

# Web Stereo Video Supervision for Depth Prediction from Dynamic Scenes

Chaoyang Wang Simon Lucey  
Carnegie Mellon University  
{chaoyanw, slucey}@cs.cmu.edu

Federico Perazzi Oliver Wang  
Adobe Inc.  
{perazzi, owang}@adobe.com

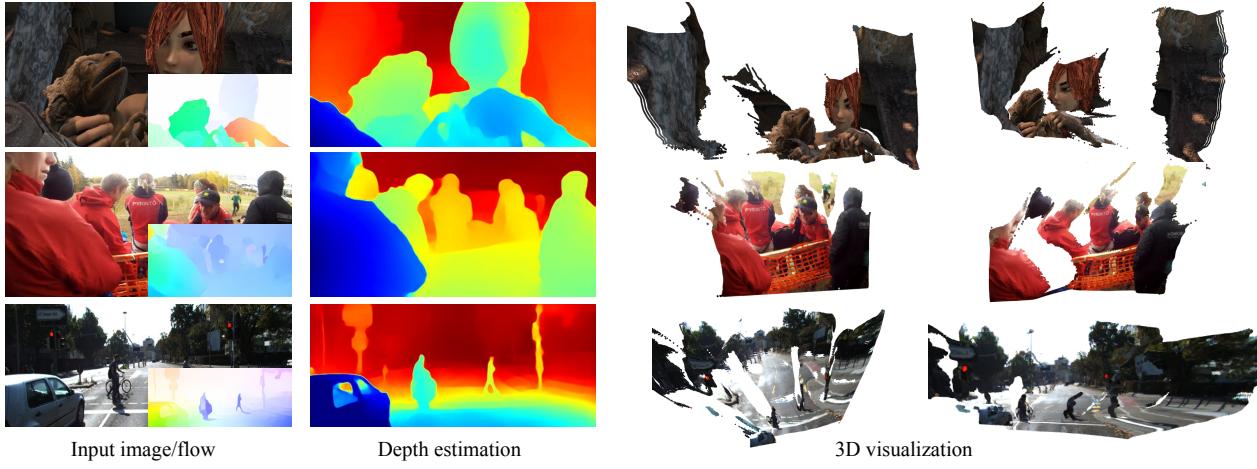


Figure 1: Depth prediction for nonrigid scenes from our multi-view depth estimator, which is trained on a new large scale database of web stereo videos.

## Abstract

We present a fully data-driven method to compute depth from diverse monocular video sequences that contain large amounts of non-rigid objects, e.g., people. In order to learn reconstruction cues for non-rigid scenes, we introduce a new dataset consisting of stereo videos scraped in-the-wild. This dataset has a wide variety of scene types, and features large amounts of nonrigid objects, especially people. From this, we compute disparity maps to be used as supervision to train our approach. We propose a loss function that allows us to generate a depth prediction even with unknown camera intrinsics and stereo baselines in the dataset. We validate the use of large amounts of Internet video by evaluating our method on existing video datasets with depth supervision, including Sintel, and KITTI, and show that our approach generalizes better to natural scenes.

## 1. Introduction

Recovering depth maps of non-rigid scenes from monocular video sequences is a challenging problem in 3D vision. While rigid scene reconstruction methods can use

geometric consistency assumptions across multiple frames, the problem becomes severely under-constrained for non-rigid scenes. As a result, nonrigid reconstruction methods must rely more heavily on strong scene priors. In fact, we see that data-driven single image reconstruction methods, which can learn *only* scene priors, sometimes outperform multi-frame geometric methods on nonrigid scenes (Sec. 5). In this work, we attempt the model nonrigid scene priors in a *data-driven* manner, by training on real-life sequences, while also learning to take advantage of geometric cues between neighboring frames of video.

Recent advances in data-driven methods have shown some advantages over traditional 3D reconstruction pipelines. However, due to restrictions on architectures and available training data, such approaches have largely been used to predict depth from single images [6, 21, 34, 4, 19] or multiple views of rigid scenes [32, 31], or have been trained on narrow domains, e.g., driving data [8].

To overcome these limitations, we introduce a new large-scale (1.5M frame) dataset collected in-the-wild from internet stereo videos, which contains large amounts of non-rigid objects and diverse scene types. We use derived stereo disparity for training, and at test time, predict depth from a pair of single-view, sequential frames. Our approach is designed

so that it learns to utilize both semantic (single-image) and geometric (multi-frame) cues to address the nonrigid reconstruction problem.

One challenge of using Internet stereo videos as a training source is that they contain unknown (and possibly temporally varying) camera intrinsics and extrinsics, and consequently, we cannot directly translate disparities to depth values (even up to scale). This prevents the use of existing regression loss including the scale invariant logarithmic depth/gradient loss [6]. Instead, we observe that in a stereo camera configuration, the *difference* in disparity between two pixels is proportional to the difference in their inverse depth. This motivates a new normalized multiscale gradient loss that allows for supervision of depth prediction networks directly from estimated disparities from uncalibrated stereo video data. Compared to the ordinal loss used in [4, 34], we show that our proposed loss has the advantage of retaining distance information between points, and yields smoother, more accurate depth maps.

While the computed disparity maps contain errors, we show that training a deep network on our dataset with the proposed loss generalizes well to other datasets, and that by using temporal information in the form of flow maps and sequential frames as input into the network, we are able to improve over single image depth prediction methods. Our code and data will be made available to the research community.

In summary, we present the following contributions:

1. A simple architecture that leverages stereo video for training, and produces depth maps from monocular sequential frames with nonrigid objects at test time. To our knowledge this is the first multi-frame data-driven approach that outperforms single-image reconstruction on nonrigid scenes.
2. A new stereo video dataset that features a wide variety of real world examples with non-rigid objects (e.g., humans) in an unconstrained setting.
3. A novel loss function that allows us to train with unknown camera parameters, while outperforming previously used ordinal losses.

## 2. Related Work

**Traditional geometric approaches** for predicting depth from video sequences rely on SLAM [25] or SfM [29] pipelines for camera pose and 3D map estimation, followed by dense multiview stereo [31] to get per-pixel depth values. These methods are mature systems that produce highly accurate reconstructions in the right circumstances. However, they rely on hand-designed features, assumptions such as brightness constancy across frames, and often require a good map initialization and sufficient parallax to converge. Additionally, while camera pose estimation can handle some amount of nonrigid motion by simply ignoring

those regions, dense multiview stereo requires a rigid scene.

Non-rigid SfM methods try to recover 3D by replacing the rigid assumption with additional constraints. Bregler et al. [3] introduces a fixed rank constraint on the shape matrix for non-rigid structures. Numerous innovations have followed, introducing additional priors to make the problem less ambiguous and to scale to dense reconstruction[14, 5, 20, 41, 2, 16]. These approaches usually assume weak perspective cameras and rely on hand-designed features for input feature tracks.

Other recent works[17, 27] explore dense reconstructions from two perspective frames, using an as-rigid-as-possible (ARAP) assumption to address the scale ambiguity issue inherent in non-rigid reconstruction. Although promising results have been shown on mostly rigid scenes, their ARAP assumption is not enough to handle complicated dynamic scenes. Our approach learns priors from data, and we show that it can often produce good result for highly dynamic scenes.

**Data driven approaches** that leverage deep networks for scene priors and feature learning, have become a potential way to overcome the limitations of traditional methods. Recent works [32, 12, 38] demonstrate promising result for rigid scene reconstruction. However, since they include explicit rigid transformation inside their network architecture, they cannot handle non-rigid motion. We propose a new data-driven method that focuses on non-rigid objects and diverse scenes, and introduce a new internet stereo video dataset to train this approach.

**Supervision using depth sensors** is a common strategy for existing methods that directly regress depth values. These approaches rely on datasets collected for example, by laser scanner [28], Kinect depth camera [6], car-mounted lidar sensor [6, 18], synthetic rendered data (for pretraining) [10], or dual-camera iPhone [22]. One challenge of requiring specialized hardware is that it is hard to acquire sufficiently diverse data for training, and as such they tend to be used only in constrained domains, e.g., driving sequences.

Recent approaches have proposed using more common stereo camera rigs for training depth prediction networks [9, 36, 37]. These approaches are trained using stereo video data on KITTI [8], by treating depth prediction as an image reconstruction problem. Our approach differs to theirs in that we are learning from stereoscopic videos with unknown camera parameters, and we also utilize temporal information at test time. Another method, Deep3D [35] uses 3D movies for training, however this approach focuses on synthesizing novel stereoscopic views rather than scene reconstruction.

**Supervision using internet images** is a powerful tool that allows for the collection of diverse datasets for learning depth reconstruction priors. MegaDepth [21] generates a set of 3D reconstructions using traditional SfM and multi-

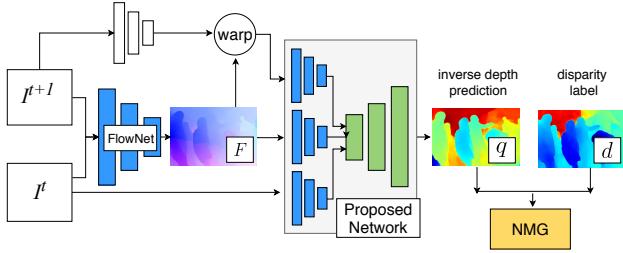


Figure 2: Our network takes input from: the frame  $I^t$ , the second frame  $I^{t+1}$  and the flow between  $I^t, I^{t+1}$ . We extract feature pyramids for both input images and the flow map. Moreover, we warp the feature pyramid for  $I^{t+1}$  with the flow. The warped feature pyramid of  $I^{t+1}$  is then fused with the feature pyramids for  $I^t$  and the flow map, fed into a decoder with skip connections and produces a depth map. This is supervised with the disparity directly using our Normalized Multi-Scale Gradient (NMG) Loss.

view stereo reconstruction pipelines from internet images. These reconstructions serve as ground truth normalized depth maps for supervision. Another recent work scrapes stereo *images* from the web, and computes disparity from these [34]. This approach uses ordinal depth constraints from these disparity estimates to train a single image depth prediction network.

In contrast, we are interested in data that allows us to take advantage of multiple frames of video. We show that reconstruction quality is improved when such data is available, indicating that it is possible to learn both scene priors from single images, and geometric information from multiple frames. In addition, we compare our loss formulation to that presented in the above method, and show that it outperforms relative depth constraints.

**Supervision using single-view video data** has been a popular recent approach for scene reconstruction. In this case, supervision is derived from temporal frames of a video by predicting depth and camera pose, and computing a loss based on warping the one frame to the next based on these values. Then at test time, a single frame is used for depth prediction [7, 23, 42, 40, 39, 24, 33, 36]. As geometric projection is only valid for rigid scenes, these methods often include a rigidity mask, where the loss is treated differently outside these regions [39, 42], and introduce regularization to prevent the method from degenerating.

While these approaches are elegant as they do not require extra supervision other than a single-view video, in practice many of these methods require both known intrinsics, and assume mostly rigid scenes. In our work we are interested in highly non-rigid scenes, and therefore do not explicitly rely on a geometric reprojection loss. Instead we use temporal flow and flow-warped images as an input to our network to learn the geometric relationship between flow and depth.

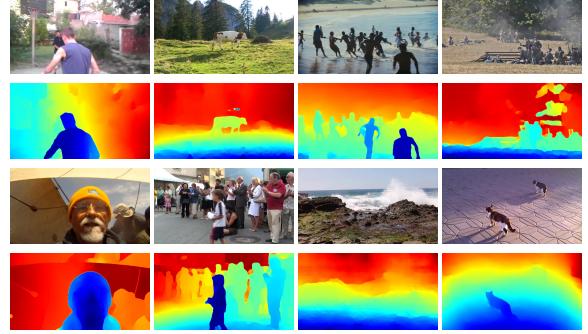


Figure 3: Random selection of frames from WSVD showing a sampling of the diversity of scenes and subjects contained. Below each image we show the computed disparity map used for supervision.

### 3. Dataset

We introduce the Web Stereo Video Dataset (WSVD), a large, in-the-wild collection of stereo videos featuring many non-rigid objects (especially people). Figure 3 shows a selection of representative frames from WSVD.

To collect WSVD, we scraped YouTube for videos that were tagged as being viewable in stereo 3D. 3D videos on YouTube are stored internally in a left/right format, meaning that each video is squeezed in half width-wise and stored side by side in a single video file. We additionally searched for videos on Vimeo with a “side-by-side 3D” query to match this format (Vimeo does not naively support 3D). In theory all videos with the proper tag should be 3D, but in practice, a large number of these videos were either regular monocular video, or different kinds of stereo, e.g., anaglyph, top/bottom, etc. We therefore conduct two stages of filtering to make this data usable.

**Filtering** Our initial scraping yielded 7k+ videos from YouTube and Vimeo. We first identified all videos that were actual left-right stereo videos from this set. To do this, we computed the MSE of the left and right half of each video, split the videos into two classes based on this metric and then manually removed outliers.

After this step, all videos remaining were left/right stereo videos, but still not all of them were usable. For further filtering, we split up each video into shots using histogram differences [1], and performed a per-shot categorization of good and bad videos. To do this, we first calculated the average brightness of the middle frame to filter out shots with black screens. We also performed text detection to remove shots with large text titles.

We then automatically computed the disparity for the middle frames per shot using a flow approach [13]. We found that many samples displayed substantial lens distortion, near-zero baselines, vertical disparities, color differ-



Figure 4: WSVD word cloud of the two hundred most frequent classes estimated with an object detector trained on OpenImages [15] bounding boxes, showing a high quantity of non-rigid objects, especially people.

ences, inverted cameras, and other poor stereo characteristics. To reject those shots, we used the following criteria as an initial step for filtering: pixels with vertical disparity  $> 1$  pixel is greater than 10%; range of horizontal disparity is  $< 5$  pixels, and the percentage of pixels passing a left-right disparity consistency check is  $< 70\%$ . This was followed by a second curation step, to remove videos with obviously incorrect disparity maps (e.g., due to radial distortion, or incorrect camera configurations).

Next, we removed static frames by filtering frames with maximum flow magnitude less than 8 pixels. For each remaining frame, we calculated disparity map as our ground truth using FlowNet2.0 [13], which we found to produce the best results. The left-right disparity check is also used to mask out outliers, which are not used for supervision. Finally, we are left with 10788 clips from 689 videos, consisting of around 1.5 million frames.

**Analysis** In order to understand what type of content is present in the WSVD dataset, we run a two-stage image classifier similar to Mask R-CNN [11] on the middle frames of each shot. The classifier is pre-trained on the 600 classes annotated in the Open Images dataset [15]. We retained bounding boxes with confidence score  $\geq 0.7$ . The analysis of the results indicates that roughly 79% of the video sequences contain either humans, animals, or vehicles, which are likely to display non-rigid motion. In Figure 4, we show the word cloud generated from the class frequencies of the detected bounding boxes. In contrast to other datasets, i.e. KITTI and NYU, WSVD reflects the diverse content that people watch on YouTube, with many non-rigid objects, especially humans.

#### **4. Approach**

Our supervision comes in the form of a disparity map. Assuming that the stereoscopic videos are captured by two horizontally placed cameras, and the vertical disparity has already been removed, then disparity can be translated to

inverse depth  $q$  using the following equation:

$$q = \frac{d - (c_x^R - c_x^L)}{fb} \quad (1)$$

Where  $d$  is the disparity,  $b$  is the camera baseline,  $f$  is the focal length, and  $c_x^R, c_x^L$  denote the horizontal coordinates of the principal points for the left/right cameras.  $c_x^R - c_x^L$  defines the minimum possible disparity  $d_{\min}$  in the camera configuration. In practice,  $d_{\min}$  can have arbitrary values depending on the stereo rig configuration and post-production. For example, to release visual fatigue, most stereo rigs for 3D movies are towed-in to place the object of interest at the 0 disparity plane, which creates negative disparity values.

Unlike most other video datasets with depth supervision, this dataset has unknown and variable focal length  $f$ , and camera stereo configuration (baseline  $b$  and  $d_{\min}$ ). Among those,  $d_{\min}$  is the key parameter preventing us from converting the estimated disparity into an inverse depth map up to scale, as is commonly done in other self-supervised learning methods [9, 21]. This term also prevents us to apply the widely used scale-invariant logarithmic depth (gradient) loss [6], due to the fact that subtracting the mean/neighboring pixel’s logarithmic depth value is not enough to cancel out  $d_{\min}$ .

Although in theory, we could estimate  $d_{\min}$  with the disparity value of pixels at infinity distance, it would not be robust due to the fact that regions in distance does not always present in videos, and that the usual choice for such regions i.e. textureless sky, are likely to have incorrect disparity values.

Due to this issue, prior work [34] only uses an ordinal relation for supervision, and does not attempt to recover the relative distance from the disparity map. We take a different approach – the proposed loss function takes supervision from the whole disparity map, and preserves the continuous distance information between points in addition to ordinal relation. In comparison, ordinal loss [34, 4] only enforce binary ordinal relation between a sparse set of pixel pairs.

#### 4.1. Normalized multiscale gradient (NMG) loss

From Eq. 1 we derive that the difference in disparity between two pixels is proportional to the difference in their inverse depth:

$$q_i - q_j = \frac{d_i - d_j}{fb} \quad (2)$$

This allows us to design a novel loss invariant to the minimum possible disparity value. The idea is to enforce the gradient of inverse depth prediction to be close to the gradient of the disparity map up to scale (normalized). The gradient is evaluated at different spacing amounts (multiscale) to include both local/global information [32]. The loss can

be written as:

$$\mathcal{L} = \sum_k \sum_i |s\nabla_x^k q_i - \nabla_x^k d_i| + |s\nabla_y^k q_i - \nabla_y^k d_i|, \quad (3)$$

where  $\nabla_x^k$ ,  $\nabla_y^k$  denote the difference evaluated with spacing  $k$  (we use  $k = \{2, 8, 32, 64\}$ ); and the scale ratio  $s$  is estimated by:

$$s = \frac{\sum_k \sum_i |\nabla_x^k d_i| + \sum_k \sum_i |\nabla_y^k d_i|}{\sum_k \sum_i |\nabla_x^k q_i| + \sum_k \sum_i |\nabla_y^k q_i|}. \quad (4)$$

We choose to scale the inverse depth prediction to match disparity, and not the other way around, as this is more robust when the signal-to-noise ratio of the disparity map is low, which often happens when the range of disparity values is narrow, either due to a small baseline, or distant scenes. Scaling such noisy disparity with low contrast would amplify noise in the depth prediction.

Compared to the scale-invariant gradient loss of Eigen et al. [6], the proposed loss NMG is different in that it is defined in terms of disparity, not the log of depth, which is inapplicable in our scenario since we do not have depth as supervision.

Compared to the ordinal loss [34, 4], our NMG loss enforces relative distance and smoothness in addition to pairwise ordinal relation. As shown in Figure 5, the NMG loss yields more accurate global structure and preserves edge details better.

## 4.2. Depth prediction network

To utilize temporal information for depth prediction, we propose a network architecture inspired by recent work [34], modified to take input from two sequential frames and their optical flow. The general architecture is a multi-scale feature fusion [42, 34] net with three feature pyramids streams (features for 1st image, warped features for 2nd frame, and features for the flow map). These features are then projected and concatenated together and fed into an decoder with skip connections. Please refer to the supplementary material for a detailed description of the architecture.

**Feature pyramids** We use a ResNet-50 network to extract feature pyramids from the 1st and 2nd input frames. These feature pyramids have 4 levels with a coarsest scale of 1/32 of the original resolution. For flow input, we first use conv7×7 and conv5×5 layers with stride 2 to get features at 1/4 resolution. Then we apply 4 blocks of 2 conv3×3 layers to get pyramid flow features with 4 scales. We also apply residual blocks identical to [34] to project each image feature map to 256-channels, and each flow feature map to 128-channels. Finally, we warp the features in the pyramid for the 2nd frame to the 1st using the flow.

At each level of the feature pyramids, we concatenate the feature map of the 1st frame and the warped feature of the 2nd frame, and then use a residual block to project it to 256-channels, and concatenated it with the flow features. A final residual block projects this tensor to 256-channels.

**Depth decoder** The fused feature pyramid is fed into a depth decoder with skip connections. Starting from the coarsest scale, the output from previous scale of the decoder is bilinearly upsampled and then added to the corresponding fused feature from the pyramid. After reaching 1/4 of the full resolution, a stack of 2 3×3 conv layers and 1 bilinear upsampling layer is applied to produce final full-resolution depth prediction in log scale.

As shown in Figure 8 and our ablation study, we find that compared to using features only from single image, fusing features from the second frame and the flow helps identify foreground objects and produces more accurate result.

## 5. Evaluation

We evaluate our network on three video datasets with non-rigid scenes, and compare to other methods trained on a variety of training sets as well as a traditional geometric method (COLMAP [31]). We seek to answer the following questions:

- How does the proposed NMG loss compared to ordinal loss used by previous methods?
- How important is the temporal information used by our network to improve depth prediction?
- How does WSVD compare to other multi-view datasets when used as training source?
- How does our method generalize to other dataset compared to state-of-the-art methods?

**Experiment setup** We hold out 50 videos, and sample 597 frame pairs as our testing set. We use all the rest videos in our dataset for training and validation. The validation set consists of 1000 frames from 500 randomly sampled clips. To avoid bias in sampling from longer videos, we randomly sample 1 clip per video at each training epoch. We train our network using Adam with default parameters and use batch-size of 6. For our ablation study, we also train a single-view depth prediction network which has identical architecture of our proposed network, but without the feature pyramids from the flow and the second input frame.

**Metrics** For comparison on Sintel, we use the commonly used mean relative error (MRE) and scale invariant logarithmic error (SILog) evaluated on inverse depth. These two errors are measured for pixels up to 20 meters away. We also use the proposed NMG as an additional metric. To avoid bias towards testing samples with large inverse depth value, we scaled the NMG error by the mean of the ground truth inverse depth.

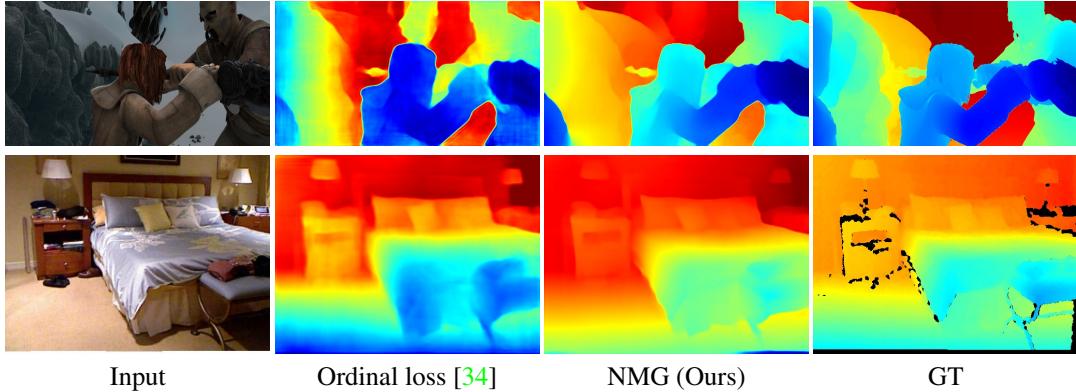


Figure 5: Comparison of NMG with an Ordinal loss. Top: trained on WSVD and tested on Sintel; Bottom: trained and tested on NYU-v2.

For KITTI, we use the 697 test samples (paired with consecutive frame for multi-view testing) from the Eigen split. We report absolute relative difference (abs rel), squared relative difference (sqr rel), relative mean squared logarithmic error (RMSE log) and percentage of error lower than a threshold (e.g.  $\delta < 1.25$ ). Since our method cannot predict metric depth, we align the prediction to the groundtruth by the median of the depth map, as is also done in [39].

In addition, we perform evaluation on our WSVD test set. We compute the NMG error excluding pixels which fail the left-right disparity consistency check (visualized as black region in Figure 5). While we do not claim that this is any indication of the generalization ability of our approach, we do so as an indication of how methods might perform on diverse non-rigid scenes.

### 5.1. Evaluation of NMG Loss

In order to evaluate the effect of our normalized multiscale gradient (NMG) loss, we compare it to the ordinal loss used in RedWeb [34]. To do this, we first compare both losses on the NYU-v2 dataset [26]. To mimic disparity maps from Internet stereoscopic images, we apply affine transformations to the ground truth depth values with uniformly sampled slope and bias parameters. These synthesized disparity maps are then used to train a single view depth estimator with different loss functions. As shown in Table 1, NMG loss has lower testing errors compared to ordinal loss, and its depth map prediction appears to be smoother as well (see Fig. 5).

Moreover, we train the proposed multi-view depth estimator on WSVD with both losses. We see in Table 2, and Figure 5, that our loss function again yields visually smoother, and more accurate depth maps when tested on Sintel.

	loss by Eigen [6]	ordinal	NMG
train label	depth	{synthesized disparity}	
rmse	0.467	0.767	<b>0.706</b>
abs rel	0.128	0.184	<b>0.164</b>
$\delta < 1.25$	0.840	0.753	<b>0.768</b>
$\delta < 1.25^2$	0.961	0.927	<b>0.945</b>
$\delta < 1.25^3$	0.990	0.979	<b>0.988</b>

Table 1: Comparison between NMG and ordinal loss on NYU-v2. Trained using synthesized disparity map with randomly sampled camera parameters. Result of directly using depth map as training label (loss by Eigen [6]) is also provided as reference.

	NMG	MRE	SILog
ordinal [34]	0.963	0.350	0.228
NMG (Ours)	<b>0.890</b>	<b>0.311</b>	<b>0.172</b>

Table 2: Comparison between NMG and ordinal loss. Trained on WSVD and tested on Sintel. Columns show different evaluation metrics.

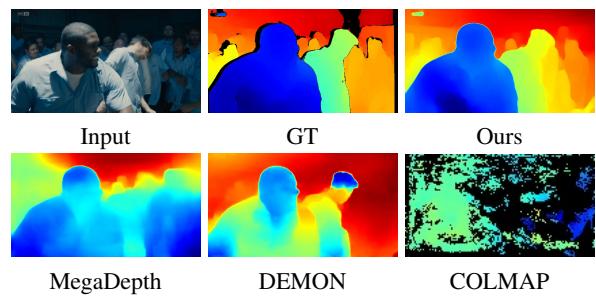


Figure 6: Qualitative comparison between our multi-view depth network and the state-of-the-art single/multi-view depth prediction methods.

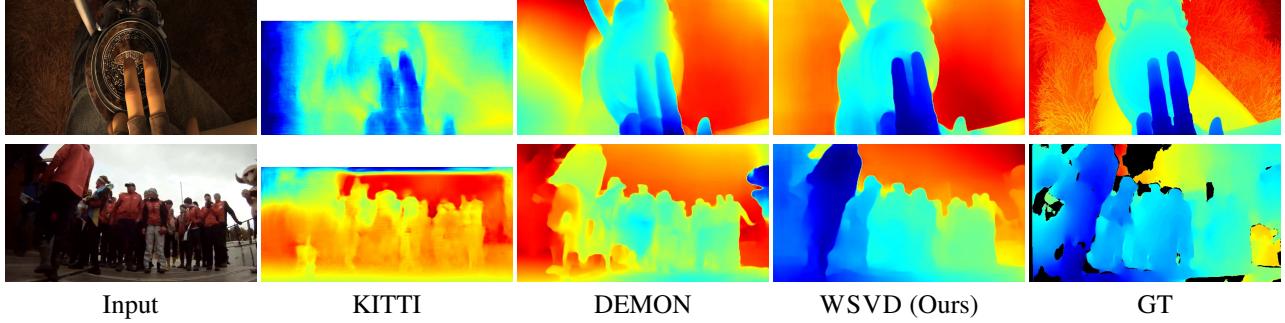


Figure 7: We show the performance of the same multi-view depth network trained on different datasets, and tested on Sintel and WSVD. Note that KITTI does not have groundtruth for the top region of the image, thus network trained with KITTI may produce arbitrary values in this area. For fairer comparison, we cropped out the top part.

Test set	Metric	single view				multi view		
		MegaDepth [21]	ReDWeb*	DDVO [33]	Ours	COLMAP	DEMON [32]	Ours
SINTEL	MRE	0.364	0.401	0.435	0.333	-	0.478	<b>0.311</b>
	SILog	0.266	0.311	0.360	0.206	-	0.397	<b>0.172</b>
	NMG	1.212	1.041	1.492	0.993	-	1.311	<b>0.890</b>
KITTI	abs rel	0.220	0.234	<b>0.148</b>	0.230	-	0.235	<u>0.213</u>
	$\delta < 1.25$	0.632	0.617	<b>0.812</b>	0.606	-	0.605	<u>0.637</u>
KITTI	abs rel	0.227	0.231	<b>0.183</b>	0.230	-	0.307	<u>0.207</u>
	pedestrian	0.625	0.622	<b>0.746</b>	0.610	-	0.458	<u>0.654</u>
WSVD	NMG	1.269	1.152	1.505	1.012	2.021 (69.8%)	1.418	<b>0.899</b>

Table 3: Quantitative comparison of methods. All metrics in this table are *lower is better* except  $\delta < 1.25$ . Best results are shown in bold, and underlining indicates the best result second only to DDVO trained on KITTI. COLMAP returns semi-dense depth values. In this case, 69.8% of pixels had valid depths computed, we compute the metric over these pixels only.  
\* numbers reported for RedWeb is from our re-implementation of [34] due to their code has not been released.

## 5.2. Multi-view v.s. Single-view Prediction

We evaluate the effect of incorporating temporal information at *test* time, by comparing our approach against two baselines: 1) a single view depth prediction network; 2) our proposed network with two identical frames ( $I_t, I_t$ ) and a zero flow map as input. This baseline is used to verify if the improvement is truly from temporal information, instead of the difference in network architecture. Table 5, and Figure 8 show that the two baselines achieve similar performance, our proposed method with additional temporal information as input can identify foreground objects better and gives more accurate depth estimation.

Training set	Multi-view test set		
	SINTEL SILog	KITTI RSME(log)	WSVD NMG
KITTI	0.3871	<b>0.180</b>	1.620
DEMON	0.222	0.356	1.205
WSVD	<b>0.172</b>	0.317	<b>0.899</b>

Table 4: Testing result of our proposed network trained on different datasets. DEMON refers to the video datasets used for training in [32], SUN3D, RGBD, and Scenes11.

## 5.3. Training on Different Multi-view Datasets

We conduct a controlled experiment of training our multi-view depth prediction network on different multi-view datasets and compare their cross dataset generalization performance. We use scale invariant logarithm loss [6] to train on KITTI; and a combination of scale invariant logarithm depth loss and gradient loss to train on the training set (SUN3D+ RGBD + Scene11) proposed by DEMON [32], consisting mostly of rigid scenes. As shown in Table 4 and Figure 7, training on our WSVD dataset has the lowest error on Sintel and outperforms the DEMON training set on KITTI, while models learned from KITTI have the worst performance on other datasets.

## 5.4. Cross Dataset Evaluation

We compare the generalization capability of the compared methods by testing on different video datasets, i.e. Sintel, KITTI and WSVD. We notice that most of the scenes in the KITTI test set are rigid, therefore, we select a subset of 24 images with pedestrians, and use it to analyze the performance of handling non-rigid scenes for the compared methods.

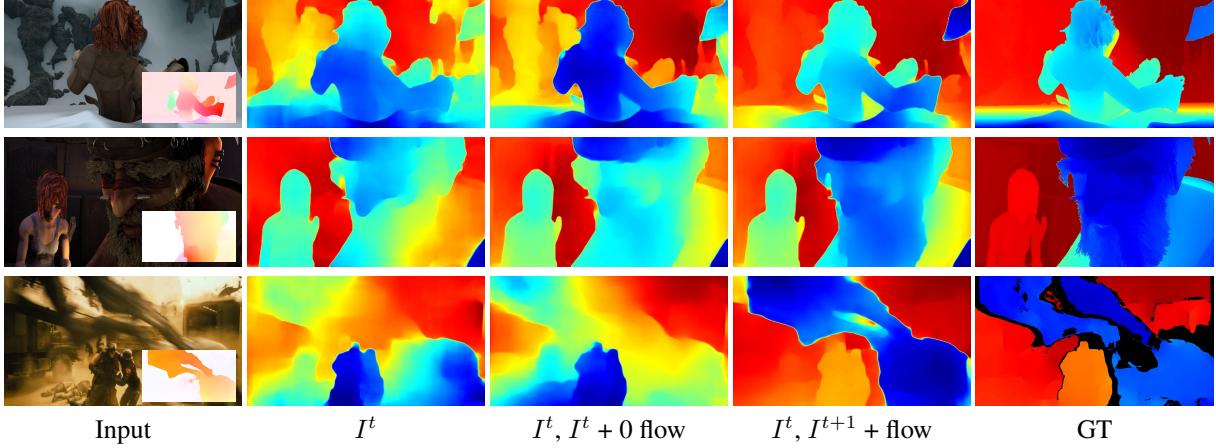


Figure 8: Depth maps predicted with different inputs, showing the impact of additional temporal information made available to the network. Using sequential frames and optical flow (4th column) improves baselines without temporal information.

Input	SINTEL			KITTI			WSVD NMG
	NMG	MRE	SILog	abs rel	sq rel	RMSE(log)	
$I^t$	0.993	0.333	0.206	0.230	1.937	0.327	1.012
$I^t, I^t, 0$ flow	0.972	0.326	0.198	0.235	2.103	0.341	0.977
$I^t, I^{t+1}, \text{flow}$	<b>0.890</b>	<b>0.311</b>	<b>0.172</b>	<b>0.213</b>	<b>1.849</b>	<b>0.317</b>	<b>0.899</b>

Table 5: Evaluate the improvement due to adding the flow and second frames as input in test time. Each row describes the input to the network. i.e.  $(I^t, I^t)$  means the same image is given twice; “0-flow”means feeding flow map filled with zeros.

Table 3 and Figure 6 show that our (single/multi-view) network compares favorably to single view depth prediction methods, i.e. MegaDepth [21], our re-implementation of RedWeb [34] and DDVO [33], which is unsupervised learned on KITTI . On the KITTI test set, MegaDepth performs similarly to ours, which could be due to their abundant training for street and landmark scenes, while our method demonstrates better result on the pedestrian subset. Not surprisingly, DDVO performs the best on KITTI since it is trained on this dataset, but it generalizes poorly on other test sets.

We also run DEMON [32] – a deep learning method for predicting depth and camera pose from two views. We find that DEMON produces plausible results for rigid scenes with sufficient parallax, but generates worse results on nonrigid scenes. In addition, we compare the quality of reconstruction on some clips from our dataset using COLMAP [30], and show quantitative results on Table 3. We note that due to running time constraints we ran COLMAP on a random subset of our data.

Throughout the evaluation above, our method consistently outperforms our single-view depth prediction baseline, which indicates that our performance gain is not solely due to our training set but also from the proposed network’s

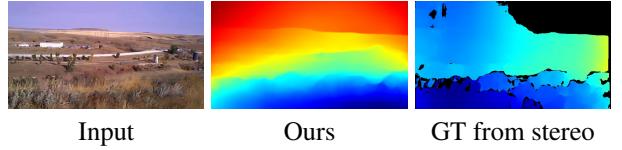


Figure 9: Limitations. Stereo supervision is less reliable at long distances or texture-less regions.

ability to take temporal information into account.

Finally, geometric-based methods by Kumar et al. [17] and Ranftl et al. [27] are related to ours, but we’re unable to provide meaningful comparison here due to their code and result is not publicly available, and the numbers reported in their paper are from an undisclosed subset of the dataset.

## 5.5. Limitations

One limitation of our approach, is that by using stereo disparity as supervision, we are restricting ourselves to scenes whose disparity can be computed. This can be problematic with large-scale scenes such as landscapes, where stereo baselines would have to be very large to have nonzero disparity. In Fig 9, we can see that our “ground-truth” is in fact incorrect in such scenes. We can also have similar difficulties in untextured regions, where reliable correspondences do not exist for supervision. In general, any biases in the reconstruction step will likely persist in our final results.

## 6. Conclusion

In conclusion, we present a step towards data-driven reconstruction of non-rigid scenes, by introducing the first in-the-wild stereo video dataset that features a wide distribution of nonrigid object types. We hope that this dataset will encourage future work in the area of diverse non-rigid video reconstruction, a topic with many exciting applications.

## References

- [1] Pyscndetect. <https://pyscenedetect.readthedocs.io>, 2008. [Online; accessed 1-July-2018]. 3
- [2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Nonrigid structure from motion in trajectory space. In *Advances in neural information processing systems*, pages 41–48, 2009. 2
- [3] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 690–696. IEEE, 2000. 2
- [4] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems*, pages 730–738, 2016. 1, 2, 4, 5
- [5] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2):101–122, 2014. 2
- [6] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 1, 2, 4, 5, 6, 7
- [7] R. Garg, V. K. BG, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016. 3
- [8] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361. IEEE, 2012. 1, 2
- [9] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017. 2, 4
- [10] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang. Learning monocular depth by distilling cross-domain stereo networks. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2
- [11] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 4
- [12] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018. 2
- [13] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE conference on computer vision and pattern recognition (CVPR)*, volume 2, page 6, 2017. 3, 4
- [14] C. Kong and S. Lucey. Prior-less compressible structure from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4123–4131, 2016. 2
- [15] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017. 4
- [16] S. Kumar, A. Cherian, Y. Dai, and H. Li. Scalable dense non-rigid structure-from-motion: A grassmannian perspective. *CVPR*, 2018. 2
- [17] S. Kumar, Y. Dai, and H. Li. Monocular dense 3d reconstruction of a complex dynamic scene from two perspective frames. In *IEEE International Conference on Computer Vision*, pages 4649–4657, 2017. 2, 8
- [18] Y. Kuznetsov, J. Stückler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6647–6655, 2017. 2
- [19] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016. 1
- [20] M. Lee, J. Cho, C.-H. Choi, and S. Oh. Procrustean normal distribution for non-rigid structure from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1280–1287, 2013. 2
- [21] Z. Li and N. Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 1, 2, 4, 7, 8
- [22] W. Lijun, S. Xiaohui, Z. Jianming, W. Oliver, H. Chih-Yao, K. Sarah, and L. Huchuan. Deeplens: Shallow depth of field from a single image. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 37(6):6:1–6:11, 2018. 2
- [23] R. Mahjourian, M. Wicke, and A. Angelova. Geometry-based next frame prediction from monocular video. In *IEEE Intelligent Vehicles Symposium, IV 2017, Los Angeles, CA, USA, June 11-14, 2017*, pages 1700–1707, 2017. 3
- [24] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, 2018. 3
- [25] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 2
- [26] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 6
- [27] R. Ranftl, V. Vineet, Q. Chen, and V. Koltun. Dense monocular depth estimation in complex dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4058–4066, 2016. 2, 8
- [28] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2009. 2
- [29] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2
- [30] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Com-*

- puter Vision and Pattern Recognition, pages 4104–4113, 2016. 8
- [31] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 5
  - [32] 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 *IEEE Conference on computer vision and pattern recognition (CVPR)*, volume 5, page 6, 2017. 1, 2, 4, 7, 8
  - [33] C. Wang, J. Miguel Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3, 7, 8
  - [34] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 311–320, 2018. 1, 2, 3, 4, 5, 6, 7, 8
  - [35] J. Xie, R. Girshick, and A. Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*, pages 842–857. Springer, 2016. 2
  - [36] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia. Lego: Learning edge with geometry all at once by watching videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 225–234, 2018. 2, 3
  - [37] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 340–349, 2018. 2
  - [38] H. Zhou, B. Ummenhofer, and T. Brox. Deeptam: Deep tracking and mapping. In *European Conference on Computer Vision (ECCV)*, 2018. 2
  - [39] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017. 3, 6
  - [40] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 3
  - [41] Y. Zhu, D. Huang, F. De La Torre, and S. Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1542–1549, 2014. 2
  - [42] Y. Zou, Z. Luo, and J.-B. Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *European Conference on Computer Vision*, 2018. 3, 5