract

The success of monocular depth estimation relies on large and diverse training sets. Due to the challenges associated with acquiring dense ground-truth depth across different environments at scale, a number of datasets with distinct characteristics and biases have emerged. We develop tools that enable mixing multiple datasets during training, even if their annotations are incompatible. In particular, we propose a training objective that is invariant to changes in depth range and scale. Armed with this objective, we explore an abundant source of training data: 3D films. We demonstrate that despite pervasive inaccuracies, 3D films constitute a useful source of data that is complementary to existing training sets. We evaluate the presented approach on diverse datasets, focusing on zero-shot cross-dataset transfer*: testing the generality of the learned model by evaluating it on datasets that were not seen during training. The experiments confirm that mixing data from complementary sources yields improved depth estimates, particularly on previously unseen datasets.*

即使标注不一致 也可以混合多个数据集 并提出尺度平移不变性 使用大量3D电影数据

混合数据集训练可以改善深度估计，提高模型泛化性

1. Introduction

Depth is among the most useful intermediate representations for action in physical environments [48]. Yet despite its utility, monocular depth estimation remains a challenging open problem. When only one image is given as input, depth is heavily underconstrained and its estimation calls for the use of multiple monocular cues along with comprehensive prior knowledge. This must rely on learning-based techniques [20, 37].

To learn models that are effective across a variety of scenarios, we need training data that provides relevant supervision and captures the diversity of the visual world. The key challenge is acquiring such data at sufficient scale. Sensors that provide dense ground-truth depth in dynamic scenes, such as structured light or time-of-flight, have limited range

*Equal contribution.

and operating conditions [22, 19, 10]. Laser scanners are expensive and most designs can only provide sparse depth measurements when the scene is in motion. Stereo cameras are a promising source of data [13, 15], but collecting stereo images in diverse environments at scale remains a challenge. Structure-from-motion (SfM) reconstruction has been used to construct training data for monocular depth estimation across a variety of scenes [28], but the result does not include independently moving objects and is incomplete due to the limitations of multi-view matching. On the whole, none of the existing datasets are sufficiently rich and unbiased to support the training of a general model that works robustly on real data from wildly diverse scenes. At present, we are faced with multiple datasets that may usefully complement each other, but are individually biased and incomplete.

In this paper, we propose an approach to mixing diverse datasets for training monocular depth estimation models. We develop a novel loss function that is invariant to the major sources of incompatibility between datasets, including unknown and inconsistent scale and baselines. Our loss enables training on data that was acquired with diverse sensing modalities such as stereo cameras (with potentially unknown calibration), laser scanners, and structured light sensors. We explore strategies for mixing datasets during training and show that a principled approach based on multi-objective optimization [39] can lead to improved generalization performance.

Equipped with the ability to combine diverse datasets, we tap into a new source of data for monocular depth estimation: 3D film. We construct a new 3D Movies dataset and show that it provides a powerful training resource that improves generalization to new and dynamic environments.

Our experiments show that a model trained on a rich and diverse set of images from different sources, with the appropriate training procedure, delivers state-of-the-art results across a variety of environments. Our primary experimental procedure is *zero-shot cross-dataset transfer*. That is, we train a model on certain datasets and then test its performance on other datasets that were not seen during training. The basic intuition is that zero-shot cross-dataset performance is a more faithful proxy for performance in the "real world" than training and testing on subsets from a single biased dataset.

Our evaluation on eight different datasets suggests that our model outperforms prior art both quantitatively and qualitatively. Example results are shown in Figure 1.

2. Related Work

Early work on monocular depth estimation used MRF-based formulations [37], simple geometric assumptions [20], and non-parametric methods [21]. More recently, significant advances have been made by leveraging the expressive power of convolutional networks. Eigen et al. [9] trained a multi-scale deep network to perform depth regression. Various architectural innovations have been proposed to enhance prediction accuracy [25, 36, 29, 12, 26]. These methods need ground-truth depth for training, which is commonly acquired using RGB-D cameras or LiDAR sensors. The performance of these methods in unconstrained scenes is limited by the lack of diverse ground-truth data at scale.

Garg et al. [13] proposed to use calibrated stereo cameras for self-supervision. While this significantly simplifies the acquisition of training data, diverse, large-scale stereo datasets are still not available.

Various approaches that leverage self-supervision have been proposed, but they either require stereo images [15, 47] or are based on apparent motion [49, 31, 2], and are thus challenging to apply to highly dynamic scenes. Other approaches leverage existing stereo matching networks to obtain supervision [17, 30].

We argue that the deployment of high-capacity deep models for monocular depth estimation in unconstrained environments is limited by the lack of large-scale, dense ground truth that spans a variety of scenes. Indeed, commonly used datasets feature homogeneous scene compositions such as street scenes in a limited geographic area [14, 32, 37] or indoor environments [40], and have only a limited number of dynamic objects. Models that are trained on data with such strong biases are prone to fail in unconstrained environments.

Efforts have been made to create datasets that overcome these limitations. Chen et al. [3] used crowdsourcing to sparsely annotate ordinal relations in images that were collected from the web. Xian et al. [45] collected a stereo dataset from the web and used off-the-shelf tools to generate dense ground-truth disparity; while this dataset is fairly diverse and provides dense ground truth, it only contains 3,600 images. Li and Snavely [28] used SfM and MVS to reconstruct many (predominantly static) scenes to obtain depth supervision. Each of these efforts contributes datasets with different characteristics and limitations. On the whole, there is no single large-scale dataset with dense ground truth for diverse dynamic scenes.

To the best of our knowledge, the controlled mixing of multiple data sources has not been explored before in this context. Ummenhofer et al. [43] presented a model for two-view structure and motion estimation and trained it on the union of several datasets that depict static scenes. However, this work did not propose strategies for optimal mixing or study the influence of combining multiple datasets. The model was further constrained to work with the intrinsic parameters of a single camera model.

Concurrent work. Several concurrent projects aim to extend the generalization capabilities of monocular depth es-

Dataset	Indoor	Outdoor	Dynamic	Video	Metric	Dense	Accuracy	Diversity	Annotation	# Images
NYUDv2 [40]	✓		(✓)	✓	✓	✓	Medium	Low	RGB-D	407K
SUN-RGBD [41]	✓			✓	✓	✓	Medium	Low	RGB-D	10K
ScanNet [7]	✓			✓	✓	✓	Medium	Low	RGB-D	2.5M
Make3D [37]		✓			✓	✓	Low	Low	Laser	534
KITTI LiDAR [14, 32]		✓	✓	✓	✓		Medium	Low	Laser	93K
KITTI Stereo [14, 32]		✓	✓	✓	✓	✓	Medium	Low	Stereo	93K
Cityscapes [6]		✓	✓	✓	✓	✓	Medium	Low	Stereo	25K
DIW [3]	✓	✓	✓				Low	High	User clicks	496K
ETH3D [38]	✓	✓			✓	✓	High	Low	Laser	454
Tanks and Temples [24]	✓	✓			✓	✓	High	Low	Laser	3290
Sintel [1]	✓	✓	✓	✓	(✓)	✓	High	Medium	Synthetic	1064
MegaDepth [28]		✓	(✓)			✓	Medium	Medium	SfM	130K
ReDWeb [45]	✓	✓	✓			✓	Medium	High	Stereo	3600
3D Movies (Ours)	✓	✓	✓	✓		✓	Medium	High	Stereo	∼ 50K/movie

Table 1. Datasets that provide relevant supervision for monocular depth estimation. No single real-world dataset features a large number of diverse scenes with dense and accurate ground truth.

timation by collecting larger and more diverse datasets. Li et al. [27] use SfM and MVS to construct a dataset from a collection of videos of people imitating mannequins (i.e. the people are frozen in action while the camera moves through the scene). Chen et al. [4] propose an approach to automatically assess the quality of sparse SfM reconstructions to enable the construction of a large dataset. Wang et al. [44] build a large and diverse dataset from stereo videos sourced from the Web, while Cho et al. [5] collect a large dataset of outdoor scenes using handheld stereo cameras. Gordon et al. [16] estimate intrinsic parameters of YouTube videos in order to leverage them for training. Our approach can be used to directly integrate these datasets in a single training procedure to train even more general and accurate models.

3. 3D Movies Dataset

Table 1 summarizes datasets that provide relevant supervision for monocular depth estimation. Most datasets feature a limited range of environments, such as indoor [40, 41, 7] or road scenes [14, 32]. Some are restricted to static scenes [37, 6, 24, 38, 28]. Only DIW [3] and ReDWeb [45] include diverse dynamic scenes in a variety of settings. However, the diversity of the data comes with drawbacks. DIW is large and diverse, but each image provides only an ordinal depth relation for a single pair of points. And the ReDWeb dataset contains only 3,600 images.

We propose to extract depth data from 3D movies. This source of data is complementary to existing datasets and comes with its own strengths. 3D movies feature diverse dynamic environments that range from human-centric imagery in story- and dialogue-driven Hollywood films to nature scenes with landscapes and animals in documentary features. While the data does not provide metric depth, we can use stereo matching to obtain relative depth. Using relative depth for supervision comes with challenges and we will present techniques for overcoming them.

Our driving inspiration is the scale and diversity of the data. 3D movies provide the largest known source of stereo pairs, presenting the possibility of tapping into millions of images from an ever-growing library of content. We note that 3D movies have been used in related tasks in isolation [18, 46]. We will show that this data reveals its full potential in combination with other, complementary data sources. 我们证明了3D电影和其他数据互补后有优势

Challenges. Movie data comes with its own challenges and imperfections. The primary objective when producing stereoscopic film is providing a visually pleasing viewing experience while avoiding discomfort for the viewer [8]. This means that the disparity range for any given scene (also known as the depth budget) is limited and depends on both artistic and psychophysical considerations. For example, disparity ranges are often increased in the beginning and the end of a movie, in order to induce a very noticeable stereoscopic effect for a short time. Depth budgets in the middle may be lower to allow for more comfortable viewing. Stereographers thus adjust their depth budget depending on the content, transitions, and even the rhythm of scenes.

In consequence, focal lengths, baseline, and convergence angle between the cameras of the stereo rig are unknown and vary between scenes even within a single film. Furthermore, in contrast to image pairs obtained directly from a standard stereo camera, stereo pairs in movies usually contain both positive and negative disparities to allow objects to be perceived either in front of or behind the screen. Additionally, the depth that corresponds to the screen is scene-dependent and is often modified in post-production by shifting the image pairs. We describe data extraction and training procedures that address these challenges.

Movie selection and preprocessing. We selected a diverse set of 23 movies. The selection was based on the follow-

DIW大且多样性大 但是只有点对的序关系

处于艺术和心理考虑，任何场景都有限制。深度范围通常有限，一般开头和结尾为了加强视差会有很强的立体效果，中间较低

因此，焦距、基线和收敛角即使在一部电影中也不知道

此外，不像标准立体相机，电影有正视差和负视差。此外，深度通常都是缩放平移了，因此需要考虑这些挑战

大部分深度数据集有场景和深度范围限制，因此用3D电影作为补充，他有很大的优势，从故事到人物、风景和动物都很丰富。尽管不提供深度，但可以用立体匹配获得相对深度

Figure 2. Sample images from the 3D Movies dataset. We show images from some of the films in the training set together with their inverse depth maps. Sky regions and invalid pixels are masked out. Each image is taken from a different film. 3D movies provide a massive source of diverse data. **逆深度=视差，天空被卡掉**

物理相机拍摄非后处理

平衡真实和多样性

选择蓝光电影

ing considerations. 1) We only selected movies that were shot using a physical stereo camera. (Some 3D films are shot with a monocular camera and the stereoscopic effect is added in post-production by artists.) 2) We tried to balance realism and diversity. 3) We only selected movies that are available in Blu-ray format and thus allow extraction of high-resolution images. The complete list of movies can be found in supplementary material. **完整名单见补充材料**

We extract stereo image pairs at 1920x1080 resolution and 24 frames per second (fps). Movies have varying aspect ratios, resulting in black bars on the top and bottom of the frame, and some movies have thin black bars along frame boundaries due to realignment of stereo images in post-production. We thus center-crop all frames to 1880x800 pixels. We use the chapter information (Blu-ray meta-data) to split each movie into individual chapters. We drop the first and last chapters since they usually include the introduction and credits.

提取1920x1080分辨率，由于会有黑边，所以中心抠出1880x800
由于电影前后视差较强，因此会扔掉前后的章节

We use the scene detection tool of FFmpeg [11] with a threshold of 0.1 to extract individual clips. We discard clips that are shorter than one second to filter out chaotic action scenes and highly correlated clips that rapidly switch between protagonists during dialogues. To balance scene diversity, we sample the first 24 frames of each clip and additionally sample 24 frames every four seconds for longer clips. Example frames from the resulting dataset are shown in Figure 2.

丢掉少于1s、动作混乱和高度相关的、快速切换的场景

Disparity extraction. The extracted image pairs can be used to estimate disparity maps using stereo matching. Unfortunately, state-of-the-art stereo matchers perform poorly when applied to movie data, since the matchers were designed and trained to match only over positive disparity ranges. This assumption is appropriate for the rectified output of a standard stereo camera, but not to image pairs extracted from stereoscopic film. Moreover, disparity ranges encountered in 3D movies are usually smaller than ranges

最好的立体匹配不适用电影因为只用正视差训练

该假设值适用于标准立体相机，不适用3D电影，此外，3D电影视差范围通常小于标准立体设置中的常见范围（平均视差范围是32像素）

that are common in standard stereo setups due to the limited depth budget. (The average disparity range in our dataset is 32 pixels.)

To alleviate these problems, we apply a modern optical flow algorithm [42] to the stereo pairs. We retain the horizontal component of the flow as a proxy for disparity. Optical flow algorithms naturally handle both positive and negative disparities and usually perform well for displacements of moderate size. For each stereo pair we use the left camera as the reference and extract the optical flow from the left to the right image and vice versa. We perform a left-right consistency check and mark pixels with a disparity difference of more than one pixel as invalid. In a final step, we detect pixels that belong to sky regions using a pre-trained semantic segmentation model [35] and set their disparity to the minimum disparity in the image.

将天空设成最小视差

Dataset statistics. We use frames from 19 movies for training and set aside two movies for validation and two movies for testing, respectively. An overview of the statistics of the resulting training set is shown in Table 2. The complete dataset contains 38,000 clips of one second each and close to one million frames. Since multiple frames are part of the same clip, the complete dataset is highly correlated. We thus subsample the dataset at 1 fps or 4 fps, respectively.

4. Training on Diverse Data

Training models for monocular depth estimation on diverse datasets presents a challenge because ground-truth data may take different forms. Ground truth may be present in the form of absolute depth (from laser-based measurements or stereo cameras with known calibration), depth up to an unknown scale (from SfM), or disparity maps (from stereo cameras with unknown calibration). The main requirement for a sensible training scheme is to carry out computations in an appropriate output space that is compatible with all ground-truth representations and is numerically well-behaved. We further need to design a loss function that is flexible enough to handle diverse sources of data while making optimal use of all available information.

需要设计loss处理不同源的数据

We identify three major challenges. 1) Inherently different representations of depth: direct depth versus inverse depth representations (such as disparity). 2) Scale ambiguity: for some data sources, depth is only given up to an unknown scale. 3) Shift ambiguity: the ground truth in the 3D Movies dataset is only given up to an unknown global shift that is a function of the unknown baseline and a possible shift of the disparity range in post-production.

不同的深度表示，放缩未知，平移未知

Movies	19
Resolution	1880x800
Number of clips	38,000
Images @4fps	152,000
Total images	912,000
Avg./Max. disparity range	32/223
Min./Max. disparity	-144/156

Table 2. Statistics of the 3D Movies training set.

Scale- and shift-invariant loss. We propose to perform prediction in inverse depth space together with a scale- and shift-invariant dense loss to handle the aforementioned ambiguities. Let M denote the number of pixels in an image with valid ground truth and let $\boldsymbol{\theta}$ be the parameters of the prediction model. Let $\mathbf{d} = \mathbf{d}(\boldsymbol{\theta}) \in \mathbb{R}^M$ be an inverse depth prediction and let $\mathbf{d}^* \in \mathbb{R}^M$ be the corresponding ground-truth inverse depth. We index individual pixels by subscripts. We define a scale- and shift-invariant loss for a single sample as

表明每张图计算一个scale和shift

$$\mathcal{L}_{ssi}(\mathbf{d}, \mathbf{d}^*) = \min_{s,t} \frac{1}{2M} \sum_{i=1}^{M} \left(s\mathbf{d}_i + t - \mathbf{d}_i^*\right)^2, \quad (1)$$

但是正常分布会被改变分布，难以学习

where $s \in \mathbb{R}^+$ accounts for the unknown scale and $t \in \mathbb{R}$ accounts for an unknown shift between the inverse depth maps. The loss effectively aligns the scale and shift of the estimate $\mathbf{d}$ to the ground truth $\mathbf{d}^*$ based on a least-squares criterion before measuring the mean squared error. The factors s and t can be efficiently determined in closed form, as follows. Let $\vec{\mathbf{d}}_i = (\mathbf{d}_i, 1)^\top$ and $\mathbf{h} = (s, t)^\top$. We can rewrite (1) as

$$\mathcal{L}_{ssi}(\mathbf{d}, \mathbf{d}^*) = \min_{\mathbf{h}} \frac{1}{2M} \sum_{i=1}^{M} \left(\vec{\mathbf{d}}_i^\top \mathbf{h} - \mathbf{d}_i^*\right)^2, \quad (2)$$

which has the closed-form solution

$$\mathbf{h}^{opt} = \left(\sum_{i=1}^{M} \vec{\mathbf{d}}_i \vec{\mathbf{d}}_i^\top\right)^{-1} \left(\sum_{i=1}^{M} \vec{\mathbf{d}}_i \mathbf{d}_i^*\right). \quad (3)$$

Substituting into (2) yields 最小二乘法求解

$$\mathcal{L}_{ssi}(\mathbf{d}, \mathbf{d}^*) = \frac{1}{2M} \sum_{i=1}^{M} \left(\vec{\mathbf{d}}_i^\top \mathbf{h}^{opt} - \mathbf{d}_i^*\right)^2. \quad (4)$$

It is straightforward to show that this loss is indeed invariant to scale and shift of the prediction. We provide a proof sketch in supplementary material.

Relation to existing loss functions. The importance of accounting for unknown or varying scale in the training of monocular depth estimation models has been recognized early. Eigen et al. [9] proposed a scale-invariant loss in log-depth space. Their loss can then be written as

$$\mathcal{L}_{silog}(\mathbf{z}, \mathbf{z}^*) = \min_{s} \frac{1}{2M} \sum_{i=1}^{M} \left(\log(e^s \mathbf{z}_i) - \log(\mathbf{z}_i^*)\right)^2, \quad (5)$$

只提出了scale不变性损失而且是逆深度

where $\mathbf{z}_i = \mathbf{d}_i^{-1}$ and $\mathbf{z}_i^* = (\mathbf{d}_i^*)^{-1}$ are depths up to unknown scale. By comparing this loss to (1) it becomes clear

我们的loss和Eigen的都考虑了scale，但是我们还考虑了未知的shift在逆深度范围
且我们是在逆深度空间定义，稳定且兼容相对深度的表示，允许对误差分布建模为
高斯分布

that both losses account for the unknown scale of the predictions, but only (1) accounts for an unknown global shift of the inverse depth range. Moreover, the two losses are evaluated on different representations of depth. Our loss (1) is defined in inverse depth space, which is numerically stable, compatible with common representations of relative depth, and allows to model the error distribution as Gaussian, which is represented well by a quadratic loss [33].

Chen et al. [3] proposed a generally applicable loss for relative depth estimation based on ordinal relations:

$$\phi(\mathbf{d}_i, \mathbf{d}_j) = \begin{cases} \log(1 + \exp(-\mathbf{d}_i + \mathbf{d}_j) l_{ij}), & l_{ij} \neq 0 \\ (\mathbf{d}_i - \mathbf{d}_j)^2, & l_{ij} = 0, \end{cases} \quad (6)$$

where $l_{ij} \in \{-1, 0, 1\}$ encodes the ground-truth ordinal relation. This loss encourages pushing points infinitely far apart when $l_{ij} \in \{+1, -1\}$ and pulling them to the same depth when $l_{ij} = 0$. Note that $\phi(\mathbf{d}_i, \mathbf{d}_j)$ measures the ordinal relation between a pair of points, which is inefficient if applied exhaustively to dense ground truth. Xian et al. [45] suggest to sparsely evaluate this loss by randomly sampling point pairs, even when dense ground truth is available. In contrast, our proposed loss takes all available data into account. While the ordinal loss can be applied to arbitrary depth representations and is thus suited for mixing diverse datasets, we will show that our scale- and shift-invariant loss leads to consistently better performance.

证明我们的损失与相对深度损失有一样好的表现

Regularization terms. We adapt the multi-scale scale-invariant gradient matching term [28] to the inverse depth space. This term biases discontinuities to be sharp and to coincide with discontinuities in the ground truth. Let

逆深度的多尺度梯度

$$R_i = \vec{\mathbf{d}}_i^{\top} \mathbf{h}^{opt} - \mathbf{d}_i^*. \quad (7)$$

We define the gradient matching term as

$$\mathcal{L}_{reg}(\mathbf{d}, \mathbf{d}^*) = \frac{1}{M} \sum_{k=1}^{K} \sum_{i=1}^{M} \left(|\nabla_x R_i^k| + |\nabla_y R_i^k| \right), \quad (8)$$

where R^k denotes the difference of inverse depth maps at scale k. As proposed by Li and Snavely [28], we use $K = 4$ scale levels, halving the image resolution at each level. Note that in inverse depth space the estimated scale needs to be applied to the prediction before measuring gradients. This is in contrast to the term proposed in [28] where predictions are given in log-depth space and no pre-scaling is required.

Mixing strategies. Our final loss for a training set l is

$$\mathcal{L}_l = \frac{1}{N_l} \sum_{n=1}^{N_l} \mathcal{L}_{ssi}\left(\mathbf{d}^n, (\mathbf{d}^*)^n\right) + \alpha \mathcal{L}_{reg}\left(\mathbf{d}^n, (\mathbf{d}^*)^n\right), \quad (9)$$

where N_l is the number of samples in the training set and α is set to 0.5.

While our choice of prediction space together with our loss enables mixing datasets, it is not immediately clear in what proportions different datasets should be integrated during training with a stochastic optimization algorithm. We explore two different strategies in our experiments.

The first, naive strategy is to mix datasets in equal parts in each minibatch. For a minibatch of size B, we sample B/L training samples from each dataset, where L denotes the number of distinct datasets. This strategy ensures that all datasets are represented equally in the effective training set, irrespective of their size. 采样策略

Our second strategy explores a more principled approach, where we adapt a recent procedure for Pareto-optimal multi-task learning to our setting [39]. We define learning on each dataset as a separate task and are thus seeking an approximate Pareto-optimum over datasets (i.e. the loss cannot be decreased on any of the training sets without increasing the loss on at least one of the other training sets). Formally, we use the algorithm presented in [39] to minimize the multi-objective optimization criterion

$$\min_{\boldsymbol{\theta}} \left(\mathcal{L}_1(\boldsymbol{\theta}), \ldots, \mathcal{L}_L(\boldsymbol{\theta}) \right)^{\top}, \quad (10)$$

where we share the parameters $\boldsymbol{\theta}$ of the deep network across all datasets.

5. Experiments

Experimental setup. We start from the experimental setup proposed by Xian et al. [45] and use their ResNet-based multi-scale architecture for single-image depth prediction. We initialize all ResNet-50 blocks with pretrained ImageNet weights and initialize other layers randomly. We use Adam [23] with a learning rate of 10^{-4} for randomly initialized layers and 10^{-5} for layers that were initialized with pretrained weights. We set the exponential decay rate parameters of the moving averages for Adam to $\beta_1 = 0.9$ and $\beta_2 = 0.999$ in all experiments. The batch size is set to 8. Images are flipped horizontally with a 50% chance, and randomly cropped to augment the data and maintain the aspect ratio across different input images. We pretrain the network for 300 epochs on 3,240 images from the ReDWeb dataset to produce a baseline model comparable to [45].

For all experiments that follow, we start from this pretrained model and fine-tune on the respective collections of different datasets. When fine-tuning we fix the learning rate to 10^{-5} for all layers. We use the same dataset augmentation and a batch size of $8L$, i.e. when mixing three datasets the batch size is 24. When comparing datasets of different sizes, the term epoch is not well-defined anymore; we thus denote an epoch as processing 72,000 images, roughly the size of the MegaDepth training dataset, and train for 60 epochs. For all datasets, we shift and scale the ground-truth inverse depth to the range $[0, 10]$.

Training datasets. We use three complementary datasets for training. ReDWeb [45] (RW) is small but features diverse, dynamic scenes with ground truth that was acquired with a relatively large stereo baseline. The 3D Movies dataset (MV) features highly dynamic scenes and is considerably larger but is biased towards dominant foreground objects due to the depth budget. MegaDepth [28] (MD) is large and shows predominantly static scenes; the ground truth in this dataset is usually more accurate in background regions due to the large-baseline multi-view stereo reconstruction. We hold out test and validation sets for all datasets (details in the supplement).

Test datasets. To benchmark the generalization performance of different models, we use a variety of datasets for testing. We split datasets in two distinct groups. The first group are datasets where the model was trained on similar data, namely our held-out validation sets of ReDWeb, MegaDepth, and Movies. The second group are entirely different datasets that were *never seen during training*. We chose five datasets based on diversity and accuracy of their ground truth. DIW [3] is highly diverse but provides ground truth only in the form of sparse ordinal relations. ETH3D [38] features highly accurate laser-scanned ground truth on static scenes. Sintel [1] features perfect ground truth for synthetic scenes. KITTI [32] and NYU [40] are commonly used datasets with characteristic biases. Note that we never fine-tune models on any of the datasets in this second group. We refer to this experimental procedure as *zero-shot cross-dataset transfer*.

Metrics. For each dataset, we use a single metric that fits the ground-truth data in that dataset. For DIW we use the Weighed Human Disagreement Rate [3]. For datasets that are based on relative depth, we measure the root mean squared error in disparity space (Movies, ReDWeb, MegaDepth). For datasets that provide accurate absolute depth we measure the mean absolute value of the relative error $(1/M)\sum_{i=1}^{M}|z_i - z_i^*|/z_i^*$ in depth space (ETH3D, Sintel). Finally, we use the percentage of pixels with $\max(\frac{z_i}{z_i^*}, \frac{z_i^*}{z_i}) > 1.25$ to evaluate models on the KITTI and NYU datasets (*c.f.* [9]). We align predictions and dense ground truth in scale and shift before measuring errors (details in the supplement). To summarize the performance of different models across test datasets, we rank methods by their performance on each dataset and compute the average rank. This is our primary performance measure for zero-shot cross-dataset transfer.

Comparison of loss functions. We show the effect of different loss functions on validation performance in Table 3. We used the ReDWeb dataset to train networks with different losses: mean squared error in disparity space (MSE); the scale-invariant loss in log-depth space (5) as defined by Eigen et al. [9]; the ordinal loss (6), where we sample 5,000 point pairs randomly (ORD); a scale-invariant loss in disparity space, where we assume a fixed offset of $t = 0$ in (1) and only estimate the scale s (SIMSE); and the full scale- and shift-invariant loss (1) (SSIMSE). Note that the model trained with the ordinal loss (ORD) corresponds to our reimplementation of Xian et al. [45]. Table 3 shows that our proposed loss yields the lowest validation error on all datasets. We thus conduct all experiments that follow using the scale- and shift-invariant loss.

Training on diverse datasets. We show the performance of models that were trained on different combinations of training sets in Table 4. We trained models with the MV dataset sampled at 1 fps as well as 4 fps. We indicate if models were trained using Pareto-optimal mixing (MGDA) in a separate column. We observe that adding more training sets consistently improves performance across validation sets. The strongest models (RW+MD+MV) were trained on all three datasets and outperform the best single-dataset model (RW) on all validation sets. Oversampling clips in the movie dataset (4 fps) typically leads to an increase in performance, likely because it provides a natural form of data augmentation due to small shifts in perspective and scene composition over consecutive frames. We can additionally see that performing principled Pareto-optimal dataset mixing (MGDA) leads to a noticeable improvement over the naive mixing strategy on most datasets. Using all three datasets (RW+MD+MV) with MGDA yields our best-performing model.

Effect of 3D Movies dataset. Our 3D Movies dataset features smaller baselines than ReDWeb and thus provides less accurate disparities in the background. However, an analysis of Table 4 shows that it consistently improves performance when used in combination with other datasets, especially for diverse test sets. Fine-tuning only on MegaDepth or 3D Movies decreases performance compared to training on ReDWeb only, while fine-tuning on MegaDepth and 3D Movies (MD+MV) leads to consistently better results. We further observe that mixing 3D Movies and ReDWeb always leads to an improvement on the ReDWeb validation set. Similar findings hold on DIW, arguably the most diverse test set. We achieve the best results when mixing all three datasets. Without Pareto-optimal mixing, RW+MD+MV (4 fps) improves performance on five out of seven datasets when compared to RW+MD, while RW+MD+MV (1 fps) improves performance on four out of seven datasets.

Comparison to the state of the art. We compare our best-performing model to various state-of-the-art baselines in Table 5. The top part of the table compares to baselines that were not fine-tuned on any of the evaluated datasets (i.e. zero-shot transfer, akin to our model). The bottom part shows baselines that were fine-tuned on a subset of the datasets for reference. In the training set column, CS refers

Loss	RW	MV	MD
MSE	1.114	0.803	0.379
Eigen	1.565	0.948	0.825
ORD	1.100	0.836	0.354
SIMSE	1.046	0.777	0.349
SSIMSE	**1.045**	**0.768**	**0.347**

Table 3. Comparison of different loss functions when pretraining on the ReDWeb dataset.

Training sets	Fps	MGDA	RW	MV	MD	DIW	ETH3D	Sintel	KITTI	NYU	Rank
RW → RW+MD+MV	4	✓	1.09	0.62	0.18	13.70	**0.160**	**0.319**	25.44	**19.05**	**3.1**
RW → RW+MD+MV	1	✓	**1.07**	0.66	0.21	**12.75**	0.167	0.343	27.27	19.89	4.7
RW → RW+MD+MV	4	✗	**1.07**	0.63	0.19	13.17	0.167	0.387	27.12	19.54	4.9
RW → RW+MD+MV	1	✗	1.09	0.64	0.18	13.60	0.166	0.368	27.75	19.28	5.1
RW → RW+MD	-	✗	1.11	0.79	0.16	13.44	0.168	0.341	**25.06**	20.23	5.1
RW → MD+MV	1	✗	1.17	0.62	0.16	14.10	0.162	0.371	28.89	20.35	5.9
RW → MD+MV	4	✗	1.19	0.62	0.17	14.31	0.167	0.413	28.37	20.40	7.2
RW → RW+MV	4	✗	1.10	0.61	0.35	14.44	0.175	0.354	31.43	22.66	7.3
RW → RW+MV	1	✗	1.09	0.62	0.35	14.59	0.170	0.331	32.59	23.25	7.4
RW → RW	-	✗	1.11	0.79	0.34	15.33	0.173	0.358	31.83	23.46	9.1
RW → MV	4	✗	1.27	**0.58**	0.37	16.05	0.191	0.364	46.91	30.98	10.2
RW → MD	-	✗	1.40	1.24	**0.13**	17.93	0.182	0.391	31.83	24.11	10.3
RW → MV	1	✗	1.27	0.60	0.37	16.13	0.185	0.392	41.87	29.98	10.8

Table 4. Performance of models when trained on different combinations of training sets: ReDWeb (RW), 3D Movies (MV), and MegaDepth (MD). MGDA refers to Pareto-optimal mixing. We evaluate zero-shot cross-dataset transfer: the five datasets to the right were never seen during training.

to Cityscapes [6], K to KITTI, and A → B indicates that a model was pretrained on A and fine-tuned on B.

Our model outperforms the baselines by a comfortable margin on most datasets. The only exception is KITTI, where a model that was trained on the visually similar CityScapes dataset achieves the lowest error [15]. It can also be seen that fine-tuning on the KITTI dataset improves accuracy on this specific dataset but often leads to worse performance on other datasets. A qualitative comparison is shown in Figure 3. The visual results correspond well to our quantitative evaluation. Only our model and the model of Xian et al. [45] adequately handle diverse scenes. The model of Xian et al. tends to miss details or place parts of the scene at the wrong depth. Interestingly, the bias of the model that was trained on street scenes [15] can be clearly observed in the sample from ETH3D, where the slope of the staircase is too flat and the thin horizontal shadow on the left side of the building is mistaken for a free-standing pole. Note, however, that this model better reconstructs the thin structures in the foreground, likely because such structures are frequently encountered in street scenes (e.g. street signs and poles). Additional results are shown in the supplement.

	Training sets	DIW	ETH3D	Sintel	KITTI	NYU	Rank
Ours	RW+MD+MV	**13.74**	**0.160**	**0.319**	25.44	19.05	2.2
Li [28]	MD	23.39	0.174	0.395	40.66	28.33	4.4
Xian [45]	RW	15.51	0.210	0.424	33.74	26.47	4.8
Godard [15]	CS	31.10	0.204	0.428	19.77	35.84	5.7
Casser [2]	CS	39.60	0.273	0.427	30.05	49.70	9.2
Godard [15]	CS → K	28.87	0.224	0.430	13.25	35.52	5.4
Fu [12]	K	33.67	0.235	0.419	**12.66**	40.49	6.1
Fu [12]	NYU	31.32	0.243	0.419	60.95	**10.26**	6.5
Chen [3]	NYU → DIW	14.81	0.236	0.444	41.43	28.08	7.0
Godard [15]	K	32.55	0.227	0.436	14.89	37.04	7.0
Casser [2]	K	41.41	0.268	0.417	19.77	46.78	7.7

Table 5. Comparison to state-of-the-art baselines. Top: Models that were not fine-tuned on any of the datasets. Bottom: Models that were fine-tuned on a subset of the tested datasets.

6. Conclusion

The success of deep networks has been driven by massive datasets. We believe that learning truly general models for monocular depth estimation will require not only innovations in network architectures, but also creative ways to boost the amount and diversity of training data. Motivated by the difficulty of capturing diverse depth datasets at scale, we have introduced tools for combining complementary sources of data. We have proposed a flexible loss function and a principled dataset mixing strategy. We have further introduced a dataset based on 3D movies that provides dense ground truth for diverse dynamic scenes. To evaluate the robustness and generality of trained models, we used *zero-shot cross-dataset transfer*: systematically testing models on datasets that were never seen during training. The results indicate that the presented ideas substantially advance monocular depth estimation in diverse environments. Our code and and pretrained models will be made available online.

References

[1] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012.

[2] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova. Unsupervised learning of depth and ego-motion: A structured approach. In *AAAI*, 2019.

[3] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. In *NIPS*, 2016.

[4] W. Chen, S. Qian, and J. Deng. Learning single-image depth from videos using quality assessment networks. In *CVPR*, 2019.

[5] J. Cho, D. Min, Y. Kim, and K. Sohn. A large RGB-D dataset for semi-supervised monocular depth estimation. *arXiv:1904.10230*, 2019.

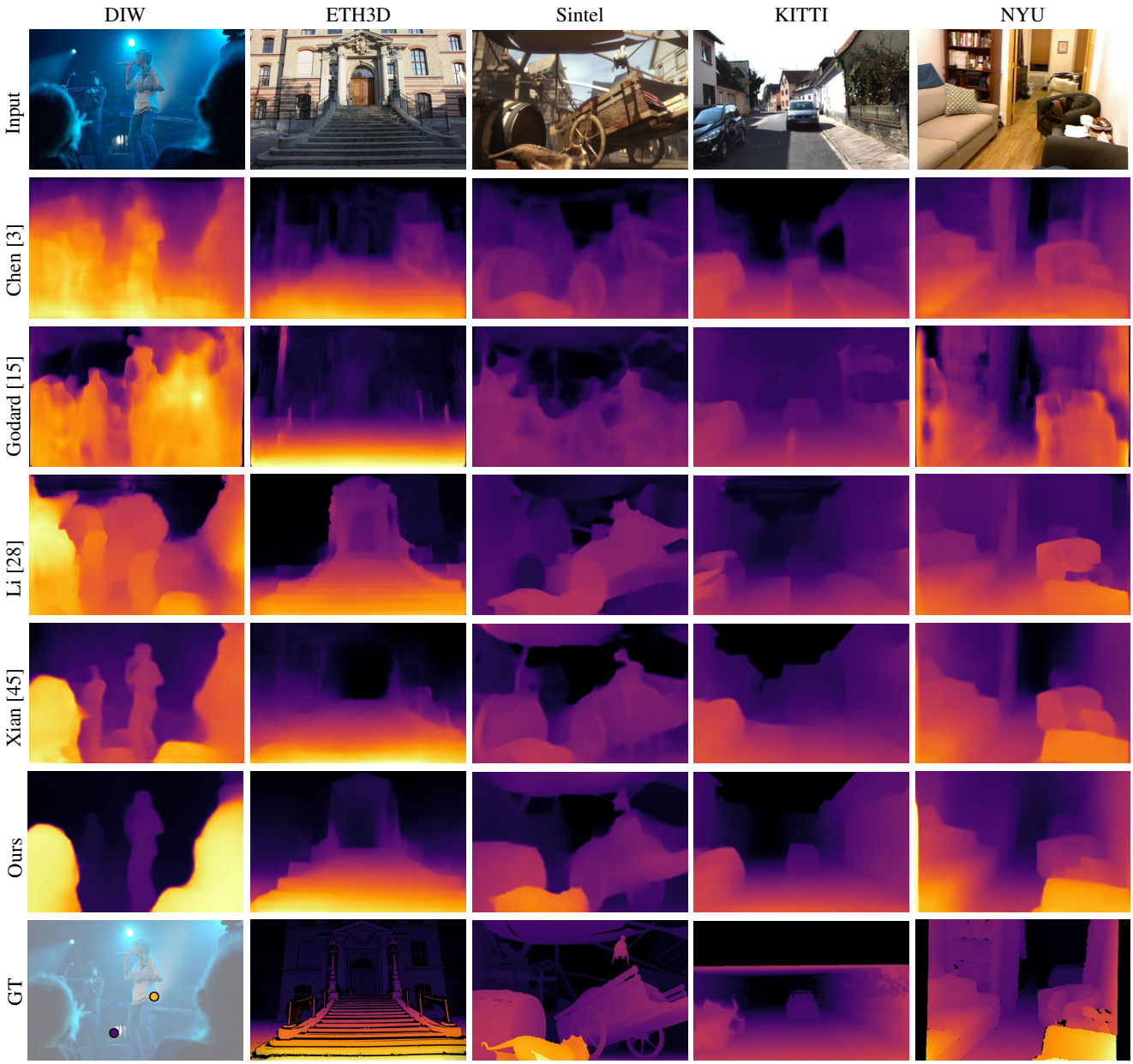

Figure 3. Qualitative comparison of our approach to various baselines. Ground truth on KITTI was interpolated from sparse LiDAR measurements for visualization. On DIW the yellow and purple dots represent sparse human annotations for close and far, respectively.

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[7] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017.

[8] F. Devernay and P. A. Beardsley. Stereoscopic cinema. In *Image and Geometry Processing for 3-D Cinematography*. Springer, 2010.

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

[10] P. Fankhauser, M. Blösch, D. Rodriguez, R. Kaestner, M. Hutter, and R. Siegwart. Kinect v2 for mobile robot navigation: Evaluation and modeling. In *International Conference on Advanced Robotics*, 2015.

[11] FFmpeg developers. FFmpeg. `https://ffmpeg.org`, 2018.

[12] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018.

[13] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016.

[14] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In

CVPR, 2012.

[15] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[16] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. *arXiv:1904.04998*, 2019.

[17] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang. Learning monocular depth by distilling cross-domain stereo networks. In *ECCV*, 2018.

[18] S. Hadfield, K. Lebeda, and R. Bowden. Hollywood 3D: What are the best 3D features for action recognition? *IJCV*, 121(1), 2017.

[19] M. Hansard, S. Lee, O. Choi, and R. Horaud. *Time-of-Flight Cameras: Principles, Methods and Applications*. Springer, 2013.

[20] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. *ACM Transactions on Graphics*, 24(3), 2005.

[21] K. Karsch, C. Liu, and S. B. Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *PAMI*, 36(11), 2014.

[22] K. Khoshelham and S. O. Elberink. Accuracy and resolution of Kinect depth data for indoor mapping applications. *Sensors*, 12(2), 2012.

[23] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[24] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

[25] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016.

[26] R. Li, K. Xian, C. Shen, Z. Cao, H. Lu, and L. Hang. Deep attention-based classification network for robust depth prediction. In *ACCV*, 2018.

[27] Z. Li, T. Dekel, F. Cole, R. Tucker, N. Snavely, C. Liu, and W. T. Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, 2019.

[28] Z. Li and N. Snavely. MegaDepth: Learning single-view depth prediction from Internet photos. In *CVPR*, 2018.

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

[30] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin. Single view stereo matching. In *CVPR*, 2018.

[31] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints. In *CVPR*, 2018.

[32] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015.

[33] J. M. M. Montiel, J. Civera, and A. J. Davison. Unified inverse depth parametrization for monocular SLAM. In *Robotics: Science and Systems*, 2006.

[34] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. J. V. Gool, M. H. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.

[35] S. Rota Bulò, L. Porzi, and P. Kontschieder. In-place activated batchnorm for memory-optimized training of DNNs. In *CVPR*, 2018.

[36] A. Roy and S. Todorovic. Monocular depth estimation using neural regression forest. In *CVPR*, 2016.

[37] A. Saxena, M. Sun, and A. Y. Ng. Make3D: Learning 3D scene structure from a single still image. *PAMI*, 31(5), 2009.

[38] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017.

[39] O. Sener and V. Koltun. Multi-task learning as multi-objective optimization. In *NeurIPS*, 2018.

[40] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012.

[41] S. Song, S. P. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, 2015.

[42] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018.

[43] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. DeMoN: Depth and motion network for learning monocular stereo. In *CVPR*, 2017.

[44] C. Wang, S. Lucey, F. Perazzi, and O. Wang. Web stereo video supervision for depth prediction from dynamic scenes. *arXiv:1904.11112*, 2019.

[45] K. Xian, C. Shen, Z. Cao, H. Lu, Y. Xiao, R. Li, and Z. Luo. Monocular relative depth perception with web stereo data supervision. In *CVPR*, 2018.

[46] J. Xie, R. B. Girshick, and A. Farhadi. Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In *ECCV*, 2016.

[47] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. D. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, 2018.

[48] B. Zhou, P. Krähenbühl, and V. Koltun. Does computer vision matter for action? *Science Robotics*, 4(30), 2019.

[49] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

Supplementary Material

A. 3D Movies Dataset

Table 6 shows the complete list of movies that were used for creating the 3D Movies dataset. We additionally state the number of extracted clips of one second (24 frames each). Note that discrepancies in the number of extracted clips per movie occur due to varying runtimes.

Movie title	# Clips
Training set	**38000**
Battle of the Year (2013)	2053
Billy Lynn's Long Halftime Walk (2016)	1645
Drive Angry (2011)	1722
Exodus: Gods and Kings (2014)	2847
Final Destination 5 (2011)	1502
A very Harold & Kumar 3D Christmas (2011)	1601
Hellbenders (2012)	1378
The Hobbit: An Unexpected Journey (2012)	2742
Hugo (2011)	2097
The Three Musketeers (2011)	1958
Nurse 3D (2013)	1397
Pina (2011)	1550
Dawn of the Planet of the Apes (2014)	2087
The Amazing Spider-Man (2012)	2240
Step Up 3D (2010)	1841
Step Up: All In (2014)	1876
Transformers: Age of Extinction (2014)	2903
Le Dernier Loup / Wolf Totem (2015)	1874
X-Men: Days of Future Past (2014)	2687
Validation set	**4435**
The Great Gatsby (2013)	2606
Step Up: Miami Heat / Revolution (2012)	1829
Test set	**2961**
Doctor Who - The Day of the Doctor (2013)	1447
StreetDance 2 (2012)	1514

Table 6. List of movies and the number of one-second clips in the 3D Movies dataset.

B. Scale- and Shift-invariant Loss

To see that the proposed loss is invariant to scale and shift of the prediction, let $\vec{\mathbf{d}}_i = (\mathbf{d}_i, 1)^\top$, $\mathbf{h} = (s, t)^\top$ and $\mathbf{A} \in \mathbb{R}^{2\times 2}$ be a diagonal matrix with strictly positive entries. We have

$$\mathcal{L}_{ssi}(\mathbf{d}, \mathbf{d}^*) = \min_{\mathbf{h}} \frac{1}{2M} \sum_{i=1}^{M} (\vec{\mathbf{d}}_i^\top \mathbf{A}\mathbf{h} - \mathbf{d}_i^*)^2 \tag{11}$$

The optimality condition is:

$$\sum_{i=1}^{M} \mathbf{A}\vec{\mathbf{d}}_i (\vec{\mathbf{d}}_i^\top \mathbf{A}\mathbf{h} - \mathbf{d}_i^*) = \mathbf{0} \tag{12}$$

$$\mathbf{A} \left(\sum_{i=1}^{M} \vec{\mathbf{d}}_i \vec{\mathbf{d}}_i^\top \right) \mathbf{A}\mathbf{h} - \mathbf{A} \left(\sum_{i=1}^{M} \vec{\mathbf{d}}_i \mathbf{d}_i^* \right) = \mathbf{0} \tag{13}$$

$$\mathbf{A}\mathbf{C}\mathbf{A}\mathbf{h} - \mathbf{A}\mathbf{b} = \mathbf{0} \tag{14}$$

$$\mathbf{h} = (\mathbf{A}\mathbf{C}\mathbf{A})^{-1}\mathbf{A}\mathbf{b} \tag{15}$$

$$\mathbf{h} = \mathbf{A}^{-1}\mathbf{C}^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{b} \tag{16}$$

$$\mathbf{h} = \mathbf{A}^{-1} \left(\sum_{i=1}^{M} \vec{\mathbf{d}}_i \vec{\mathbf{d}}_i^\top \right)^{-1} \left(\sum_{i=1}^{M} \vec{\mathbf{d}}_i \mathbf{d}_i^* \right) \tag{17}$$

Substitute $\mathbf{h}$ into (11) to see that $\mathbf{A}$ cancels out.

C. Details of Evaluation

Our evaluation in Table 4 of the main paper has been performed on the validation sets of MegaDepth [28] (2963 images) and the 3D movies dataset (4435 images). For DIW [3] we created a validation set of 10000 images from the DIW training set. For ReDWeb [45] we left out 360 images of the training set for validation. For KITTI [32] we used the Eigen test split of 697 images [9]. For NYU [40] we used the official test split of 654 images. For ETH3D

[38] and the MPI Sintel depth dataset [1] we used all images from the respective datasets with publicly available ground truth (454 and 1064 images, respectively). For comparisons to the state of the art (Table 5), we used the test set of DIW [3] with 74441 images.

Alignment. We align the scale and shift of all predictions (our models as well as baselines) to the ground truth before conducting evaluations. We perform the alignment in inverse depth space based on a least-squares criterion.

Depth cap. Following [15], we cap predictions at an appropriate maximum value for datasets that are evaluated in depth space (ETH3D, Sintel, KITTI, NYU). For ETH3D, KITTI, and NYU, the depth cap was set to the maximum ground truth depth value (72, 80, and 10 meters, respectively). For Sintel we evaluate on areas with ground truth depth smaller than 72 meters and accordingly use a depth cap of 72 meters.

Input resolution for evaluation. For the 3D Movies dataset, we use the center crop of size 320×192 pixels after downscaling the original image by a factor of four. For all other datasets, we downscale input images to a size that is as close as possible to the training size (320×192), while maintaining aspect ratio and ensuring that each dimension is divisible by 32 (a constraint imposed by the network architecture).

Resolution of evaluation. On the 3D Movies dataset, we evaluate at the resolution of the prediction (320×192). On the MegaDepth dataset, we follow the original evaluation protocol and evaluate at a resolution of 320×240 for landscape images and 240×320 for portrait images. On the remaining datasets (DIW, ReDWeb, ETH3D, Sintel, KITTI, and NYU), the prediction is upscaled to the original resolution of the input images and evaluated at the full resolution. For ETH3D, we rendered ground-truth depth maps from the 3D point clouds at a resolution of 1512×1008 pixels.

D. Additional Results

We show additional results of our best-performing model that was trained on all three datasets (RW+MD+MV) with the multi-task learning strategy (MGDA).

Supplementary video. In the supplementary video, we show qualitative results on the DAVIS video dataset [34]. Note that every frame was processed individually, i.e. no temporal information was used in any way. For each clip, the inverse depth maps were jointly scaled and shifted for visualization. The dataset consists of a diverse set of videos and includes humans, animals, and cars in action. This dataset was filmed with monocular cameras, hence, no ground truth depth information is available.

Additional qualitative results. We show qualitative results from the two movies in the test set (*Doctor Who - The Day of the Doctor* and *Streetdance 2*) in Figures 4 and 5. The ground truth (obtained from stereo matching [42]) is shown for reference. Invalid pixels have been masked out in the ground truth. Note that our method is able to handle various camera angles, complex scenes depicting multiple people or animals, atypical objects such as robots, and various static objects.

To further showcase the generalization ability of our model, Figure 6 provides qualitative results on the DIW test set [3]. We again show results on a diverse set of input images depicting various objects and scenes, including humans, mammals, birds, cars, helicopters in flight, and other man-made and natural objects. The images feature indoor, street and nature scenes, various lighting conditions, and various camera angles. Additionally, subject areas vary from close-up to long-range shots.

Failure cases. We identify common failure cases and biases of our model. As observed by [3], images have a natural bias where the lower parts of the image are closer to the camera than the higher image regions. When randomly sampling two points and classifying the lower point as closer to the camera, [3] achieved an agreement rate of 85.8% with human annotators. To some extent, this bias has also been learned by our network and can be observed in some extreme cases that are shown in Figure 7. In the example on the top, the model fails to recover the ground plane; likely because the input image was rotated by 90 degrees. In the bottom image, pellets at approximately the same distance to the camera are reconstructed closer to the camera in the lower part of the image. Such cases could be prevented by augmenting training data with rotated images. However, it is not clear if invariance to image rotations is a necessary or desired property for this task.

Some particularly interesting failure cases are shown in Figure 8. Paintings, photos, and mirrors are often not recognized as such, especially if they are very prominent in the image. The network estimates depth based on the content that is depicted on the reflector rather than predicting the depth of the reflector itself.

A selection of additional failure cases is shown in Figure 9. Strong edges in RGB space can lead to hallucinated depth discontinuities, resulting, for example, in heads being detached from the lower body. Thin structures can be missed by the network and relative depth arrangement between disconnected objects might fail in some situations (e.g. relative placement of people). The results tend to get blurred in background areas, which might be explained by the limited resolution of the input images and imperfect ground truth in the far range.

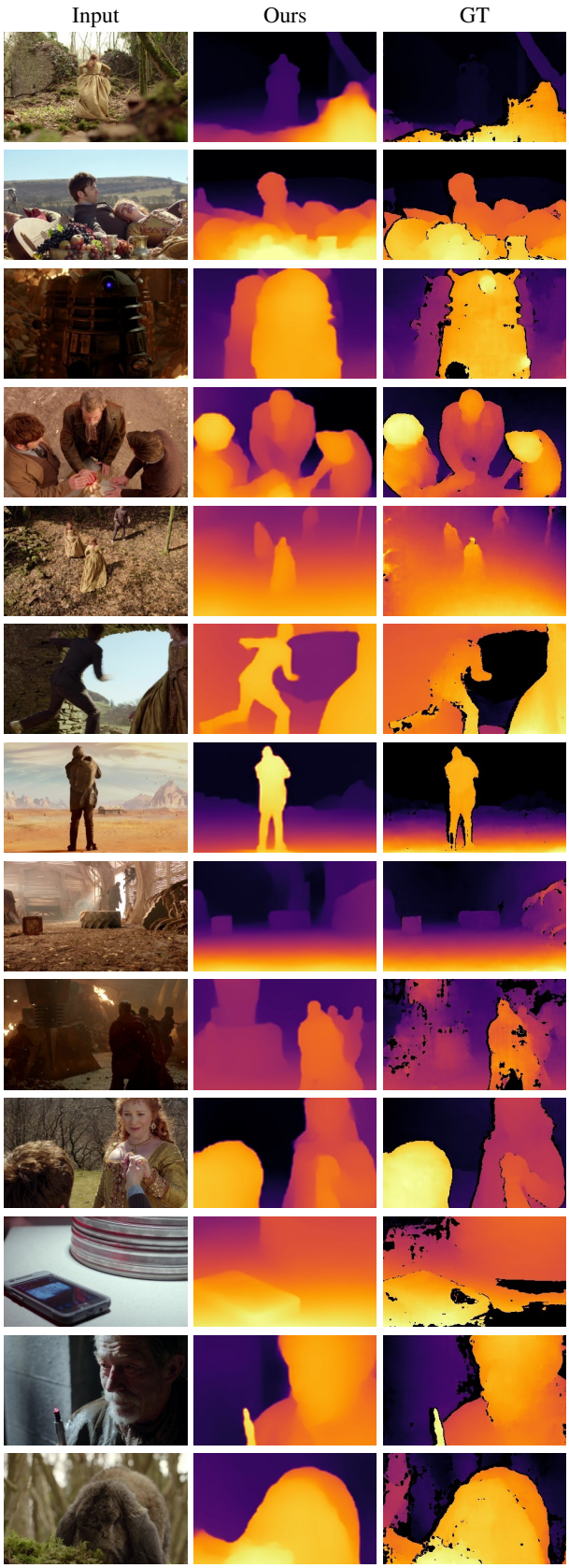

Figure 4. Qualitative results on the movies test set (Doctor Who).

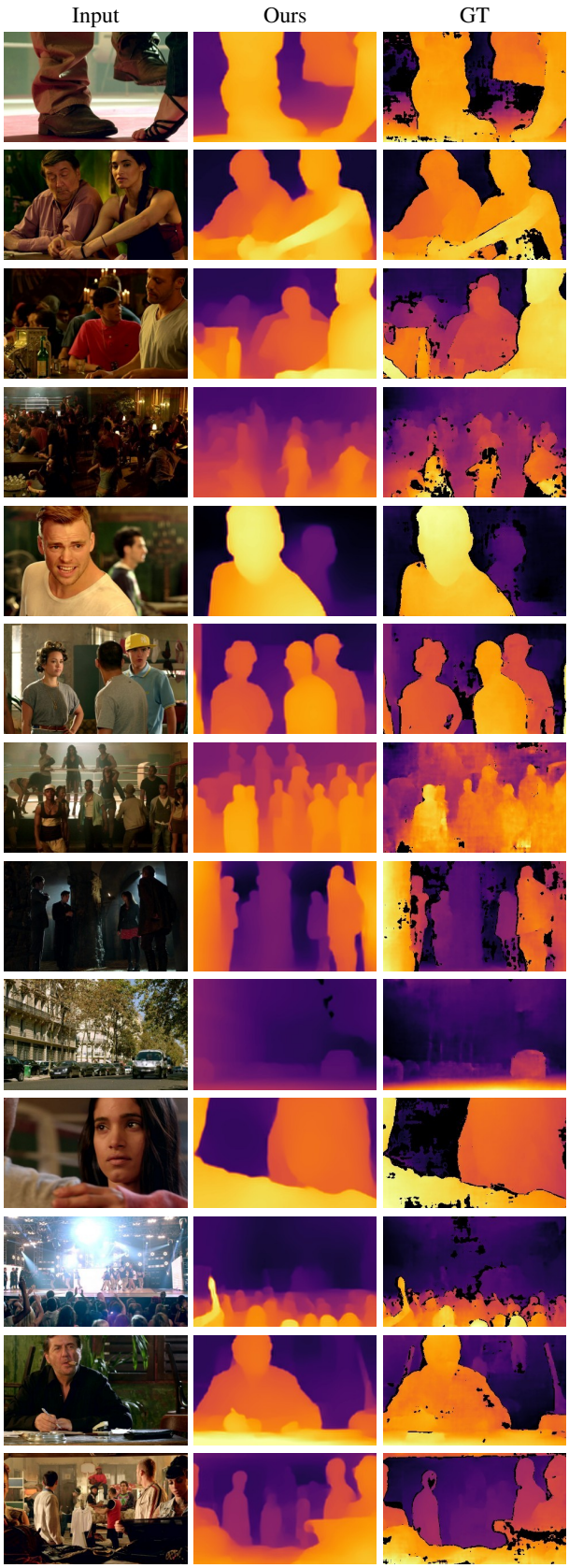

Figure 5. Qualitative results on the movies test set (Streetdance 2).

Figure 6. Qualitative results on the DIW test set.

Figure 6 (cont.). Qualitative results on the DIW test set.

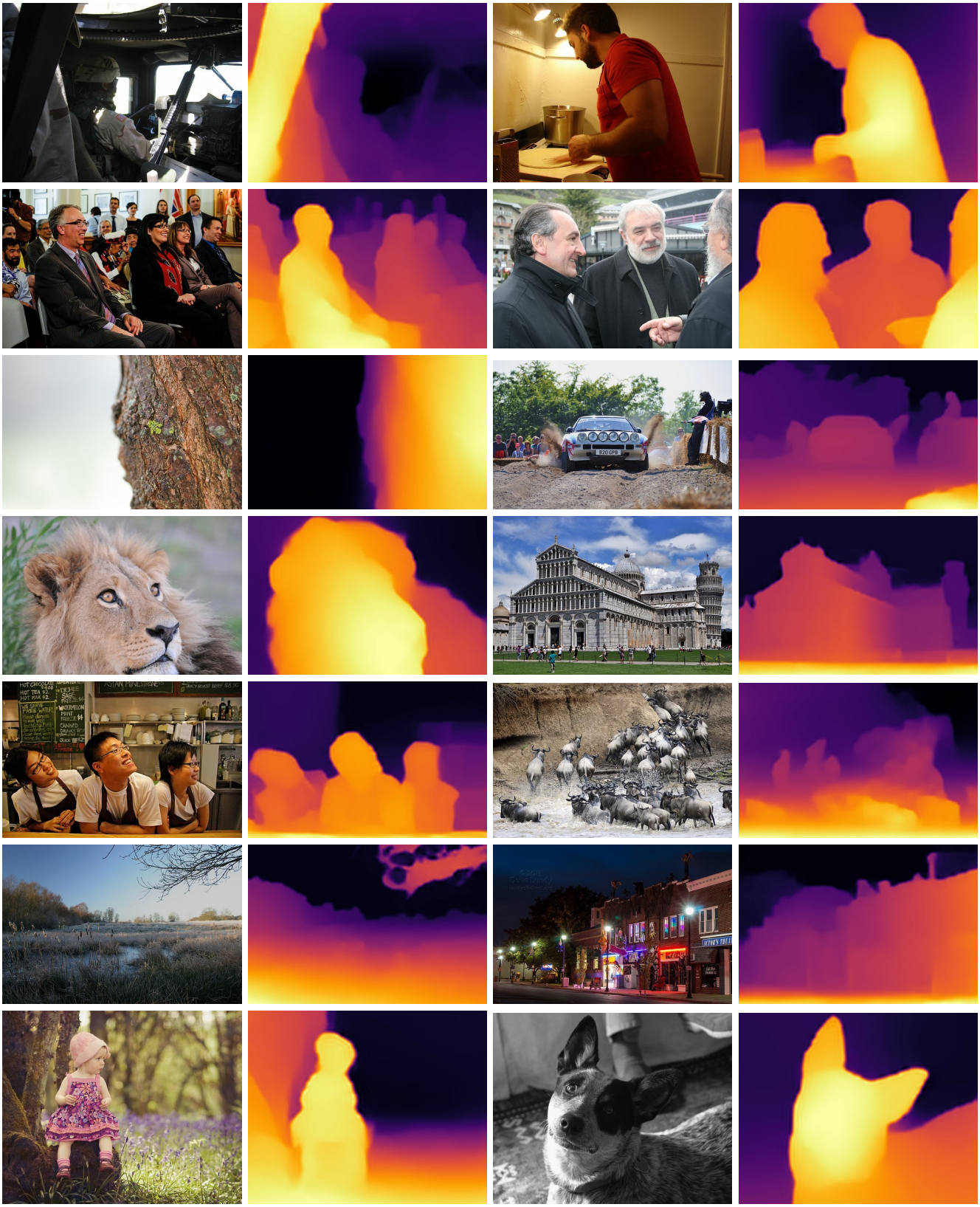

Figure 6 (cont.). Qualitative results on the DIW test set.

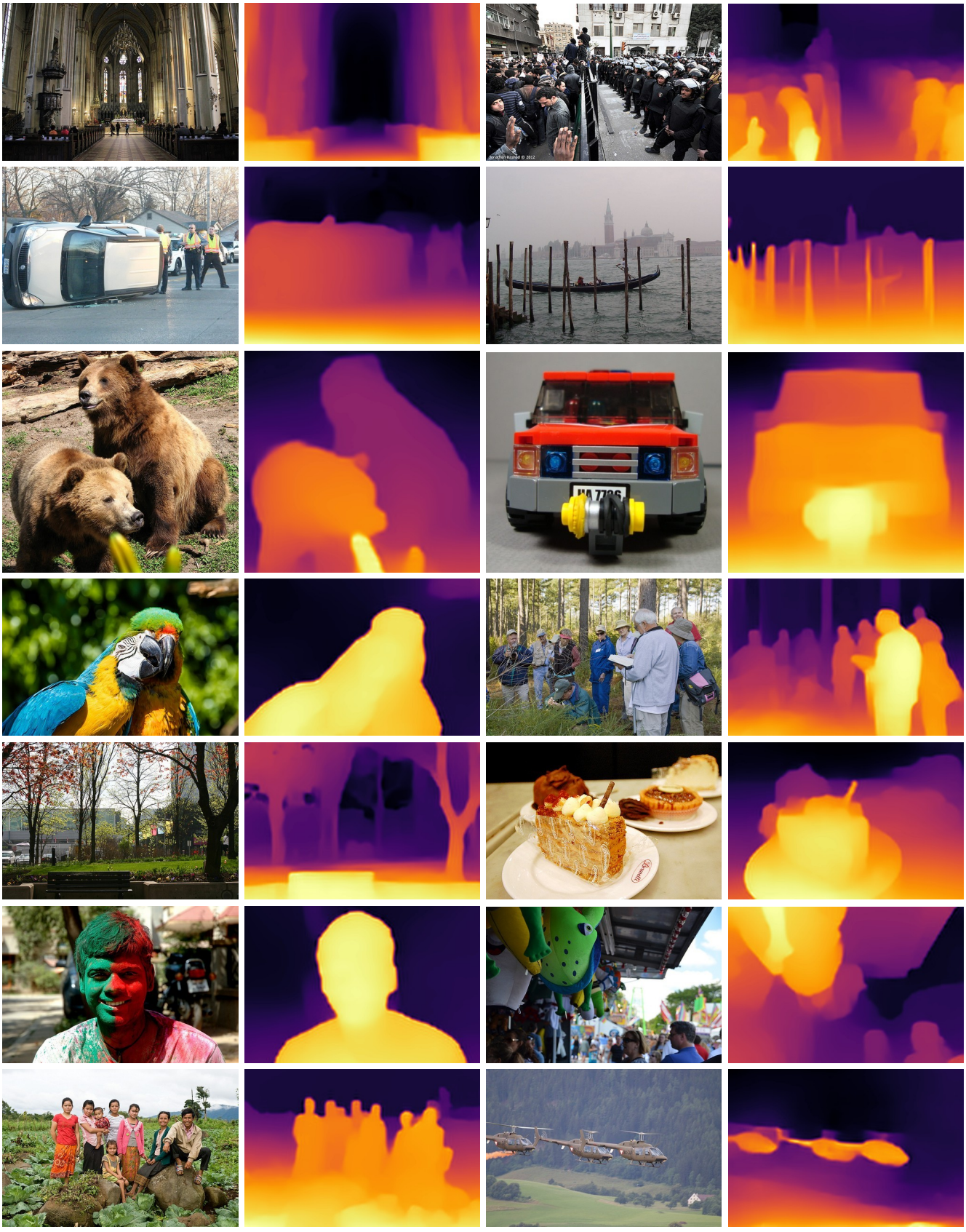

Figure 6 (cont.). Qualitative results on the DIW test set.

Figure 6 (cont.). Qualitative results on the DIW test set.

Figure 7. Failure cases: Bias of lower regions being closer to the camera.

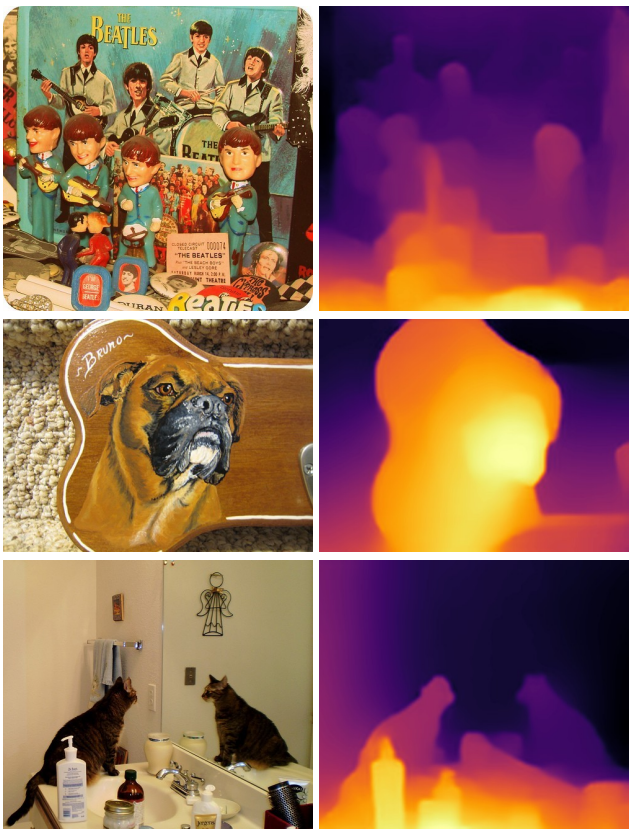

Figure 8. Failure cases: Pictures and mirrors.

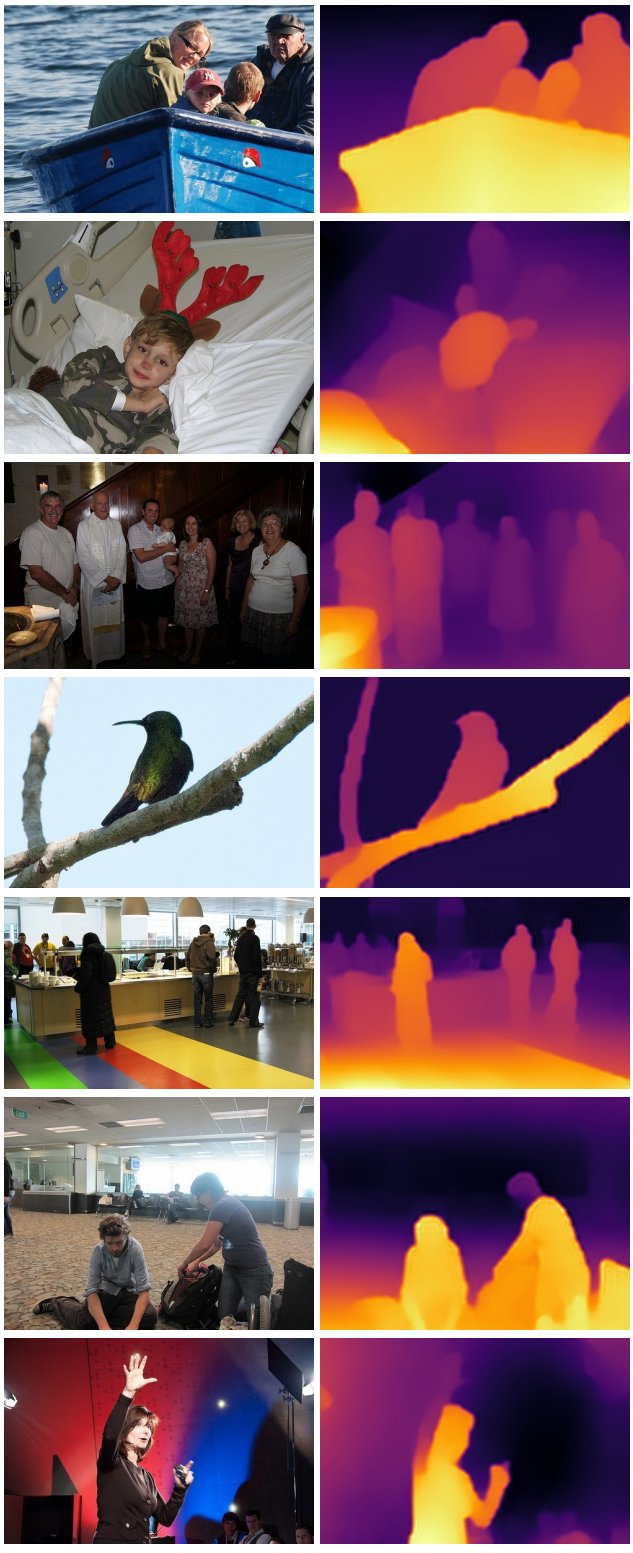

Figure 9. Failure cases: Relative depth arrangement, spurious depth discontinuities at strong RGB edges, and related problems.