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

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, *Senior Member, IEEE*, and Vladlen Koltun

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 robust training objective that is invariant to changes in depth range and scale, advocate the use of principled multi-objective learning to combine data from different sources, and highlight the importance of pretraining encoders on auxiliary tasks. Armed with these tools, we experiment with five diverse training datasets, including a new, massive data source: 3D films. To demonstrate the generalization power of our approach we use *zero-shot cross-dataset transfer*, i.e. we evaluate on datasets that were not seen during training. The experiments confirm that mixing data from complementary sources greatly improves monocular depth estimation. Our approach clearly outperforms competing methods across diverse datasets, setting a new state of the art for monocular depth estimation.

Index Terms—Monocular depth estimation, single-image depth prediction, zero-shot cross-dataset transfer, multi-dataset training

1 Introduction

Depth is among the most useful intermediate representations for action in physical environments [1]. Despite its utility, monocular depth estimation remains a challenging problem that is heavily underconstrained. To solve it, one must exploit many, sometimes subtle, visual cues, as well as long-range context and prior knowledge. This calls for learning-based techniques [2], [3].

To learn models that are effective across a variety of scenarios, we need training data that is equally varied and captures the diversity of the visual world. The key challenge is to acquire 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 and operating conditions [6], [7], [8]. Laser scanners are expensive and can only provide sparse depth measurements when the scene is in motion. Stereo cameras are a promising source of data [9], [10], but collecting suitable 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 [11], 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 is sufficiently rich to support the training of a model that works robustly on real images of 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 investigate ways to train robust monocular depth estimation models that are expected to perform across diverse environments. We develop novel loss functions that are invariant to the major sources of incompatibility between datasets, including unknown and inconsistent scale and baselines. Our losses enable 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 also quantify the value of a variety of existing datasets for monocular depth estimation and explore optimal strategies for mixing datasets during training. In particular, we show that a principled approach based on multi-objective optimization [12] leads to improved results compared to a naive mixing strategy. We further empirically highlight the importance of high-capacity encoders, and show the unreasonable effectiveness of pretraining the encoder on a large-scale auxiliary task.

Our extensive experiments, which cover approximately six GPU months of computation, show that a model trained on a rich and diverse set of images from different sources, with an appropriate training procedure, delivers state-of-the-art results across a variety of environments. To demonstrate this, we use the experimental protocol of *zero-shot cross-dataset transfer*. That is, we train a model on certain datasets and then test its performance on other datasets that were never seen during training. The intuition is that zero-shot

- *René Ranftl and David Hafner are with the Intelligent Systems Lab, Intel Labs, 85579 Munich, Germany. E-mail: {rene.ranftl, david.hafner}@intel.com.*
- *Vladlen Koltun is with the Intelligent Systems Lab, Intel Labs, Santa Clara, CA 95054 USA. E-mail: vladlen.koltun@intel.com.*
- *Katrin Lasinger and Konrad Schindler are with the Institute of Geodesy and Photogrammetry, ETH, 8093 Zürich, Switzerland. E-mail: {katrin.lasinger, konrad.schindler}@geod.baug.ethz.ch.*

Manuscript received 13 Mar. 2020; revised 21 July 2020; accepted 21 Aug. 2020. Date of publication 27 Aug. 2020; date of current version 3 Feb. 2022. (Corresponding author: René Ranftl.) Recommended for acceptance by T. Hassner. Digital Object Identifier no. 10.1109/TPAMI.2020.3019967

Fig. 1. We show how to leverage training data from multiple, complementary sources for single-view depth estimation, in spite of varying and unknown depth range and scale. Our approach enables strong generalization across datasets. Top: input images. Middle: inverse depth maps predicted by the presented approach. Bottom: corresponding point clouds rendered from a novel view-point. Point clouds rendered via Open3D [4]. Input images from the Microsoft COCO dataset [5], which was not seen during training.

cross-dataset performance is a more faithful proxy of "real world" performance than training and testing on subsets of a single data collection that largely exhibit the same biases [13].

In an evaluation across six different datasets, we outperform prior art both quantitatively and qualitatively, and set a new state of the art for monocular depth estimation. Example results are shown in Fig. 1.

2 Related Work

Early work on monocular depth estimation used MRF-based formulations [3], simple geometric assumptions [2], or non-parametric methods [14]. More recently, significant advances have been made by leveraging the expressive power of convolutional networks to directly regress scene depth from the input image [15]. Various architectural innovations have been proposed to enhance prediction accuracy [16], [17], [18], [19], [20]. These methods need ground-truth depth for training, which is commonly acquired using RGB-D cameras or LiDAR sensors. Others leverage existing stereo matching methods to obtain ground truth for supervision [21], [22]. These methods tend to work well in the specific type of scenes used to train them, but do not generalize well to unconstrained scenes, due to the limited scale and diversity of the training data.

Garg *et al.* [9] proposed to use calibrated stereo cameras for self-supervision. While this significantly simplifies the acquisition of training data, it still does not lift the restriction to a very specific data regime. Since then, various approaches leverage self-supervision, but they either require stereo images [10], [23], [24] or exploit apparent motion [24], [25], [26], [27], and are thus difficult to apply to dynamic scenes.

We argue that high-capacity deep models for monocular depth estimation can in principle operate on a fairly wide and unconstrained range of scenes. What limits their performance is the lack of large-scale, dense ground truth that spans such a wide range of conditions. Commonly used datasets feature homogeneous scene layouts, such as street scenes in a specific geographic region [3], [28], [29] or indoor environments [30]. We note in particular that these datasets show only a small number of dynamic objects. Models that are trained on data with such strong biases are prone to fail in less constrained environments.

Efforts have been made to create more diverse datasets. Chen *et al.* [34] used crowd-sourcing to sparsely annotate ordinal relations in images collected from the web. Xian *et al.* [32] collected a stereo dataset from the web and used off-the-shelf tools to extract dense ground-truth disparity; while this dataset is fairly diverse, it only contains 3,600 images. Li and Snavely [11] used SfM and multi-view stereo (MVS) to reconstruct many (predominantly static) 3D scenes for supervision. Li *et al.* [38] used SfM and MVS to construct a dataset from videos of people imitating mannequins (*i.e.* they are frozen in action while the camera moves through the scene). Chen *et al.* [39] propose an approach to automatically assess the quality of sparse SfM reconstructions in order to construct a large dataset. Wang *et al.* [33] build a large dataset from stereo videos sourced from the web, while Cho *et al.* [40] collect a dataset of outdoor scenes with handheld stereo cameras. Gordon *et al.* [41] estimate the intrinsic parameters of YouTube videos in

TABLE 1
Datasets Used in Our Work

Dataset	Indoor	Outdoor	Dynamic	Video	Dense	Accuracy	Diversity	Annotation	Depth	# Images
DIML Indoor [31]	✓			✓	✓	Medium	Medium	RGB-D	**Metric**	220K
MegaDepth [11]		✓	(✓)		(✓)	Medium	Medium	SfM	No scale	130K
ReDWeb [32]	✓	✓	✓		✓	Medium	**High**	Stereo	No scale & shift	3600
WSVD [33]	✓	✓	✓	✓	✓	Medium	**High**	Stereo	No scale & shift	1.5M
3D Movies	✓	✓	✓	✓	✓	Medium	**High**	Stereo	No scale & shift	75K
DIW [34]	✓	✓	✓			Low	**High**	User clicks	Ordinal pair	496K
ETH3D [35]	✓	✓			✓	**High**	Low	Laser	**Metric**	454
Sintel [36]	✓	✓	✓	✓	✓	**High**	Medium	Synthetic	(Metric)	1064
KITTI [28], [29]		✓	(✓)	✓	(✓)	Medium	Low	Laser/Stereo	**Metric**	93K
NYUDv2 [30]	✓		(✓)	✓	✓	Medium	Low	RGB-D	**Metric**	407K
TUM-RGBD [37]	✓		(✓)	✓	✓	Medium	Low	RGB-D	**Metric**	80K

Top: Our training sets. Bottom: Our test sets. No single real-world dataset features a large number of diverse scenes with dense and accurate ground truth.

order to leverage them for training. Large-scale datasets that were collected from the Internet [33], [38] require a large amount of pre- and post-processing. Due to copyright restrictions, they often only provide links to videos, which frequently become unavailable. This makes reproducing these datasets challenging.

To the best of our knowledge, the controlled mixing of multiple data sources has not been explored before in this context. Ummenhofer *et al.* [42] presented a model for two-view structure and motion estimation and trained it on a dataset of (static) scenes that is the union of multiple smaller datasets. However, they did not consider strategies for optimal mixing, or study the impact of combining multiple datasets. Similarly, Facil *et al.* [43] used multiple datasets with a naive mixing strategy for learning monocular depth with known camera intrinsics. Their test data is very similar to half of their training collection, namely RGB-D recordings of indoor scenes.

3 Existing Datasets

Various datasets have been proposed that are suitable for monocular depth estimation, *i.e.* they consist of RGB images with corresponding depth annotation of some form [3], [11], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [40], [44], [45], [46], [47], [48]. Datasets differ in captured environments and objects (indoor/outdoor scenes, dynamic objects), type of depth annotation (sparse/dense, absolute/relative depth), accuracy (laser, time-of-flight, SfM, stereo, human annotation, synthetic data), image quality and camera settings, as well as dataset size.

Each single dataset comes with its own characteristics and has its own biases and problems [13]. High-accuracy data is hard to acquire at scale and problematic for dynamic objects [35], [47], whereas large data collections from Internet sources come with limited image quality and depth accuracy as well as unknown camera parameters [33], [34]. Training on a single dataset leads to good performance on the corresponding test split of the same dataset (same camera parameters, depth annotation, environment), but may have limited generalization capabilities to unseen data with different characteristics. Instead, we propose to train on a collection of datasets, and demonstrate that this approach leads to strongly enhanced generalization by testing on diverse datasets that were not seen during training. We list our training and test datasets, together with their individual characteristics, in Table 1.

Training Datasets. We experiment with five existing and complementary datasets for training. ReDWeb [32] (RW) is a small, heavily curated dataset that features diverse and dynamic scenes with ground truth that was acquired with a relatively large stereo baseline. MegaDepth [11] (MD) is much larger, but shows predominantly static scenes. The ground truth is usually more accurate in background regions since wide-baseline multi-view stereo reconstruction was used for acquisition. WSVD [33] (WS) consists of stereo videos obtained from the web and features diverse and dynamic scenes. This dataset is only available as a collection of links to the stereo videos. No ground truth is provided. We thus recreate the ground truth according to the procedure outlined by the original authors. DIML Indoor [31] (DL) is an RGB-D dataset of predominantly static indoor scenes, captured with a Kinect v2.

Test Datasets. To benchmark the generalization performance of monocular depth estimation models, we chose six datasets based on diversity and accuracy of their ground truth. DIW [34] is highly diverse but provides ground truth only in the form of sparse ordinal relations. ETH3D [35] features highly accurate laser-scanned ground truth on static scenes. Sintel [36] features perfect ground truth for synthetic scenes. KITTI [29] and NYU [30] are commonly used datasets with characteristic biases. For the TUM dataset [37], we use the *dynamic* subset that features humans in indoor environments [38]. Note that we never fine-tune models on any of these datasets. We refer to this experimental procedure as *zero-shot cross-dataset transfer*.

4 3D Movies

To complement the existing datasets we propose a new data source: 3D movies (MV). 3D movies feature high-quality video frames in a variety of 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 (similar to RW and WS). Our driving motivation is the scale and diversity of the data. 3D movies provide the largest known source of stereo pairs that were captured in carefully controlled conditions. This offers the

possibility of tapping into millions of high-quality images from an ever-growing library of content. We note that 3D movies have been used in related tasks in isolation [49], [50]. We will show that their full potential is unlocked by combining them with other, complementary data sources. In contrast to similar data collections in the wild [32], [33], [38], no manual filtering of problematic content was required with this data source. Hence, the dataset can easily be extended or adapted to specific needs (*e.g.* focus on dancing humans or nature documentaries).

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 [51]. 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 [52].

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 following 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.

We extract stereo image pairs at 1920 x 1080 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 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.

We use the scene detection tool of FFmpeg [53] 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. Since multiple frames are part of the same clip, the complete dataset is highly correlated. Hence, we further subsample the training set at 4 fps and the test and validation sets at 1 fps.

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 that are common in standard stereo setups due to the limited depth budget.

To alleviate these problems, we apply a modern optical flow algorithm [54] 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 2 pixels as invalid. We automatically filter out frames of bad disparity quality following the guidelines of Wang *et al.* [33]: frames are rejected if more than 10 percent of all pixels have a vertical disparity >2 pixels, the horizontal disparity range is <10 pixels, or the percentage of pixels passing the left-right consistency check is <70 percent. In a final step, we detect pixels that belong to sky regions using a pre-trained semantic segmentation model [55] and set their disparity to the minimum disparity in the image.

The complete list of selected movies together with the number of frames that remain after filtering with the automatic cleaning pipeline is shown in Table 2. Note that discrepancies in the number of extracted frames per movie occur due to varying runtimes as well as varying disparity quality. We use frames from 19 movies for training and set aside two movies for validation and two movies for testing, respectively. Example frames from the resulting dataset are shown in Fig. 2.

5 Training on Diverse Data

Training models for monocular depth estimation on diverse datasets presents a challenge because the ground truth comes in different forms (see Table 1). It may be 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.

We identify three major challenges. 1) Inherently different representations of depth: direct versus inverse depth representations. 2) Scale ambiguity: for some data sources, depth is only given up to an unknown scale. 3) Shift ambiguity: some datasets provide disparity only up to an unknown

TABLE 2
List of Films and the Number of Extracted Frames in the 3D Movies Dataset After Automatic Processing

Movie title	# frames
Training set	**75074**
Battle of the Year (2013)	4821
Billy Lynn's Long Halftime Walk (2016)	4178
Drive Angry (2011)	328
Exodus: Gods and Kings (2014)	8063
Final Destination 5 (2011)	1437
A very Harold & Kumar 3D Christmas (2011)	3690
Hellbenders (2012)	120
The Hobbit: An Unexpected Journey (2012)	8874
Hugo (2011)	3189
The Three Musketeers (2011)	5028
Nurse 3D (2013)	492
Pina (2011)	1215
Dawn of the Planet of the Apes (2014)	5571
The Amazing Spider-Man (2012)	5618
Step Up 3D (2010)	509
Step Up: All In (2014)	2187
Transformers: Age of Extinction (2014)	8740
Le Dernier Loup / Wolf Totem (2015)	4843
X-Men: Days of Future Past (2014)	6171
Validation set	**3058**
The Great Gatsby (2013)	1815
Step Up: Miami Heat / Revolution (2012)	1243
Test set	**788**
Doctor Who - The Day of the Doctor (2013)	508
StreetDance 2 (2012)	280

scale and global disparity shift that is a function of the unknown baseline and a horizontal shift of the principal points due to post-processing [33].

Scale- and Shift-Invariant Losses. We propose to perform prediction in disparity space (inverse depth up to scale and shift) together with a family of scale- and shift-invariant dense losses 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 a disparity prediction and let $\mathbf{d}^* \in \mathbb{R}^M$ be the corresponding ground-truth disparity. Individual pixels are indexed by subscripts.

We define the scale- and shift-invariant loss for a single sample as

$$\mathcal{L}_{ssi}(\hat{\mathbf{d}}, \hat{\mathbf{d}}^*) = \frac{1}{2M} \sum_{i=1}^{M} \rho\Big(\hat{\mathbf{d}}_i - \hat{\mathbf{d}}_i^*\Big), \tag{1}$$

where $\hat{\mathbf{d}}$ and $\hat{\mathbf{d}}^*$ are scaled and shifted versions of the predictions and ground truth, and ρ defines the specific type of loss function.

Let $s : \mathbb{R}^M \rightarrow \mathbb{R}_+$ and $t : \mathbb{R}^M \rightarrow \mathbb{R}$ denote estimators of the scale and translation. To define a meaningful scale- and shift-invariant loss, a sensible requirement is that prediction and ground truth should be appropriately aligned with respect to their scale and shift, *i.e.* we need to ensure that $s(\hat{\mathbf{d}}) \approx s(\hat{\mathbf{d}}^*)$ and $t(\hat{\mathbf{d}}) \approx t(\hat{\mathbf{d}}^*)$. We propose two different strategies for performing this alignment.

The first approach aligns the prediction to the ground truth based on a least-squares criterion:

$$\begin{gathered}(s, t) = \arg\min_{s,t} \sum_{i=1}^{M} \big(s\mathbf{d}_i + t - \mathbf{d}_i^*\big)^2, \\ \hat{\mathbf{d}} = s\mathbf{d} + t, \quad \hat{\mathbf{d}}^* = \mathbf{d}^*,\end{gathered} \tag{2}$$

where $\hat{\mathbf{d}}$ and $\hat{\mathbf{d}}^*$ are the aligned prediction and ground truth, respectively. The factors s and t can be efficiently determined in closed form by rewriting (2) as a standard least-squares problem: Let $\vec{\mathbf{d}}_i = (\mathbf{d}_i, 1)^\top$ and $\mathbf{h} = (s, t)^\top$, then we can rewrite the objective as

$$\mathbf{h}^{opt} = \arg\min_{\mathbf{h}} \sum_{i=1}^{M} \Big(\vec{\mathbf{d}}_i^\top \mathbf{h} - \mathbf{d}_i^*\Big)^2, \tag{3}$$

which has the closed-form solution

$$\mathbf{h}^{opt} = \left(\sum_{i=1}^{M} \vec{\mathbf{d}}_i \vec{\mathbf{d}}_i^\top\right)^{-1} \left(\sum_{i=1}^{M} \vec{\mathbf{d}}_i \mathbf{d}_i^*\right). \tag{4}$$

We set $\rho(x) = \rho_{mse}(x) = x^2$ to define the scale- and shift-invariant mean-squared error (MSE). We denote this loss as $\mathcal{L}_{ssimse}$.

The MSE is not robust to the presence of outliers. Since all existing large-scale datasets only provide imperfect ground truth, we conjecture that a robust loss function can improve training. We thus define alternative, robust loss functions based on robust estimators of scale and shift:

$$t(\mathbf{d}) = \text{median}(\mathbf{d}), \quad s(\mathbf{d}) = \frac{1}{M} \sum_{i=1}^{M} |\mathbf{d} - t(\mathbf{d})|. \tag{5}$$

We align both the prediction and the ground truth to have zero translation and unit scale:

$$\hat{\mathbf{d}} = \frac{\mathbf{d} - t(\mathbf{d})}{s(\mathbf{d})}, \quad \hat{\mathbf{d}}^* = \frac{\mathbf{d}^* - t(\mathbf{d}^*)}{s(\mathbf{d}^*)}. \tag{6}$$

We define two robust losses. The first, which we denote as $\mathcal{L}_{ssimae}$, measures the absolute deviations $\rho_{mae}(x) = |x|$. We define the second robust loss by trimming the 20 percent largest residuals in every image, irrespective of their magnitude:

$$\mathcal{L}_{ssitrim}(\hat{\mathbf{d}}, \hat{\mathbf{d}}^*) = \frac{1}{2M} \sum_{j=1}^{U_m} \rho_{mae}\Big(\hat{\mathbf{d}}_j - \hat{\mathbf{d}}_j^*\Big), \tag{7}$$

with $|\hat{\mathbf{d}}_j - \hat{\mathbf{d}}_j^*| \leq |\hat{\mathbf{d}}_{j+1} - \hat{\mathbf{d}}_{j+1}^*|$ and $U_m = 0.8M$ (set empirically based on experiments on the ReDWeb dataset). Note that this is in contrast to commonly used M-estimators, where the influence of large residuals is merely down-weighted. Our reasoning for trimming is that outliers in the ground truth should never influence training.

Related 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.* [15] proposed a scale-invariant loss in log-depth space. Their loss can be written as

$$\mathcal{L}_{silog}(\mathbf{z}, \mathbf{z}^*) = \min_{s} \frac{1}{2M} \sum_{i=1}^{M} \big(\log\left(e^s \mathbf{z}_i\right) - \log\left(\mathbf{z}_i^*\right)\big)^2, \tag{8}$$

where $\mathbf{z}_i = \mathbf{d}_i^{-1}$ and $\mathbf{z}_i^* = (\mathbf{d}_i^*)^{-1}$ are depths up to unknown scale. Both (8) and $\mathcal{L}_{ssimse}$ account for the unknown scale of

Fig. 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.

the predictions, but only $\mathcal{L}_{ssimse}$ accounts for an unknown global disparity shift. Moreover, the losses are evaluated on different depth representations. Our loss is defined in disparity space, which is numerically stable and compatible with common representations of relative depth.

Chen *et al.* [34] proposed a generally applicable loss for relative depth estimation based on ordinal relations:

$$\rho_{ord}(\mathbf{z}_i - \mathbf{z}_j) = \begin{cases} \log\left(1 + \exp(-(\mathbf{z}_i - \mathbf{z}_j) l_{ij}\right), & l_{ij} \neq 0 \\ (\mathbf{z}_i - \mathbf{z}_j)^2, & l_{ij} = 0, \end{cases} \tag{9}$$

where $l_{ij} \in \{-1, 0, 1\}$ encodes the ground-truth ordinal relation of point pairs. This encourages pushing points as far apart as possible when $l_{ij} \neq 0$ and pulling them to the same depth when $l_{ij} = 0$. Xian *et al.* [32] suggest to sparsely evaluate this loss by randomly sampling point pairs from the dense ground truth. In contrast, our proposed losses take all available data into account.

Recently, Wang *et al.* [33] proposed the normalized multiscale gradient (NMG) loss. To achieve shift invariance in addition to scale invariance in disparity space, they evaluate the gradient difference between ground-truth and rescaled estimates at multiple scales k:

$$\mathcal{L}_{nmg}(\mathbf{d}, \mathbf{d}^*) = \sum_{k=1}^{K} \sum_{i=1}^{M} |s\nabla_x^k \mathbf{d} - \nabla_x^k \mathbf{d}^*| + |s\nabla_y^k \mathbf{d} - \nabla_y^k \mathbf{d}^*|. \tag{10}$$

In contrast, our losses are evaluated directly on the ground-truth disparity values, while also accounting for unknown scale and shift. While both the ordinal loss and NMG can, conceptually, be applied to arbitrary depth representations and are thus suited for mixing diverse datasets, we will show that our scale- and shift-invariant loss variants lead to consistently better performance.

Final Loss. To define the complete loss, we adapt the multi-scale, scale-invariant gradient matching term [11] to the disparity space. This term biases discontinuities to be sharp and to coincide with discontinuities in the ground truth. We define the gradient matching term as

$$\mathcal{L}_{reg}(\hat{\mathbf{d}}, \hat{\mathbf{d}}^*) = \frac{1}{M} \sum_{k=1}^{K} \sum_{i=1}^{M} \left(|\nabla_x R_i^k| + |\nabla_y R_i^k| \right), \tag{11}$$

where $R_i = \hat{\mathbf{d}}_i - \hat{\mathbf{d}}_i^*$, and R^k denotes the difference of disparity maps at scale k. We use $K = 4$ scale levels, halving the image resolution at each level. Note that this term is similar to $\mathcal{L}_{nmg}$, but with different approaches to compute the scaling s.

Our final loss for a training set l is

$$\mathcal{L}_l = \frac{1}{N_l} \sum_{n=1}^{N_l} \mathcal{L}_{ssi}\left(\hat{\mathbf{d}}^n, (\hat{\mathbf{d}}^*)^n\right) + \alpha\, \mathcal{L}_{reg}\left(\hat{\mathbf{d}}^n, (\hat{\mathbf{d}}^*)^n\right), \tag{12}$$

where N_l is the training set size and α is set to 0.5.

Mixing Strategies. While our loss and choice of prediction space enable 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, regardless of their individual size.

Our second strategy explores a more principled approach, where we adapt a recent procedure for Pareto-optimal multi-task learning to our setting [12]. We define learning on each dataset as a separate task and seek an approximate Pareto optimum over datasets (*i.e.* a solution where the loss cannot be decreased on any training set without increasing it for at least one of the others). Formally, we use the algorithm presented in [12] to minimize the multi-objective optimization criterion

$$\min_{\boldsymbol{\theta}} \left(\mathcal{L}_1(\boldsymbol{\theta}), \ldots, \mathcal{L}_L(\boldsymbol{\theta})\right)^\top, \tag{13}$$

where model parameters $\boldsymbol{\theta}$ are shared across datasets.

6 Experiments

We start from the experimental setup of Xian *et al.* [32] and use their ResNet-based [56] multi-scale architecture for single-image depth prediction. We initialize the encoder with pretrained ImageNet [57] weights and initialize other layers randomly. We use Adam [58] with a learning rate of 10^{-4} for randomly initialized layers and 10^{-5} for pretrained layers, and set the exponential decay rate to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Images are flipped horizontally with a 50 percent chance, and randomly cropped and resized to 384×384 to augment the data and maintain the aspect ratio across different input images. No other augmentations are used.

Subsequently, we perform ablation studies on the loss function and, since we conjecture that pretraining on ImageNet data has significant influence on performance, also the encoder architecture. We use the best-performing pretrained model as the starting point for our dataset mixing experiments. We use 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; we thus denote an epoch as processing $72,000$ images, roughly the size of MD and MV, and train for 60 epochs. We shift and scale the ground-truth disparity to the range [0,1] for all datasets.

Test Datasets and Metrics. For ablation studies of loss and encoders, we use our held-out validation sets of RW (360 images), MD (2,963 images – official validation set), and MV (3,058 images – see Table 2). For all training dataset mixing experiments and comparisons to the state of the art, we test on a collection of datasets that were never seen during training: DIW, ETH3D, Sintel, KITTI, NYU, and TUM. For DIW [34] we created a validation set of 10,000 images from the DIW training set for our ablation studies and used the official test set of 74,441 images when comparing to the state of the art. For NYU we used the official test split (654 images). For KITTI we used the intersection of the official validation set for depth estimation (with improved ground-truth depth [59]) and the Eigen test split [60] (161 images). For ETH3D and Sintel we used the whole dataset for which ground truth is available (454 and 1,064 images, respectively). For the TUM dataset, we use the *dynamic* subset that features humans in indoor environments [38] (1,815 images).

For each dataset, we use a single metric that fits the ground truth in that dataset. For DIW we use the Weighted Human Disagreement Rate (WHDR) [34]. For datasets that are based on relative depth, we measure the root mean squared error in disparity space (MV, RW, MD). For datasets that provide accurate absolute depth (ETH3D, Sintel), 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 (AbsRel). Finally, we use the percentage of pixels with $\delta = \max(\frac{z_i}{z_i^*}, \frac{z_i^*}{z_i}) > 1.25$ to evaluate models on KITTI, NYU, and TUM [15]. Following [10], we cap predictions at an appropriate maximum value for datasets that are evaluated in depth space. For ETH3D, KITTI, NYU, and TUM, the depth cap was set to the maximum ground-truth depth value (72, 80, 10, and 10 meters, respectively). For Sintel, we evaluate on areas with ground-truth depth below 72 meters and accordingly use a depth cap of 72 meters. For all our models and baselines, we align predictions and ground truth in scale and shift for each image before measuring errors. We perform the alignment in inverse-depth space based on the least-squares criterion. Since absolute numbers quickly become hard to interpret when evaluating on multiple datasets, we also present the relative change in performance compared to an appropriate baseline method.

Input Resolution for Evaluation. We resize test images so that the larger axis equals 384 pixels while the smaller axis is resized to a multiple of 32 pixels (a constraint imposed by the encoder), while keeping an aspect ratio as close as possible to the original aspect ratio. Due to the wide aspect ratio in KITTI this strategy would lead to very small input images. We thus resize the *smaller* axis to be equal to 384 pixels on this dataset and adopt the same strategy otherwise to maintain the aspect ratio. Most state-of-the-art methods that we compare to are specialized to a specific dataset (with fixed image dimensions) and thus did not specify how to handle different image sizes and aspect ratios during inference. We tried to find the best-performing setting for all methods, following their evaluation scripts and training dimensions. For approaches trained on square patches [32], we follow our setup and set the larger axis to the training image axis length and adapt the smaller one, keeping the aspect ratio as close as possible to the original. For approaches with non-square patches [11], [33], [34], [38] we fix the smaller axis to the smaller training image axis dimension. For DORN [19] we followed their tiling protocol, resizing the images to the dimensions stated for their NYU and KITTI evaluation, respectively. For Monodepth2 [24] and Struct2Depth [27], which were both trained on KITTI and thus expect a very wide aspect ratio, we pad the input image using reflection padding to obtain the same aspect ratio, resize to their specific input dimension, and crop the resulting prediction to the original target dimensions. For methods where model weights were available for different training resolutions we evaluated all of them and report numbers for the best-performing variant.

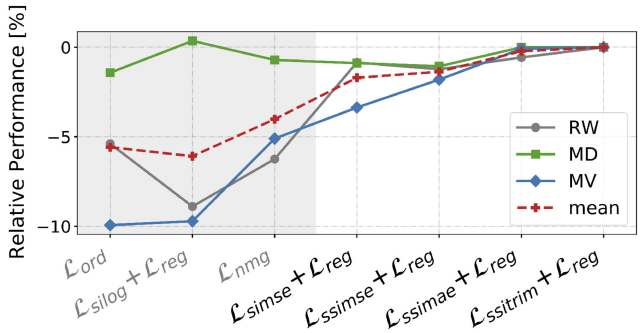

Fig. 3. Relative performance of different loss functions (higher is better) with the best performing loss $\mathcal{L}_{ssitrim}+\mathcal{L}_{reg}$ used as reference. All our four proposed losses (white area) outperform current state-of-the-art losses (gray area).

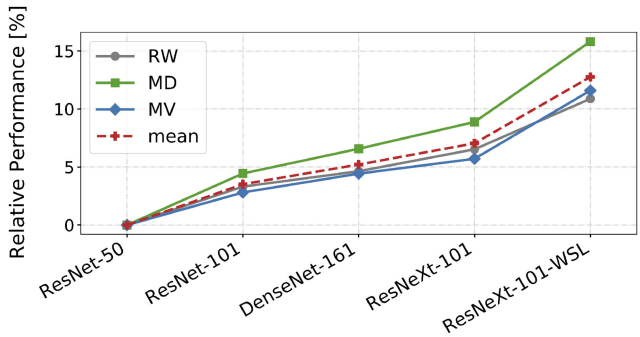

Fig. 4. Relative performance of different encoders across datasets (higher is better). ImageNet performance of an encoder is predictive of its performance in monocular depth estimation.

All predictions were rescaled to the resolution of the ground truth for evaluation.

Comparison of Loss Functions. We show the effect of different loss functions on the validation performance in Fig. 3. We used RW to train networks with different losses. For the ordinal loss (*cf.* Equation (9)), we sample 5,000 point pairs randomly [32]. Where appropriate, we combine losses with the gradient regularization term (11). We also test a scale-invariant, but not shift-invariant, MSE in disparity space $\mathcal{L}_{simse}$ by fixing $t=0$ in (1). The model trained with $\mathcal{L}_{ord}$ corresponds to our reimplementation of Xian *et al.* [32]. Fig. 3 shows that our proposed trimmed MAE loss yields the lowest validation error over all datasets. We thus conduct all experiments that follow using $\mathcal{L}_{ssitrim}+\mathcal{L}_{reg}$.

Comparison of Encoders. We evaluate the influence of the encoder architecture in Fig. 4. We define the model with a ResNet-50 [56] encoder as used originally by Xian *et al.* [32] as our baseline and show the relative improvement in performance when swapping in different encoders (higher is better). We tested ResNet-101, ResNeXt-101 [61] and DenseNet-161 [62]. All encoders were pretrained on ImageNet [57]. For ResNeXt-101, we additionally use a variant that was pretrained with a massive corpus of weakly-supervised data (WSL) [63] before training on ImageNet. All models were fine-tuned on RW.

We observe that a significant performance boost is achieved by using better encoders. Higher-capacity encoders perform better than the baseline. The ResNeXt-101 encoder that was pretrained on weakly-supervised data performs significantly better than the same encoder that was only trained on ImageNet. We found pretraining to be crucial. A network with a ResNet-50 encoder with random initialization performs on average 35 percent worse than its pretrained counterpart. In general, we find that ImageNet performance of an encoder is a strong predictor for its performance in monocular depth estimation. This is encouraging, since advancements made in image classification can directly yield gains in robust monocular depth estimation. The performance gain over the baseline is remarkable: up to 15 percent relative improvement, without any task-specific adaptations. We use ResNeXt-101-WSL for all subsequent experiments.

Training on Diverse Datasets. We evaluate the usefulness of different training datasets for generalization in Tables 3 and 4. While more specialized datasets reach better performance on similar test sets (DL for indoor scenes or MD for ETH3D), performance on the remaining datasets declines. Interestingly, every single dataset used in isolation leads to worse generalization performance on average than just using the small, but curated, RW dataset, *i.e.* the gains on compatible datasets are offset on average by the decrease on the other datasets.

TABLE 3
Relative Performance With Respect to the Baseline in Percent When Fine-Tuning on Different Single Training Sets (Higher is Better)

	DIW	ETH3D	Sintel	KITTI	NYU	TUM	Mean [%]
RW → RW	**14.6**	0.2	**0.3**	28.0	18.7	21.7	—
RW → DL	−37.6	2.0	−4.3	−73.0	**32.3**	**19.4**	−10.2
RW → MV	−26.1	−15.9	−15.5	**10.1**	−10.2	−3.5	−10.2
RW → MD	−31.5	**4.0**	−9.7	−24.3	−1.7	−52.0	−19.2
RW → WS	−32.4	−29.8	−2.9	−34.5	−31.9	3.2	−21.4

Performance better than the baseline in green, worse performance in red. Best performance is bold, second best is underlined. The absolute errors of the RW baseline are shown on the top row. While some datasets provide better performance on individual, similar datasets, average performance for zero-shot cross-dataset transfer degrades.

TABLE 4
Absolute Performance When Fine-Tuning on Different Single Training Sets – Lower is Better

	DIW	ETH3D	Sintel	KITTI	NYU	TUM
	WHDR	AbsRel	AbsRel	$\delta > 1.25$	$\delta > 1.25$	$\delta > 1.25$
RW → RW	**14.59**	0.151	**0.349**	27.95	18.74	21.69
RW → DL	20.08	0.148	0.364	48.35	**12.68**	**17.48**
RW → MV	18.39	0.175	0.403	**25.12**	20.65	22.44
RW → MD	19.18	**0.145**	0.383	34.73	19.05	32.96
RW → WS	19.31	0.196	0.359	37.59	24.72	20.99

This table corresponds to Table 3.

TABLE 5
Combinations of Datasets Used for Training

Mix	RW	DL	MV	MD	WS
MIX 1	✓	✓			
MIX 2	✓	✓	✓		
MIX 3	✓	✓	✓	✓	
MIX 4	✓	✓		✓	✓
MIX 5	✓	✓	✓	✓	✓

TABLE 6
Relative Performance of Naive Dataset Mixing With Respect to the RW Baseline (Top Row) – Higher is Better

	DIW	ETH3D	Sintel	KITTI	NYU	TUM	Mean [%]
RW	14.6	0.2	0.3	28.0	18.7	21.7	—
MIX 1	10.9	9.9	−3.7	**18.0**	41.4	33.0	18.3
MIX 2	6.7	8.6	3.2	9.2	40.8	35.7	17.3
MIX 3	**13.5**	10.6	4.9	13.9	**43.8**	29.1	19.3
MIX 4	11.7	11.3	5.2	11.3	38.8	35.5	19.0
MIX 5	12.3	**12.6**	**7.2**	9.1	38.5	**37.2**	**19.5**

While we usually see an improvement when adding datasets, adding datasets can hurt generalization performance with naive mixing.

TABLE 8
Relative Performance of Dataset Mixing With Multi-Objective Optimization With Respect to the RW Baseline (Top Row) – Higher is Better

	DIW	ETH3D	Sintel	KITTI	NYU	TUM	Mean [%]
RW	14.6	0.2	0.3	28.0	18.7	21.7	—
MIX 1	9.4	7.3	−7.7	13.2	44.1	33.2	16.6
MIX 2	14.1	8.6	0.9	**17.5**	45.5	32.0	19.8
MIX 3	15.8	11.9	5.2	11.7	47.8	32.4	20.8
MIX 4	15.4	13.9	1.7	17.2	43.4	**38.2**	21.6
MIX 5	**15.9**	**14.6**	**6.3**	14.5	**49.0**	34.1	**22.4**

Principled mixing dominates the solutions found by naive mixing.

TABLE 7
Absolute Performance of Naive Dataset Mixing – Lower is Better

	DIW	ETH3D	Sintel	KITTI	NYU	TUM
	WHDR	AbsRel	AbsRel	$\delta > 1.25$	$\delta > 1.25$	$\delta > 1.25$
RW	14.59	0.151	0.349	27.95	18.74	21.69
MIX 1	13.00	0.136	0.362	**22.91**	10.98	14.53
MIX 2	13.62	0.138	0.338	25.39	11.10	13.94
MIX 3	**12.62**	0.135	0.332	24.06	**10.54**	15.39
MIX 4	12.88	0.134	0.331	24.78	11.46	14.00
MIX 5	12.79	**0.132**	**0.324**	25.41	11.52	**13.62**

This table corresponds to Table 6.

TABLE 9
Absolute Performance of Dataset Mixing With Multi-Objective Optimization – Lower is Better

	DIW	ETH3D	Sintel	KITTI	NYU	TUM
	WHDR	AbsRel	AbsRel	$\delta > 1.25$	$\delta > 1.25$	$\delta > 1.25$
RW	14.59	0.151	0.349	27.95	18.74	21.69
MIX 1	13.22	0.140	0.376	24.26	10.48	14.50
MIX 2	12.54	0.138	0.346	**23.05**	10.21	14.76
MIX 3	12.29	0.133	0.331	24.68	9.78	14.66
MIX 4	12.35	0.130	0.343	23.13	10.61	**13.41**
MIX 5	**12.27**	**0.129**	**0.327**	23.90	**9.55**	14.29

This table corresponds to Table 8.

The difference in performance for RW, MV, and WS is especially interesting since they have similar characteristics. Although substantially larger than RW, both MV and WS show worse individual performance. This could be explained partly by redundant data due to the video nature of these datasets and possibly more rigorous filtering in RW (human experts pruned samples that had obvious flaws). Comparing WS and MV, we see that MV leads to more general models, likely because of higher-quality stereo pairs due to the more controlled nature of the images.

For our subsequent mixing experiments, we use Table 3 as reference, *i.e.* we start with the best performing individual training dataset and consecutively add datasets to the mix. We show which datasets are included in the individual training sets in Table 5. To better understand the influence of the Movies dataset, we additionally show results where we train on all datasets except Movies (MIX 4). We always start training from the pretrained RW baseline.

Tables 6 and 7 show that, in contrast to using individual datasets, mixing multiple training sets consistently improves performance with respect to the baseline. However, we also see that adding datasets does not unconditionally improve performance when naive mixing is used (see MIX 1 versus MIX 2). Tables 8 and 9 report the results of an analogous experiment with Pareto-optimal dataset mixing. We observe that this approach improves over the naive mixing strategy. It is also more consistently able to leverage additional datasets. Combining all five datasets with Pareto-optimal mixing yields our best-performing model. We show a qualitative comparison of the resulting models in Fig. 5.

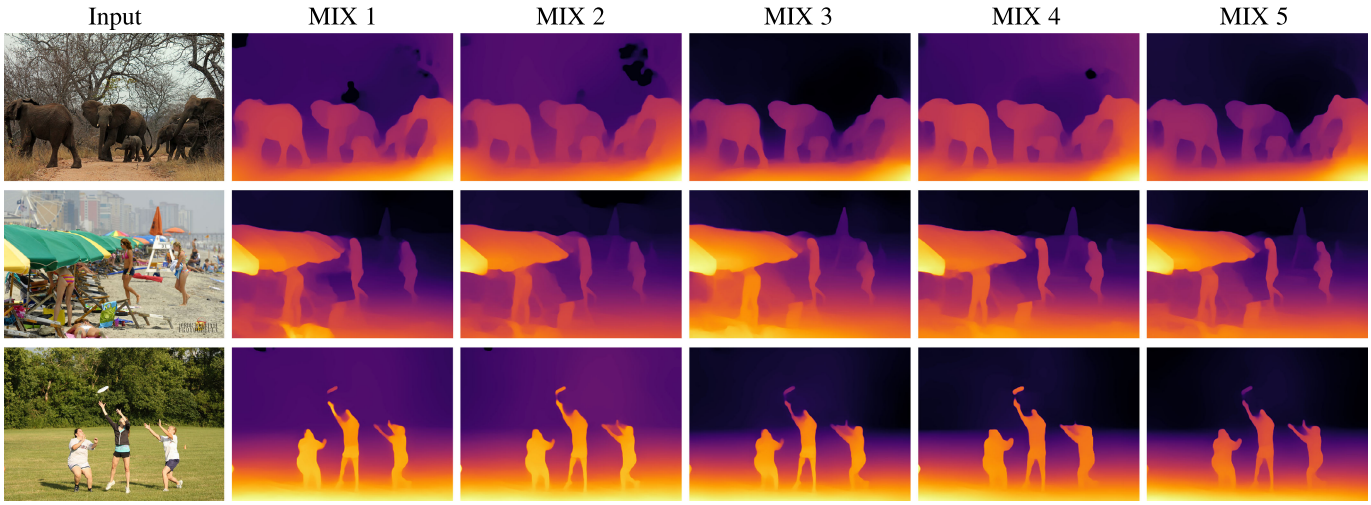

Fig. 5. Comparison of models trained on different combinations of datasets using pareto-optimal mixing. Images from Microsoft COCO [5].

TABLE 10
Relative Performance of State of the Art Methods With Respect to Our Best Model (Top Row) – Higher is Better

	Training sets	DIW	ETH3D	Sintel	KITTI	NYU	TUM	Mean [%]
Ours	MIX 5	**12.46**	**0.129**	**0.327**	23.90	9.55	**14.29**	—
Ours – small	MIX 5	-0.2	-20.2	-0.9	8.7	-64.7	-19.0	−16.0
Xian [32]	RW	-17.1	-44.2	-29.1	-42.6	-182.7	-75.1	−65.1
Li [38]	MC	-112.8	-41.9	-23.9	-100.6	-94.5	-23.9	−66.2
Wang [33]	WS	-53.2	-58.9	-19.3	-33.6	-209.6	-41.2	−69.3
Li [11]	MD	-85.8	-41.1	-17.7	-51.8	-188.2	-106.7	−81.9
Casser [27]	CS	-163.2	-82.2	-29.1	11.5	-314.5	-160.2	−122.9
Fu [19]	NYU	-131.1	-51.2	-32.4	-157.8	**9.0**	-72.5	−72.6
Chen [34]	NYU→DIW	-16.1	-71.3	-34.6	-51.9	-196.6	-111.1	−80.3
Godard [24]	KITTI	-138.1	-46.5	-24.2	**76.9**	-248.6	-152.1	−88.8
Casser [27]	KITTI	-168.8	-68.2	-25.1	50.1	-277.8	-159.1	−108.2
Fu [19]	KITTI	-143.9	-67.4	-32.1	70.2	-325.2	-180.8	−113.2

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.

TABLE 11
Absolute Performance of State of the Art Methods, Sorted by Average Rank

	Training sets	DIW WHDR	ETH3D AbsRel	Sintel AbsRel	KITTI δ>1.25	NYU δ>1.25	TUM δ>1.25	Rank
Ours	MIX 5	**12.46**	**0.129**	**0.327**	23.90	9.55	**14.29**	2.0
Ours – small	MIX 5	12.48	0.155	0.330	21.81	15.73	17.00	2.7
Li [11]	MD	23.15	0.181	0.385	36.29	27.52	29.54	5.7
Li [38]	MC	26.52	0.183	0.405	47.94	18.57	17.71	5.7
Wang [33]	WS	19.09	0.205	0.390	31.92	29.57	20.18	6.0
Xian [32]	RW	14.59	0.186	0.422	34.08	27.00	25.02	6.1
Casser [27]	CS	32.80	0.235	0.422	21.15	39.58	37.18	9.6
Godard [24]	KITTI	29.67	0.189	0.406	**5.53**	33.29	36.03	6.7
Fu [19]	NYU	28.79	0.195	0.433	61.61	**8.69**	24.65	7.3
Chen [34]	NYU → DIW	14.47	0.221	0.440	36.30	28.33	30.16	8.5
Casser [27]	KITTI	33.49	0.217	0.409	11.93	36.08	37.03	8.7
Fu [19]	KITTI	30.39	0.216	0.432	7.13	40.61	40.13	9.2

This table corresponds to Table 10.

Comparison to the State of the Art. We compare our best-performing model to various state-of-the-art approaches in Tables 10 and 11. The top part of each 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 parts show baselines that were fine-tuned on a subset of the datasets for reference. In the training set column, MC refers to Mannequin Challenge [38] and CS to

Fig. 6. Qualitative comparison of our approach to the four best competitors on images from the Microsoft COCO dataset [5].

Fig. 7. Qualitative results on the DIW test set.

Fig. 8. Results on paintings and drawings. Top row: *A Friend in Need*, Cassius Marcellus Coolidge, and *Bathers at Asniéres*, Georges Pierre Seurat. Bottom row: *Mittagsrast*, Vincent van Gogh, and *Vector drawing of central street of old european town, Vilnius,* @Misha

Cityscapes [45]. A → B indicates pretraining on A and fine-tuning on B.

Our model outperforms the baselines by a comfortable margin in terms of zero-shot performance. Note that our model outperforms the Mannequin Challenge model of Li *et al.* [38] on a subset of the TUM dataset that was specifically curated by Li *et al.* to showcase the advantages of their model. We show additional results on a variant of our model that has a smaller encoder based on ResNet-50 (Ours – small). This architecture is equivalent to the network proposed by Xian *et al.* [32]. The smaller model also outperforms the state of the art by a comfortable margin. This shows that the strong performance of our model is not only due to increased network capacity, but fundamentally due to the proposed training scheme.

Some models that were trained for one specific dataset (*e.g.* KITTI or NYU in the lower part of the table) perform very well on those individual datasets but perform significantly worse on all other test sets. Fine-tuning on individual datasets leads to strong priors about specific environments. This can be desirable in some applications, but is ill-suited if the model needs to generalize. A qualitative comparison of our model to the four best-performing competitors is shown in Fig. 6.

Additional Qualitative Results. Fig. 7 shows additional qualitative results on the DIW test set [34]. We show results on a diverse set of input images depicting various objects and scenes, including humans, mammals, birds, cars, 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.

We show qualitative results on the DAVIS video dataset [64] in our supplementary video, `https://youtu.be/D46FzVyL9I8`. 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.

Hertzmann [65] recently observed that our publicly available model provides plausible results even on abstract line drawings. Similarly, we show results on drawings and paintings with different levels of abstraction in Fig. 8. We can qualitatively confirm the findings in [65]: The model shows a surprising capability to estimate plausible relative depth even on relatively abstract inputs. This seems to be true as long as some (coarse) depth cues such as shading or vanishing points are present in the artwork.

Failure Cases. We identify common failure cases and biases of our model. 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, [34] achieved an agreement rate of 85.8 percent with human annotators. This bias has also been learned by our network and can be observed in some extreme cases that are shown in the first row of Fig. 9. In the example on the left, the model fails to recover the ground plane, likely because the input image was rotated by 90 degrees. In the right 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 desired property for this task.

Another interesting failure case is shown in the second row of Fig. 9. Paintings, photos, and mirrors are often not recognized as such. The network estimates depth based on the content that is depicted on the reflector rather than predicting the depth of the reflector itself.

Additional failure cases are shown in the remaining rows. Strong edges can lead to hallucinated depth discontinuities. Thin structures can be missed and relative depth arrangement between disconnected objects might fail in some situations. Results tend to get blurred in background regions, which might be explained by the limited resolution of the input images and imperfect ground truth in the far range.

7 Conclusion

The success of deep networks has been driven by massive datasets. For monocular depth estimation, we believe that

Fig. 9. Failure cases. Subtle failures in relative depth arrangement or missing details are highlighted in green.

existing datasets are still insufficient and likely constitute the limiting factor. 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.

We have evaluated the robustness and generality of models via zero-shot cross-dataset transfer. We find that systematically testing models on datasets that were never seen during training is a better proxy for their performance "in the wild" than testing on a held-out portion of even the most diverse datasets that are currently available.

Our work advances the state of the art in generic monocular depth estimation and indicates that the presented ideas substantially improve performance across diverse environments. We hope that this work will contribute to the deployment of monocular depth models that meet the requirements of practical applications. Our models are freely available at `https://github.com/intel-isl/MiDaS`.

Acknowledgments

René Ranftl and Katrin Lasinger contributed equally to this work.

References

[1] B. Zhou, P. Krähenbühl, and V. Koltun, "Does computer vision matter for action?," *Sci. Robot.*, vol. 4, no. 30, 2019. [Online]. Available: https://doi.org/10.1126/scirobotics.aaw6661

[2] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 577–584, 2005.

[3] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009.

[4] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," 2018, *arXiv: 1801.09847*.

[5] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp 740–755.

[6] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of Kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012.

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

[8] P. Fankhauser, M. Blösch, D. Rodriguez, R. Kaestner, M. Hutter, and R. Siegwart, "Kinect v2 for mobile robot navigation: Evaluation and modeling," in *Proc. Int. Conf. Adv. Robot.*, 2015, pp. 388–394.

[9] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 740–756.

[10] C. Godard, O. MacAodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 270–279.

[11] Z. Li and N. Snavely, "MegaDepth: Learning single-view depth prediction from Internet photos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2041–2050.

[12] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 525–536.

[13] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1521–1528.

[14] K. Karsch, C. Liu, and S. B. Kang, "Depth transfer: Depth extraction from video using non-parametric sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2144–2158, Nov. 2014.

[15] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 2366–2374.

[16] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 239–248.

[17] A. Roy and S. Todorovic, "Monocular depth estimation using neural regression forest," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5506–5514.

[18] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5162–5170.

[19] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2002–2011.

[20] R. Li, K. Xian, C. Shen, Z. Cao, H. Lu, and L. Hang, "Deep attention-based classification network for robust depth prediction," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 663–678.

[21] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 484–500.

[22] Y. Luo *et al.*, "Single view stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 155–163.

[23] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 340–349.

[24] C. Godard, O. MacAodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3828–3838.

[25] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6612–6619.

[26] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5667–5675.

[27] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," in *Proc. AAAI*, 2019, pp. 8001–8008.

[28] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[29] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3061–3070.

[30] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.

[31] Y. Kim, H. Jung, D. Min, and K. Sohn, "Deep monocular depth estimation via integration of global and local predictions," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4131–4144, Aug. 2018.

[32] K. Xian *et al.*, "Monocular relative depth perception with web stereo data supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 311–320.

[33] C. Wang, O. Wang, F. Perazzi, and S. Lucey, "Web stereo video supervision for depth prediction from dynamic scenes," in *Proc. Int. Conf. 3D Vis.*, 2019, pp. 348–357.

[34] W. Chen, Z. Fu, D. Yang, and J. Deng, "Single-image depth perception in the wild," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 730–738.

[35] T. Schöps *et al.*, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2538–2547.

[36] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 611–625.

[37] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 573–580.

[38] Z. Li *et al.*, "Learning the depths of moving people by watching frozen people," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4516–4525.

[39] W. Chen, S. Qian, and J. Deng, "Learning single-image depth from videos using quality assessment networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5597–5606.

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

[41] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8976–8985.

[42] B. Ummenhofer *et al.*, "DeMoN: Depth and motion network for learning monocular stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5622–5631.

[43] J. M. Facil, B. Ummenhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera, "CAM-Convs: Camera-aware multi-scale convolutions for single-view depth," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11818–11827.

[44] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 567–576.

[45] M. Cordts *et al.*, "The Cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[46] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2432–2443.

[47] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, 2017, Art. no. 78.

[48] I. Vasiljevic *et al.*, "DIODE: A dense indoor and outdoor DEpth dataset," 2019, *arXiv: 1908.00463*.

[49] S. Hadfield, K. Lebeda, and R. Bowden, "Hollywood 3D: What are the best 3D features for action recognition?," *Int. J. Comput. Vis.*, vol. 121, no. 1, pp. 95–110, 2017.

[50] J. Xie, R. B. Girshick, and A. Farhadi, "Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016.

[51] F. Devernay and P. A. Beardsley, "Stereoscopic cinema," in, *Image and Geometry Processing for 3-D Cinematography*. Berlin, Germany: Springer, 2010.

[52] R. Neuman, "Bolt 3D: A casestudy," in, *Stereoscopic Displays and Applications XX*. Bellingham, Washington USA: SPIE, vol. 7237, 2009.

[53] FFmpeg developers, "FFmpeg," 2018, [Online]. Available: https://ffmpeg.org

[54] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8934–8943.

[55] S. Rota Bulò, L. Porzi, and P. Kontschieder, "In-place activated batchnorm for memory-optimized training of DNNs," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5639–5647.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[57] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[58] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Proc. Int. Conf. Learn. Representations*, 2015.
[59] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant CNNs," in *Proc. Int. Conf. 3D Vis.*, 2017, pp. 11–20.
[60] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2650–2658.
[61] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5987–5995.
[62] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.
[63] D. Mahajan *et al.* "Exploring the limits of weakly supervised pretraining," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp 185–201.
[64] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 724–732.
[65] A. Hertzmann, "Why do line drawings work? A realism hypothesis," *Perception*, vol. 49, pp. 439–451, 2020.

René Ranftl received the MSc and Ph.D. degrees from the Graz University of Technology, Austria, in 2010 and 2015, respectively. He is currently a senior research scientist at the Intelligent Systems Lab at Intel, in Munich, Germany. His research interests include topics in computer vision, machine learning, and robotics.

Katrin Lasinger received the master's degree in computer science from TU Wien, in 2015. She is currently working toward the PhD degree in computer vision at the group of Photogrammetry and Remote Sensing at ETH Zurich. Her research interests include 3D computer vision, including volumetric fluid flow estimation and dense depth estimation from single or multiple views.

David Hafner received the master's and the PhD degree from Saarland University, Germany, in 2012 and 2018, respectively. Since 2019, he has been a research engineer at the Intelligent Systems Lab at Intel, in Munich, Germany.

Konrad Schindler (Senior Member, IEEE) received the Diplomingenieur MTech degree from the Vienna University of Technology, Vienna, Austria, in 1999, and the PhD degree from the Graz University of Technology, Graz, Austria, in 2003. He was a photogrammetric engineer in the private industry and held researcher positions at the Graz University of Technology, Monash University, Melbourne, VIC, Australia, and ETH Zuürich, Zürich, Switzerland. He was an assistant professor of Image Understanding with TU Darmstadt, Darmstadt, Germany, in 2009. Since 2010, he has been a tenured professor of photogrammetry and remote sensing with ETH Zürich. His research interests include computer vision, photogrammetry, and remote sensing.

Vladlen Koltun is currently the chief scientist for Intelligent Systems at Intel. He directs the Intelligent Systems Lab, which conducts high-impact basic research in computer vision, machine learning, robotics, and related areas. He has mentored more than 50 PhD students, postdocs, research scientists, and PhD student interns, many of whom are now successful research leaders.