

Learning Single-Image Depth from Videos using Quality Assessment Networks

Weifeng Chen^{1,2} Shengyi Qian¹ Jia Deng²

¹University of Michigan, Ann Arbor

{wfchen, syqian}@umich.edu

²Princeton University

jiadeng@cs.princeton.edu

Abstract

Depth estimation from a single image in the wild remains a challenging problem. One main obstacle is the lack of high-quality training data for images in the wild. In this paper we propose a method to automatically generate such data through Structure-from-Motion (SfM) on Internet videos. The core of this method is a Quality Assessment Network that identifies high-quality reconstructions obtained from SfM. Using this method, we collect single-view depth training data from a large number of YouTube videos and construct a new dataset called YouTube3D. Experiments show that YouTube3D is useful in training depth estimation networks and advances the state of the art of single-view depth estimation in the wild. Project website: <http://www-personal.umich.edu/~wfchen/youtube3d>.

1. Introduction

This paper addresses the problem of single-image depth estimation, a fundamental computer vision problem that remains challenging. Despite significant recent progress [45, 15, 35, 24, 17, 27, 46, 49, 11, 25, 22, 50, 23, 13, 43, 54, 20, 44], current systems still perform poorly on arbitrary images in the wild [6]. One major obstacle is the lack of diverse training data, as most existing RGB-D datasets were collected via depth sensors and are limited to rooms [39, 10, 5] and roads [14]. As shown by recent work [6], systems trained on such data are unable to generalize to diverse scenes in the real world.

One way to address this data issue is crowdsourcing, as demonstrated by Chen et al. [6], who crowdsourced human annotations of depth and constructed a dataset called “Depth-in-the-Wild (DIW)” that captures a broad range of scenes. One drawback, though, is that it requires a large amount of manual labor. Another possibility is to use synthetic data [4, 28, 34, 21], but it remains unclear how to automatically generate scenes that match the diversity of real-world images.

In this paper we explore a new approach that automat-

ically collects single-view training data on natural in-the-wild images, without the need for crowdsourcing or computer graphics. The idea is to reconstruct 3D points from Internet videos using Structure-from-Motion (SfM), which matches feature points across video frames and infers depth using multiview geometry. The reconstructed 3D points can then be used to train single-view depth estimation. Because there is a virtually unlimited supply of Internet videos, this approach is especially attractive for generating a large amount of single-view training data.

However, to implement such an approach in practice, there remains a significant technical hurdle—despite great successes [1, 19, 36, 37, 30], existing SfM systems are still far from reliable when applied to arbitrary Internet videos. This is because SfM operates by matching features across video frames and reconstructing depth assuming a static scene, but feature matches are often unreliable and scenes often contain moving objects, both of which cause SfM to produce erroneous 3D reconstructions. That is, if we simply apply an off-the-shelf SfM system to arbitrary Internet videos, the resulting single-view training data will have poor quality.

To address this issue, we propose to train a deep network to automatically assess the quality of a SfM reconstruction. The network predicts a quality score of a SfM construction by examining the operation of the entire SfM pipeline—the input, the final output, along with intermediate outputs generated inside the pipeline. We call this network a *Quality Assessment Network (QANet)*. Using a QANet, we filter out unreliable reconstructions and obtain high-quality single-view training data. Fig. 1 illustrates our data collection method.

It is worth noting that because Internet videos are virtually unlimited, it is sufficient for a QANet to be able to reliably identify a small proportion of high-quality reconstructions. In other words, high precision is necessary but high recall is not. This means that training a QANet will not be hopelessly difficult because we do not need to detect every good reconstruction, only *some* good reconstructions.

We experiment using Internet videos in the wild. Our experiments show that with QANet integrated with SfM, we

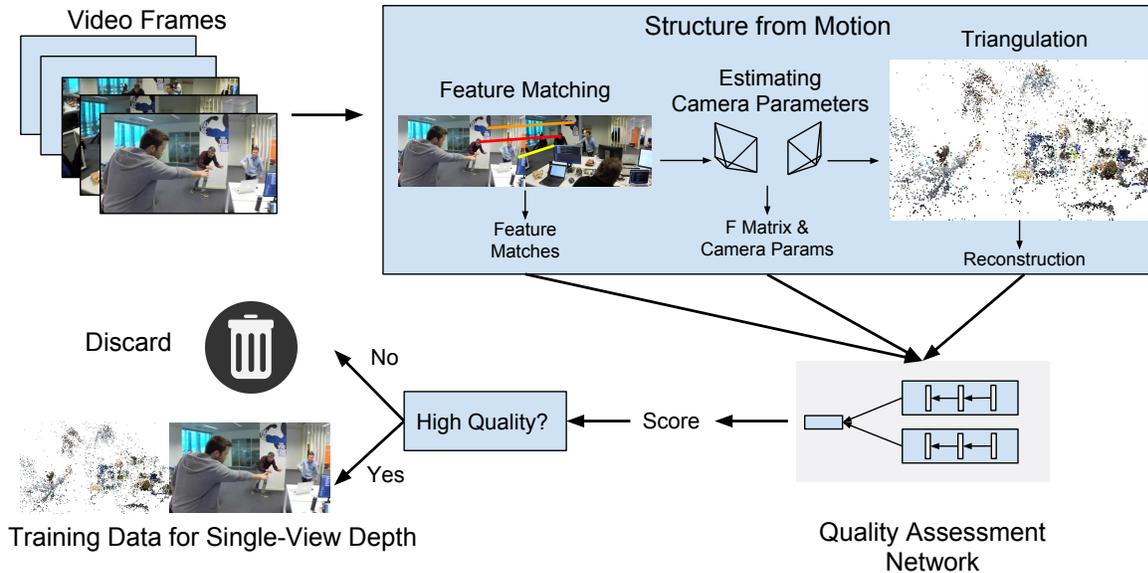


Figure 1. An overview of our data collection method. Given an arbitrary video, we follow standard steps of structure-from-motion: extracting feature points and matching them across frames, estimating the camera parameters, and performing triangulation to obtain a reconstruction. A Quality Assessment Network (QANet) examines the operation of the SfM pipeline and assigns a score to the reconstruction. If the score is above a certain threshold, this reconstruction is deemed of high quality, and we use it as single-view depth training data. Otherwise, the reconstruction is discarded.

can collect high-quality single-view training data from unlabeled videos, and such training data can supplement existing data to significantly improve the performance of single-image depth estimation.

Using our proposed method, we constructed a new dataset called YouTube3D, which consists of 795K in-the-wild images, each associated with depth annotations generated from SfM reconstructions filtered by a QANet. We show that as a standalone training set for in-the-wild depth estimation, YouTube3D is superior to existing datasets constructed with human annotation. YouTube3D also outperforms MegaDepth [26], a recent dataset automatically collected through SfM on Internet images. In addition, we show that as a supplement to existing RGB-D data, YouTube3D advances the state-of-the-art of single-image depth estimation in the wild.

Our contributions are two fold: (1) we propose a new method to automatically collect high-quality training data for single-view depth by integrating SfM and a quality assessment network; (2) using this method we construct YouTube3D, a large-scale dataset that advances the state of the art of single-view depth estimation in the wild.

2. Related Work

RGB-D from depth sensors A large amount of RGB-D data from depth sensors has played a key role in driving recent research on single-image depth estimation [14, 39, 5,

10, 38]. But due to the limitations of depth sensors and the manual effort involved in data collection, these datasets lack the diversity needed for arbitrary real world scenes. For example, KITTI [14] consists mainly of road scenes; NYU Depth [39], ScanNet [10] and Matterport3D [5] consist of only indoor scenes. Our work seeks to address this drawback by focusing on diverse images in the wild.

RGB-D from computer graphics RGB-D from computer graphics is an attractive option because the depth will be of high quality and it is easy to generate a large amount. Indeed, synthetic data has been used in computer vision with much success [16, 42, 28, 41, 4, 12, 8, 47, 33]. In particular, SUNCG [40] has been shown to improve single-view surface normal estimation on natural indoor images from the NYU Depth dataset [53]. However, the diversity of synthetic data is limited by the availability of 3D “assets”, i.e. shapes, materials, layouts, etc., and it remains difficult to automatically compose diverse scenes representative of the real world.

RGB-D from crowdsourcing Crowdsourcing depth annotations [6, 7] has recently received increasing attention. It’s appealing because it can be applied to a truly diverse set of in-the-wild images. Chen et al. [6] crowdsourced annotations of relative depth and constructed Depth in the Wild (DIW), a large-scale dataset for single-view depth in the wild. The main drawback of crowdsourcing is, obviously, the cost of manual labor, and our work attempts to mitigate

or avoid this cost through an automatic method.

RGB-D from multiview geometry When multiple images of the same scene are available, depth can be reconstructed through multiview geometry. Prior work has exploited this fact to collect RGB-D data. Xian et al. [48] perform stereopsis on stereo images, i.e. pairs of images taken by two calibrated cameras, to collect a dataset called “ReDWeb”. Li et al. [26] perform SfM on unordered collections of on-line images of the same scenes to collect a dataset called “MegaDepth”.

Our work differs from prior work in two ways. First, we use a new source of RGB data—monocular videos—which likely offer better availability and diversity—stereo images have limited availability because they must be taken by stereo cameras. Multiple images of the same scene tend to be biased toward well-known sites frequented by tourists.

Second, our method of quality assessment is new. Both prior works performed some form of quality assessment, but neither used learning. Xian et al. [48] manually remove some poor reconstructions; Li et al. [26] use handcrafted criteria based on semantic segmentation. In contrast, our quality assessment network can learn criteria and patterns beyond those that are easy to handcraft.

Predicting failure Our work is also related to prior work on predicting failures for vision systems [52, 9, 3, 2]. For example, Zhang et al. [52] predict failure for a variety of vision tasks based solely on the input. Daftry et al. [9] predict failures in an autonomous navigation system directly from the input video stream. Our method is different in that we predict failure in a SfM system to filter reconstructions, based not on the input images but on the outputs of the SfM system.

3. Approach

Our method consists of two main steps: SfM followed by quality assessment, as illustrated by Fig. 1. SfM produces candidate 3D reconstructions, which are then filtered by a QANet before we use them to generate single-view training data.

3.1. Structure from Motion

The SfM component of our method is standard. We first detect and match features across frames. We then estimate the fundamental matrix and perform triangulation to produce 3D points.

It is worth noting that SfM produces only a sparse reconstruction. Although we can generate a dense point cloud by a subsequent step of multiview stereopsis, we choose to forgo it, because stereopsis in unconstrained settings tends to contain a large amount of error, especially in the presence of low-texture surfaces or moving objects.

Our SfM component also involves a couple minor modifications compared to a standard full-fledged SfM system. First, we only perform two-view reconstruction. This is to simplify the task of quality assessment—the quality assessment network only needs to examine two input images as opposed to many. Second, we do not perform bundle adjustment [18], because we observe that with unknown focal length of Internet videos (we assume a centered principal point and focal length is the only unknown intrinsic parameter), it often leads to poor results. This is because bundle adjustment is sensitive to initialization, and tends to converge to an incorrect local minimum if the initialization of focal length is not already close to correct. Instead, we search a range of focal lengths and pick the one that leads to the smallest reprojection error after triangulation. This approach does not get stuck in local minima, and is justified by the fact that focal length can be uniquely determined when it is the only unknown intrinsic parameter of a fixed camera across two views [31].

3.2. Quality Assessment Network (QANet)

The task of a quality assessment network is to identify good SfM reconstructions and filter out bad ones. In this section we discuss important design decisions including the input, output, architecture, and training of a QANet.

Input to QANet The input to a QANet should include a variety of cues from the operation of a SfM pipeline on a particular input. Recall that we consider only two-view reconstruction; thus the input to SfM is only two video frames.

We consider cues associated with the entire reconstruction (reconstruction-wise cues) as well as those associated with each reconstructed 3D point (point-wise cues). Our reconstruction-wise cues include the inferred focal length and the average reprojection error. Our point-wise cues include the 2D coordinates of a feature match, the Sampson distance of a feature match under the recovered fundamental matrix, and the angle between the two rays connecting the reconstructed 3D point and the camera centers.

Note that we do not use any information from the pixel values. The QANet only has access to geometrical information of the matched features. This is to allow better generalization by preventing overfitting to image content.

Also note that in a SfM pipeline RANSAC is typically used to handle outliers. That is, multiple reconstructions are attempted on random subsets of the feature matches. Here we apply the QANet only to the best subset free from outliers.

Output of QANet The output of a QANet is a quality score for the entire reconstruction, i.e. a sparse point cloud.

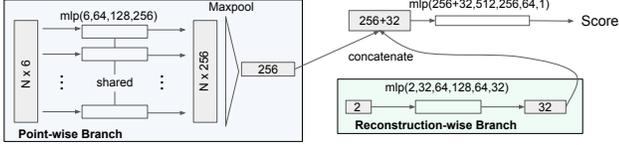


Figure 2. Architecture of the Quality Assessment Network (QANet).

Ideally, this score should correspond to a similarity metric between two point clouds, the reconstructed one and the ground truth.

There are many possible choices of the similarity metric, with different levels of invariance and robustness (e.g. invariance to scale, and robustness to deformation and outliers). Which one to use should be application dependent and is not the main concern of this work. And it is sufficient to note that our method is general and not tied to a particular similarity metric.

QANet architecture Fig. 2 illustrates the architecture of our QANet. It consists of two branches. The *reconstruction-wise branch* processes the reconstruction-wise cues (the focal length and overall reprojection error). The *point-wise branch* processes features associated with each reconstructed point. The outputs from the two branches are then concatenated and fed into multiple fully connected layers to produce a quality score.

Point-wise cues need a separate branch because they involve an unordered set of feature vectors with a variable size. To be invariant to the number and ordering of the vectors, we employ an architecture similar to that of PointNet [32]. In this architecture, each vector is independently processed by shared subnetwork and the results are max-pooled at the end.

QANet training To train a QANet, a straightforward approach is to use a regression loss that minimizes the difference between the predicted quality score and the ground truth score—the similarity between the reconstructed 3D point cloud and the ground truth.

However, using a regression loss makes learning harder than necessary. In fact, the absolute value of the score matters much less than the ordering of the score, because when we use a QANet for filtering, we remove all reconstructions with scores below a threshold, which can be chosen by cross-validation. In other words, the network just needs to tell that one construction is better than another, but does not need to quantify the exact degree. Moreover, the precision of top-ranked reconstructions is much more important than the rest, and should be given more emphasis in the loss.

This observation motivates us to use a ranking loss. Let s_1 be the “ground truth quality score” (i.e. similarity to the

ground truth reconstruction) of a reconstruction in the training set. Let s'_1 be its predicted quality score by the QANet. Similarly, let s_2 be the ground truth quality of another reconstruction, and let s'_2 be the predicted quality score. We define a ranking loss $h(s'_1, s'_2, s_1, s_2)$ on this pair of reconstructions:

$$h(s'_1, s'_2, s_1, s_2) = \begin{cases} \ln(1 + \exp(s'_2 - s'_1)), & \text{if } s_1 > s_2 \\ \ln(1 + \exp(s'_1 - s'_2)), & \text{if } s_1 < s_2 \end{cases} \quad (1)$$

This loss imposes a penalty if the score ordering of the pair is incorrect. When applied to all possible pairs, it generates a very large total penalty if a bad reconstruction is ranked top, because many pairs will have the wrong ordering. Obviously, in practice we cannot afford to train with all possible pairs. Instead, we uniformly sample random pairs whose difference in ground truth quality scores are larger than some threshold.

4. Experiments

Relative depth One implementation question we have left open in the previous sections is the choice of the “ground truth” quality score for the QANet. Specifically, to train an actual QANet, we need a similarity metric that compares a reconstructed point cloud with the ground truth point cloud (the clouds have the same number of points and known correspondence).

In our experiments we define the similarity metric based on relative depth. We consider all pairs of points in the reconstructed cloud, and calculate the percentage of pairs that have the same depth ordering as the ground truth. Note that depth ordering is view dependent, and because our SfM component performs two-view reconstruction, we take the average from both views.

Our choice of relative depth as the quality measure is motivated by two reasons. First, relative depth is more robust to outliers. Unlike metrics based on metric difference such as RMSE, with relative depth a single outlier point will not be able to dominate the error. Second, relative depth has been used as a standard evaluation metric for depth prediction in the wild [6, 24, 48, 51], partly because it would be difficult to obtain ground truth for arbitrary Internet images except to use humans, which are good at annotating relative depth but not metric depth.

Another implementation question is how to train a single-view depth network with the single-view data generated by our method, i.e. 3D points from SfM filtered by the QANet. Here we opt to also derive relative depth from the 3D points. In other words, the final form of our automatically collected training data is a set of video frames, each associated with a set of 2D points with their “ground truth” depth ordering.

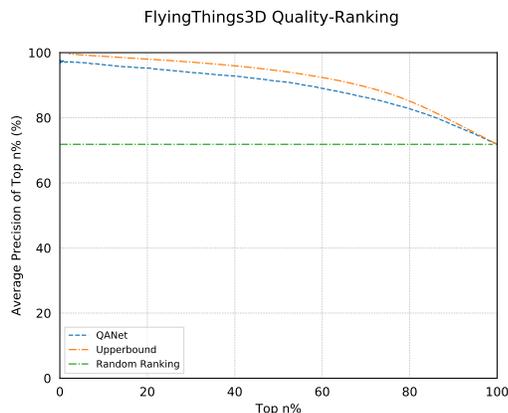


Figure 3. The quality-ranking curve on the FlyingThings3D dataset.

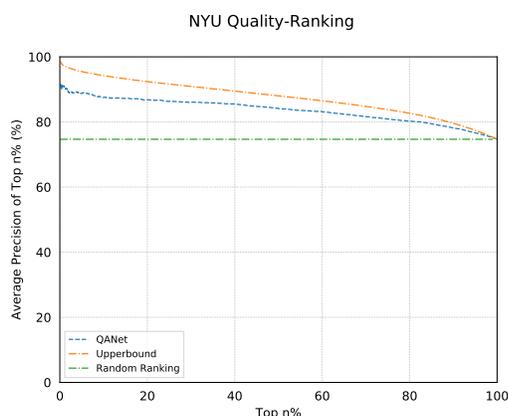


Figure 4. The quality-ranking curve on the NYU dataset.

One advantage of using relative depth as training data is that it is scale-invariant and sidesteps the issue of scale ambiguity in our SfM reconstructions. In addition, prior work [6] has shown that relative depth can serve as a good source of supervision even when the goal is to predict dense metric depth. Last but not least, using relative depth allows us to compare our automatically collected data with prior work such as MegaDepth [26], which also generates training data in the form of relative depth.

4.1. Evaluating QANet

We first evaluate whether the QANet, as a standalone component, can be successfully trained to identify high-quality reconstructions.

We train the QANet using a combination of existing RGB-D video datasets: NYU Depth [39], FlyingThings3D [28], and SceneNet [29]. We use the RGB videos to produce SfM reconstructions and use the depth maps to compute the ground truth quality score for each reconstruction.

We measure the performance of our QANet by plotting a quality-ranking curve—the Y-axis is the average ground-

| QANet Variants | AUC | |
|----------------|---------------|----------------|
| | NYU | FlyingThings3D |
| -2D | 80.53% | 85.34% |
| -Sam | 83.20% | 88.66% |
| -Ang | 82.09% | 85.00% |
| -Focal | 82.54% | 88.37% |
| -RepErr | 83.37% | 88.50% |
| Full | 83.56% | 89.02% |
| Upperbound | 87.49% | 91.28% |
| Random Ranking | 75.09% | 71.41% |

Table 1. AUC (area under curve) for different ablated versions of the QANet.

truth quality (i.e. percentage of correct relative depth orderings) of the top $n\%$ reconstructions ranked by QANet, and the X-axis is the number n . At the same n , a better QANet would have a better average quality.

We test our QANet on the test splits of FlyingThings3D and NYU Depth. The results are shown in Fig. 3 and Fig. 4. In both figures, we provide an *Upperbound* curve from a perfect ranking of the reconstructions, and a *Random Ranking* curve from a random ranking of the reconstructions.

From Fig. 3 and Fig. 4 we see that our QANet can successfully rank reconstructions by quality. On FlyingThings3D, the average quality of unfiltered (or randomly ranked) reconstructions is 71.41%, whereas the top 20% reconstructions ranked by QANet have an average quality of 95.26%. On NYU Depth, the numbers are 75.09% versus 86.80%.

In addition, we see that the QANet curve is quite close to the upperbound curve. On FlyingThings3D, the AUC (area under curve) of the upperbound curve is 91.28%, and the AUC of QANet is 89.02%. On NYU Depth, the numbers are 87.49% and 83.56%.

Ablative Studies We next study the contributions of different cues to quality assessment. We train five ablated versions of QANet by (1) removing 2D coordinate feature (-2D); (2) removing Sampson distance feature (-Sam); (3) removing angle feature (-Ang); (4) removing focal length (-Focal); (5) removing reprojection error (-RepErr).

We compare their performances in terms of AUC with the full QANet in Tab. 1. They all underperform the full QANet, indicating that all cues contribute to successful quality assessment.

4.2. Evaluating the full method

We now turn to evaluating our full data collection method. To this end, we need a way to compare our dataset with those collected by alternative methods.

Note that it is insufficient to compare datasets using the accuracy of the ground truth labels, because the datasets



Figure 5. Examples of automatically collected relative depth annotations in YouTube3D. The relative depth pairs are visualized as two connected points, with red point being closer than the blue point. These relative depth annotations are mostly correct.

may have different numbers of images, different images, or different annotations on the same images (e.g. different pairs of points for relative depth). A dataset may have less accurate labels, but may still end up more useful due to other reasons such as better diversity or more informative annotations.

Instead, we compare datasets by their usefulness for training. In our case, a dataset is better if it trains a better deep network for single-view depth estimation. Given a dataset of relative depth, we use the method of Chen et al. [6] to train an image-to-depth network by imposing a ranking loss on the output depth values to encourage agreement with the ground truth orderings. We measure the performance of the trained network by the weighted human disagreement rate (*WHDR*) [6], i.e. the percentage of incorrectly ordered point pairs.

YouTube3D We crawled 0.9 million YouTube videos using random keywords. Pairs of frames are randomly sampled and selected if feature matches exist between them. We apply our method to these pairs and obtain 2 million filtered reconstructions spanning 121,054 videos. From these reconstructions we construct a dataset called *YouTube3D*, which consists of 795,066 images, with an average of 281 relative depth pairs per image. Example images and annotations of YouTube3D are shown in Fig. 5.

As a baseline, we construct another dataset called *YT_{UF}*. It is built from all reconstructions that are used in constructing YouTube3D but without applying the QANet filtering. Note that *YT_{UF}* is a superset of YouTube3D, and contains 3.5M images.

Colmap Our implementation of SfM is adapted from Colmap [36], a state-of-the-art SfM system. We use the same feature matches generated by Colmap, and modified the remaining steps as described in Sec. 3.1. In our experiments, we also include the original unmodified Colmap system as a baseline. To generate relative depth from the sparse point clouds given by Colmap, we randomly sample point pairs and project them into different views.

We run Colmap on the same set of features and matches

| Training Sets | WHDR |
|-------------------|---------------|
| NYU | 31.31% [6] |
| DIW | 22.14% [6] |
| MegaDepth | 22.97% [26] |
| YT _{Col} | 34.47% |
| YT _{UF} | 25.11% |
| QA_train | 31.77% |
| NYU + QA_train | 31.22% |
| YouTube3D | 19.01% |

Table 2. Error rate on the DIW test set by the Hourglass Network [6] trained on different standalone datasets.

as used in constructing YouTube3D and *YT_{UF}*, obtaining 647,143 reconstructions that span 486,768 videos. From them we construct a dataset called *YT_{Col}*. It contains 3M images, with an average of 4,755 relative depth pairs per image.

Depth-in-the-Wild (DIW) We use the Depth-in-the-Wild (DIW) dataset [6] to evaluate the performance of a single-view depth network. DIW consists of Internet images that cover diverse types of scenes. It has 74,000 test and 420,000 train images; each image has human annotated relative depth for one pair of points. In addition to using the test split of DIW for evaluation, we also use its training split as a standalone training set.

Evaluation as standalone dataset We evaluate YouTube3D as a standalone dataset and compare it with other datasets. That is, we train a single-view depth network from scratch using each dataset and measure the performance on DIW. To directly compare with existing results in the literature, we use the same hourglass network that has been used in a number of prior works [6, 26].

Tab. 2 compares the DIW performance of a hourglass network trained on YouTube3D against those trained on three other datasets: MegaDepth [24], NYU Depth [39], and the training split of DIW [6]. The results are shown in Tab. 2. We see that YouTube3D not only outperforms

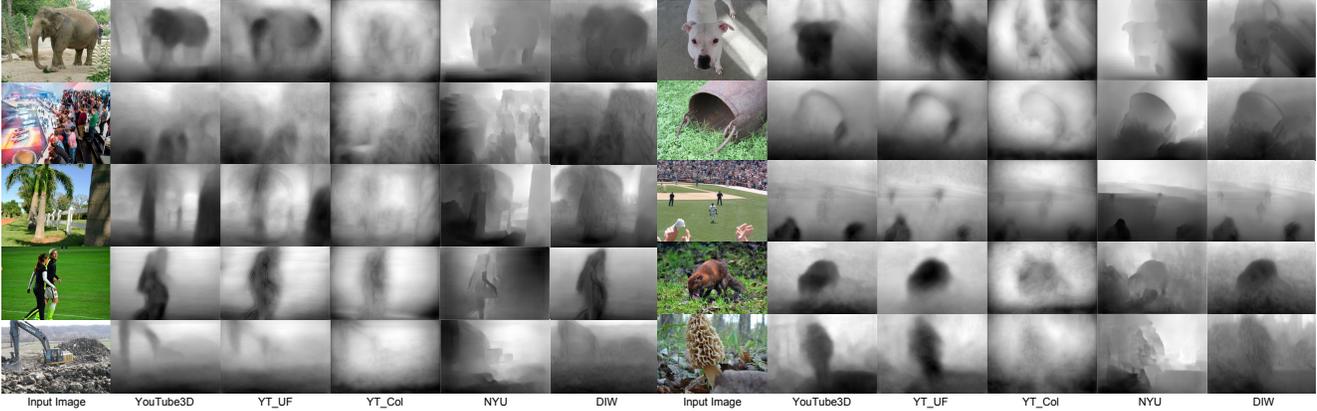


Figure 6. Qualitative results on the DIW test set by the Hourglass Network [6] trained with different datasets. Column names denote the datasets used for training.

NYU Depth, which was acquired with depth sensors, but also MegaDepth, another high-quality depth dataset collected via SfM. Most notably, even though the evaluation is on DIW, YouTube3D outperforms the training split of DIW, showing that our automatic data collection method is a viable substitute for manual annotation.

Tab. 2 also compares YouTube3D against YT_{UF} (YouTube3D without QANet filtering) and YT_{Col} (off-the-shelf SfM). We see that YouTube3D outperforms the unfiltered set YT_{UF} by a large margin, even though YT_{UF} is a much larger superset of YouTube3D. This underscores the effectiveness of QANet filtering. Moreover, YouTube3D outperforms YT_{Col} by an even larger margin, indicating our method is much better than a direct application of off-the-shelf state-of-the-art SfM to Internet videos. Notably, YT_{UF} already outperforms YT_{Col} significantly. This is a result of our modifications described in Sec. 3.1: (1) we require the estimate of the fundamental matrix to have zero outliers during RANSAC; (2) we replace bundle adjustment with a grid-search of focal length.

Fig. 6 shows a qualitative comparison of depth estimation by networks trained with different datasets. We can see that training on YouTube3D generally produces better results than others, especially compared to YT_{Col} and NYU.

We also include a comparison between YouTube3D and QA_{train} , the data used to train QANet. This is to answer the question whether a naive use of this extra data—using it directly to train a single-view depth network—would give the same advantage enjoyed by YouTube3D, rendering our method unnecessary. We see in Tab. 2 that training single-view depth directly from QA_{train} is much worse than YouTube3D (31.77% vs. 19.01%), showing that QA_{train} itself is a not a good training set for mapping pixels to depth. In addition, adding QA_{train} to NYU Depth (NYU + QA_{train} in Tab. 2) barely improves the performance of NYU Depth alone. This shows that a naive use of this extra data will not result in the improvement achiev-

| Network | Training Sets | WHDR |
|---------------------------------|--|---------------|
| Hourglass [6] | NYU + DIW | 14.39% [6] |
| | NYU + DIW + YouTube3D | 13.50% |
| EncDecResNet [48] | ImageNet + ReDWeb | 14.33% |
| | ImageNet + ReDWeb + DIW | 11.37% |
| EncDecResNet (Our Impl of [48]) | ImageNet + ReDWeb | 16.31% |
| | ImageNet + YouTube3D | 16.21% |
| | ImageNet + ReDWeb + DIW | 12.03% |
| | ImageNet + ReDWeb + DIW + YouTube3D | 10.59% |

Table 3. Error rate on the DIW test set by networks trained with and without YouTube3D as supplement.

able by our method. It also shows that QANet generalizes well to images in the wild, even when trained on data that is quite different in terms of pixel content. It is worth noting that this result should not be surprising, because QANet does not use pixel values to assess quality and only uses the geometry of the feature matches.

Evaluation as supplemental dataset We evaluate YouTube3D as supplemental data. Prior works have demonstrated state-of-the-art performance on DIW by combining multiple sources of training data [6, 48]. We investigate whether adding YouTube3D as additional data would improve state-of-the-art systems.

We first add YouTube3D to NYU + DIW, the combined training set used by Chen et al. [6] to train the first state-of-the-art system for single-view depth in the wild. We train the same hourglass network used in [6]. Results in Tab. 3 show that with the addition of YouTube3D, the network is able to achieve a significant improvement.

We next evaluate whether YouTube3D can improve the best existing result on DIW, achieved by an encoder-decoder network based on ResNet50 [48] (which we will refer to as an EncDecResNet subsequently). The network is trained on a combination of ImageNet, DIW, and ReDWeb, a relative depth dataset collected by performing stereopsis on stereo images with manual removal of poor-quality reconstructions. Tab. 3 summarizes our results, which we

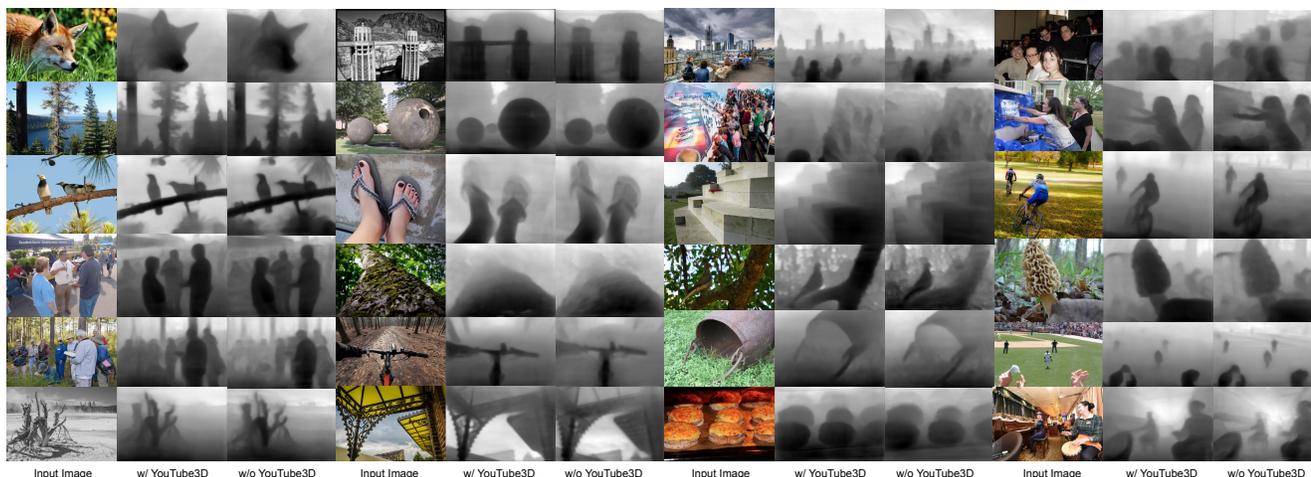


Figure 7. Qualitative results on the DIW test set by the EncDecResNet [6] trained on ImageNet + ReDWeb + DIW (*w/o YouTube3D*), and fine-tuned on YouTube3D (*w/ YouTube3D*).

elaborate below.

We implement our own version of the EncDecResNet used in [48], because there is no public code available as of writing. As a validation of our implementation, we train the network on ImageNet and ReDWeb, and achieve an error rate of 16.31%, which is slightly worse than but sufficiently close to the 14.33% reported in [48]¹. This discrepancy is likely because certain details (e.g. the exact number of channels at each layer) are different in our implementation because they are not available in their paper.

As an aside, we train the same EncDecResNet on ImageNet and YouTube3D, which gives an error rate of 16.21%, which is comparable with the 16.31% given by ImageNet and ReDWeb. This suggests that YouTube3D is as useful as ReDWeb. This is noteworthy because unlike ReDWeb, YouTube3D is not restricted to stereo images and does not involve any manual filtering. Note that it is not meaningful to compare with the 14.33% reported in [48]—to compare two training datasets we need to train the exact same network, but the 14.33% is likely from a slightly different network due to the unavailability of some details in [48].

Finally, we train an EncDecResNet on the combination of ImageNet, DIW, and ReDWeb, which has produced the current state of the art on DIW in [48]. With our own implementation we achieve an error rate of 12.03%, slightly worse than the 11.37% reported in [48]. Adding YouTube3D to the mix, we achieve an error rate of 10.59%, a new state of the art performance on DIW (see Fig. 7 for example depth estimates). This result demonstrates the effectiveness of YouTube3D as supplemental single-view training data.

¹All results in [48] are with ImageNet.

Discussion The above results suggest that our proposed method can generate high-quality training data for single-view depth in the wild. Such results are significant, because our dataset is gathered by a *completely automatic* method, while datasets like DIW [6] and ReDWeb [48] are constrained by manual labor and/or the availability of stereo images. Our automatic method can be readily applied to a much larger set of Internet videos and thus has potential to advance the state of the art of single-view depth even more significantly.

5. Conclusion

In this paper we propose a fully automatic and scalable method for collecting training data for single-view depth from Internet videos. Our method performs SfM and uses a Quality Assessment Network to find high-quality reconstructions, which are used to produce single-view depth ground truths. We apply the proposed method on YouTube videos and construct a single-view depth dataset called YouTube3D. We show that YouTube3D is useful both as a standalone and as a supplemental dataset in training depth predictors. With it, we obtain state-of-the-art results on single-view depth estimation in the wild.

6. Acknowledgment

This publication is based upon work partially supported by National Science Foundation under Grant No. 1617767, the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. OSR-2015-CRG4-2639 and a gift from Google.

References

- [1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski.

- Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.
- [2] Aayush Bansal, Ali Farhadi, and Devi Parikh. Towards transparent systems: Semantic characterization of failure modes. In *European Conference on Computer Vision*, pages 366–381. Springer, 2014.
- [3] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.
- [4] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision*, Part IV, LNCS 7577, pages 611–625, Oct. 2012.
- [5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.
- [6] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems*, pages 730–738, 2016.
- [7] Weifeng Chen, Donglai Xiang, and Jia Deng. Surface normals in the wild. In *Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy*, pages 22–29, 2017.
- [8] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016.
- [9] Shreyansh Daftry, Sam Zeng, J Andrew Bagnell, and Martial Hebert. Introspective perception: Learning to predict failures in vision systems. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 1743–1750. IEEE, 2016.
- [10] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *arXiv preprint arXiv:1702.04405*, 2017.
- [11] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.
- [12] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, volume 2, page 6, 2017.
- [13] Ravi Garg, BG Vijay Kumar, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016.
- [14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.
- [15] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. *arXiv preprint arXiv:1609.03677*, 2016.
- [16] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. Atlasnet: A paper-maché approach to learning 3d surface generation. *arXiv preprint arXiv:1802.05384*, 2018.
- [17] Christian Hane, Lubor Ladicky, and Marc Pollefeys. Direction matters: Depth estimation with a surface normal classifier. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 381–389, 2015.
- [18] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.
- [19] Jared Heinly, Johannes L Schonberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3287–3295, 2015.
- [20] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. *arXiv preprint arXiv:1708.05375*, 2017.
- [21] Philipp Krähenbühl. Free supervision from video games. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2955–2964, 2018.
- [22] Yevhen Kuznetsov, Jörg Stückler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. *arXiv preprint arXiv:1702.02706*, 2017.
- [23] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.
- [24] Bo Li, Chunhua Shen, Yuchao Dai, Anton van den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1119–1127, 2015.
- [25] Jun Li, Reinhard Klein, and Angela Yao. Learning fine-scaled depth maps from single rgb images. *arXiv preprint arXiv:1607.00730*, 2016.
- [26] Zhengqi Li and Noah 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.
- [27] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170, 2015.
- [28] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016.

- [29] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth. *arXiv preprint arXiv:1612.05079*, 2016.
- [30] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.
- [31] Marc Pollefeys, Reinhard Koch, and Luc Van Gool. Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. *International Journal of Computer Vision*, 32(1):7–25, 1999.
- [32] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016.
- [33] Konstantinos Rematas, Ira Kemelmacher-Shlizerman, Brian Curless, and Steve Seitz. Soccer on your tabletop. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4738–4747, 2018.
- [34] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118. Springer, 2016.
- [35] Anirban Roy and Sinisa Todorovic. Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5506–5514, 2016.
- [36] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [37] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2016.
- [38] Thomas Schöps, Johannes L Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proc. CVPR*, volume 3, 2017.
- [39] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.
- [40] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *arXiv preprint arXiv:1611.08974*, 2016.
- [41] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In *European Conference on Computer Vision*, pages 322–337. Springer, 2016.
- [42] Shubham Tulsiani, Saurabh Gupta, David Fouhey, Alexei A Efros, and Jitendra Malik. Factoring shape, pose, and layout from the 2d image of a 3d scene. *arXiv preprint arXiv:1712.01812*, 2017.
- [43] Benjamin Ummehofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 5, 2017.
- [44] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfmmnet: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017.
- [45] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018.
- [46] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2809, 2015.
- [47] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, pages 82–90, 2016.
- [48] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 311–320, 2018.
- [49] Junyuan Xie, Ross Girshick, and Ali 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.
- [50] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. *arXiv preprint arXiv:1704.02157*, 2017.
- [51] Xiangyu Xu, Deqing Sun, Sifei Liu, Wenqi Ren, Yu-Jin Zhang, Ming-Hsuan Yang, and Jian Sun. Rendering portraits from monocular camera and beyond. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–50, 2018.
- [52] Peng Zhang, Jiuling Wang, Ali Farhadi, Martial Hebert, and Devi Parikh. Predicting failures of vision systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3566–3573, 2014.
- [53] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5057–5065. IEEE, 2017.
- [54] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.