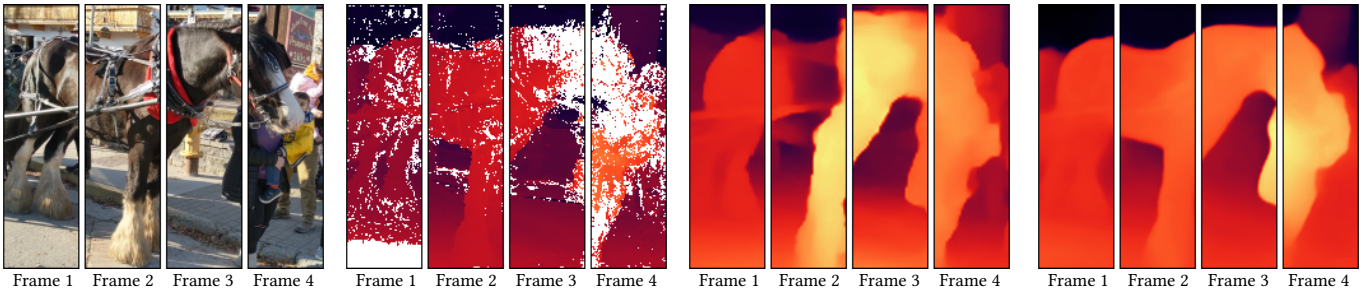
Consistent Video Depth Estimation

XUAN LUO*, University of Washington
JIA-BIN HUANG, Virginia Tech
RICHARD SZELISKI, Facebook
KEVIN MATZEN, Facebook
JOHANNES KOPF, Facebook



(a) Input video (b) COLMAP depth (c) Mannequin Challenge depth (d) Our result

Fig. 1. We present a system for estimating temporally coherent and geometrically consistent depth from a casually captured video. Conventional multi-view stereo methods such as COLMAP [Schonberger and Frahm 2016] often produce incomplete depth on moving objects or poorly textured areas. Learning-based methods (e.g., [Li et al. 2019]) predict dense depth for each frame but the video reconstruction is flickering and geometrically inconsistent. Our video depth estimation is fully dense, globally scale-consistent, and capable of handling dynamically moving objects. We evaluate our method on a wide variety of challenging videos and show that our results enable new video special effects.

We present an algorithm for reconstructing dense, geometrically consistent depth for all pixels in a monocular video. We leverage a conventional structure-from-motion reconstruction to establish geometric constraints on pixels in the video. Unlike the ad-hoc priors in classical reconstruction, we use a learning-based prior, i.e., a convolutional neural network trained for single-image depth estimation. At test time, we fine-tune this network to satisfy the geometric constraints of a particular input video, while retaining its ability to synthesize plausible depth details in parts of the video that are less constrained. We show through quantitative validation that our method achieves higher accuracy and a higher degree of geometric consistency than previous monocular reconstruction methods. Visually, our results appear more stable. Our algorithm is able to handle challenging hand-held captured input videos with a moderate degree of dynamic motion. The improved quality of the reconstruction enables several applications, such as scene reconstruction and advanced video-based visual effects.

CCS Concepts: • **Computing methodologies** → **Reconstruction**; *Computational photography*.

Additional Key Words and Phrases: video, depth estimation

*This work was done while Xuan was an intern at Facebook.

Authors' addresses: Xuan Luo, University of Washington; Jia-Bin Huang, Virginia Tech; Richard Szeliski, Facebook; Kevin Matzen, Facebook; Johannes Kopf, Facebook.

0730-0301/2020/7-ART71 $15.00
https://doi.org/10.1145/3386569.3392377

ACM Reference Format:
Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. 2020. Consistent Video Depth Estimation. *ACM Trans. Graph.* 39, 4, Article 71 (July 2020), 13 pages. https://doi.org/10.1145/3386569.3392377

1 INTRODUCTION

3D scene reconstruction from image sequences has been studied in our community for decades. Until a few years ago, the structure from motion systems for solving this problem were not very robust, and practically only worked "in the lab", with highly calibrated and predictable setups. They also, often, produced only sparse reconstructions, i.e., resolving depth at only a few isolated tracked point features. But in the last decade or so, we have seen good progress towards enabling more *casual* capture and producing *denser* reconstructions, driven by high-quality open-source reconstruction systems and recent advances in learning-based techniques, as discussed in the next section.

Arguably the *easiest* way to capture for 3D reconstruction is using hand-held cell phone video, since these cameras are so readily and widely available, and enable truly spontaneous, impromptu capture, as well as quickly covering large spaces. If we could achieve fully dense and accurate reconstruction from such input it would be immensely useful—however, this turns out to be quite difficult.

Besides the typical problems that any reconstruction system has to deal with, such as poorly textured areas, repetitive patterns, and occlusions, there are several additional challenges with video: higher noise level, shake and motion blur, rolling shutter deformations, small baseline between adjacent frames, and, often, the presence of dynamic objects, such as people. For these reasons, existing methods

often suffer from a variety of problems, such as missing regions in the depth maps (Figure 1b) and inconsistent geometry and flickering depth (Figure 1c).

Traditional reconstruction methods [Szeliski 2010] combine sparse structure-from-motion with dense multi-view stereo—essentially matching patches along epipolar lines. When the matches are correct, this results in geometrically accurate reconstructions. However, due to the before-mentioned complications, the matches are often noisy, and typically need to be regularized with heuristic smoothness priors. This often induces incorrect geometry in the affected regions, so that many methods drop pixels with low confidence altogether, leaving "holes" in the reconstruction (Figure 1b).

There has recently been immense progress on *learning-based* methods that operate on single images. These methods do not require heuristic regularization, but instead learn scene priors from data, which results in better ability to synthesize plausible depth in parts of the scene that would be weakly or even incorrectly constrained in traditional reconstruction approaches. They excel, in particular, at the reconstruction of dynamic scenes, since static and dynamic objects appear the same when we consider a single frame at a time. However, the estimated depth often flickers erratically due to the independent per-frame processing (Figure 1c), and it is not metric (i.e., not related to true depth by a single scale factor). This causes a video reconstruction to be *geometrically inconsistent*: objects appear to be attached to the camera and "swimming" in world-space.

Several video-based depth estimation methods have also been developed. These methods address the geometrical consistency of the reconstruction over time either implicitly via recurrent neural networks [Patil et al. 2020; Wang et al. 2019b] or explicitly using multi-view reconstruction [Liu et al. 2019; Teed and Deng 2020]. State-of-the-art video-based depth estimation methods [Liu et al. 2019; Teed and Deng 2020], however, handle only static scenes.

In this work, we present a new video-based reconstruction system that combines the strengths of traditional and learning-based techniques. It uses traditionally-obtained geometric constraints where they are available to achieve accurate and consistent depth, and leverages learning-based priors to fill in the weakly constrained parts of the scene more plausibly than prior heuristics. Technically, this is implemented by fine-tuning the weights of a single-image depth estimation network at test time, so that it learns to satisfy the geometry of a particular scene while retaining its ability to synthesize plausible new depth details where necessary. Our test-time training strategy allows us to use both short-term and long-term constraints and prevent drifting over time. The resulting depth videos are fully dense and detailed, with sharp object boundaries. The reconstruction is flicker-free and geometrically consistent throughout the video. For example, static objects appear rock-steady when projected into world space. The method even supports a gentle amount of dynamic scene motion, such as hand-waving (Figure 9), although it still breaks down for extreme object motion.

The improved quality and consistency of our depth videos enable interesting new applications, such as fully-automatic video special effects that interact with the dense scene content (Figure 9). We extensively evaluate our method quantitatively and show numerous qualitative results. The source code of our method is publicly available.[1]

2 RELATED WORK

Supervised monocular depth estimation. Early learning-based approaches regress local image features to depth [Saxena et al. 2008] or discrete geometric structures [Hoiem et al. 2005], followed by some post-processing steps (e.g., a MRF). Deep learning based models have been successfully applied to single image depth estimation [Eigen and Fergus 2015; Eigen et al. 2014; Fu et al. 2018; Laina et al. 2016; Liu et al. 2015]. However, training these models requires ground truth depth maps that are difficult to acquire. Several efforts have been made to address this issue, e.g., training on synthetic dataset [Mayer et al. 2016a] followed by domain adaptation [Atapour-Abarghouei and Breckon 2018], collecting relative depth annotations [Chen et al. 2016], using conventional structure-from-motion and multi-view stereo algorithms to obtain pseudo ground truth depth maps from Internet images [Chen et al. 2019a; Li et al. 2019; Li and Snavely 2018], or 3D movies [Ranftl et al. 2019; Wang et al. 2019a]. Our method builds upon recent advances in single image depth estimation and further improves the geometric consistency of the depth estimation on videos.

Self-supervised monocular depth estimation. Due to challenges of scaling up training data collection, self-supervised learning methods have received considerable attention for their ability to learn a monocular depth estimation model directly from raw stereo pairs [Godard et al. 2017] or monocular video [Zhou et al. 2017]. The core idea is to apply differentiable warp and minimize photometric reprojection error. Recent methods improve the performance through incorporating *coupled training* with optical flow [Ranjan et al. 2019; Yin and Shi 2018; Zou et al. 2018], object motion [Dai et al. 2019; Vijayanarasimhan et al. 2017], surface normal [Qi et al. 2018], edge [Yang et al. 2018], and visual odometry [Andraghetti et al. 2019; Shi et al. 2019; Wang et al. 2018b]. Other notable efforts include using stereo information [Guo et al. 2018; Watson et al. 2019], better network architecture and training loss design [Gordon et al. 2019; Guizilini et al. 2019], scale-consistent ego-motion network [Bian et al. 2019], incorporating 3D geometric constraints [Mahjourian et al. 2018], and learning from unknown camera intrinsics [Chen et al. 2019b; Gordon et al. 2019].

Many of these self-supervised methods use a *photometric* loss. However, these losses can be satisfied even if the geometry is not consistent (in particular, in poorly textured areas). In addition, they do not work well for temporally distant frames because of larger appearance changes. In our ablation study, however, we show that long-range temporal constraints are important for achieving good results.

Multi-view reconstruction. Multi-view stereo algorithms estimate scene depth using multiple images captured from arbitrary viewpoints [Furukawa et al. 2015; Schonberger and Frahm 2016; Seitz et al. 2006]. Recent learning-based methods [Huang et al. 2018; Im et al. 2019; Kusupati et al. 2019; Ummenhofer et al. 2017; Yao et al.

[1] https://roxanneluo.github.io/Consistent-Video-Depth-Estimation/

2018] leverage well-established principles in traditional geometry-based approaches (e.g., cost aggregation and plane-sweep volume) and show state-of-the-art performance in multi-view reconstruction. However, these multi-view stereo techniques assume a *static* scene. For dynamic objects, these methods either produce erroneous estimates or drop pixels with low confidence. In contrast, our method produces dense depth even in the presence of moderate dynamic scene motion.

Depth from video. Recovering dense depth from monocular video is a challenging problem. To handle moving objects, existing techniques rely on motion segmentation and explicit motion modeling for the moving objects in the scene [Casser et al. 2019; Karsch et al. 2014; Ranftl et al. 2016]. Several methods estimate depth by integrating motion estimation and multi-view reconstruction using two frames [Ummenhofer et al. 2017; Wang et al. 2019a] or a varying number of frames [Bloesch et al. 2018; Valentin et al. 2018; Zhou et al. 2018]. The state-of-the-art video-to-depth methods [Liu et al. 2019; Teed and Deng 2020] regress depth (or predict a distribution over depth) based on the cost volume constructed by warping nearby frames to a reference viewpoint. Such model designs thus do not account for dynamically moving objects. In contrast, while we also leverage constraints derived from multi-view geometry, our depth is estimated from (fine-tuned) *single-image* depth estimation models, and thereby handle dynamic object naturally and without the need for explicit motion segmentation.

Temporal consistency. Applying single-image based methods independently to each frame in a video often produce flickering results. In light of this, various approaches for enforcing temporal consistency have been developed in the context of style transfer [Chen et al. 2017; Huang et al. 2017; Ruder et al. 2016], image-based graphics applications [Lang et al. 2012], video-to-video synthesis [Wang et al. 2018a], or application-agnostic post-processing algorithms [Bonneel et al. 2015; Lai et al. 2018]. The core idea behind these methods is to introduce a "temporal consistency loss" (either at training or testing time) that encourages similar values along the temporal correspondences estimated from the input video. In the context of depth estimation from video, several efforts have been made to make the estimated depth more temporally consistent by explicitly applying optical flow-based consistency loss [Karsch et al. 2014] or implicitly encouraging temporal consistency using recurrent neural networks [Patil et al. 2020; Wang et al. 2019b; Zhang et al. 2019b]. Our work differs in that we aim to produce depth estimates from a video that are *geometrically* consistent. This is particularly important for casually captured videos because the actual depth may *not* be temporally consistent due to camera motion over time.

Depth-aware visual effects. Dense depth estimation facilitates a wide variety of visual effects such as synthetic depth-of-field [Wadhwa et al. 2018], novel view synthesis [Hedman et al. 2017; Hedman and Kopf 2018; Hedman et al. 2018; Shih et al. 2020], and occlusion-aware augmented reality [Holynski and Kopf 2018]. Our work on consistent depth estimation from causally captured videos enables several new video special effects.

Test-time training. Learning on testing data has been used in several different problem contexts: online update in visual tracking [Kalal et al. 2011; Ross et al. 2008], adapting object detectors from images to videos [Jain and Learned-Miller 2011; Tang et al. 2012], and learning video-specific features for person re-identification [Cinbis et al. 2011; Zhang et al. 2019a]. The work most closely related to ours is that of [Casser et al. 2019; Chen et al. 2019b] where they improve monocular depth estimation results by fine-tuning a pre-trained model using the testing video sequence. Note that any self-supervised method can be trained at test time (as in [Casser et al. 2019; Chen et al. 2019b]). However, the focus of previous methods is largely on achieving *per-frame accuracy*, while our focus is on achieving an accurate prediction with *global geometric consistency*. Our method achieves accurate and detailed reconstructions with a higher level of temporal smoothness than previous methods, which is important for many video-based applications.

Aside from these goals, there are important technical differences between our method and prior ones. The method in [Casser et al. 2019] performs a binary object-level segmentation and estimates rigid per-object transformations. This is appropriate for rigid objects such as cars in a street scene, but less so for highly deformable subjects such as people. The method in [Chen et al. 2019b] uses a geometric loss, similar to ours. However, they only train on consecutive frame pairs and relative poses. We use absolute poses and long-term temporal connections, which our ablation shows is critical for achieving good results (Figure 6).

3 OVERVIEW

Our method takes a monocular video as input and estimates a camera pose as well as a dense, *geometrically consistent* depth map (up to scale ambiguity) for each video frame. The term geometric consistency not only implies that the depth maps do not flicker over time but also, that all the depth maps are in mutual agreement. That is, we may project pixels via their depth and camera pose accurately amongst frames. For example, all observations of a static point should be mapped to a single common 3D point in the world coordinate system without drifting.

Casually captured input videos exhibit many characteristics that are challenging for depth reconstruction. Because they are often captured with a handheld, uncalibrated camera, the videos suffer from motion blur and rolling shutter deformations. The poor lighting conditions may cause increased noise level and additional blur. Finally, these videos usually contain dynamically moving objects, such as people and animals, thereby breaking the core assumption of many reconstruction systems designed for static scenes.

As we explained in the previous sections, in problematic parts of a scene, traditional reconstruction methods typically produce "holes" (or, if forced to return a result, estimate very noisy depth.) In areas where these methods are confident enough to return a result, however, it is typically fairly accurate and consistent, because they rely strongly on geometric constraints.

Recent learning-based methods [Liu et al. 2019; Ranftl et al. 2019] have complementary characteristics. These methods handle the challenges described above just fine because they leverage a strong data-driven prior to predict *plausible* depth maps from any input image. However, applying these methods independently for each

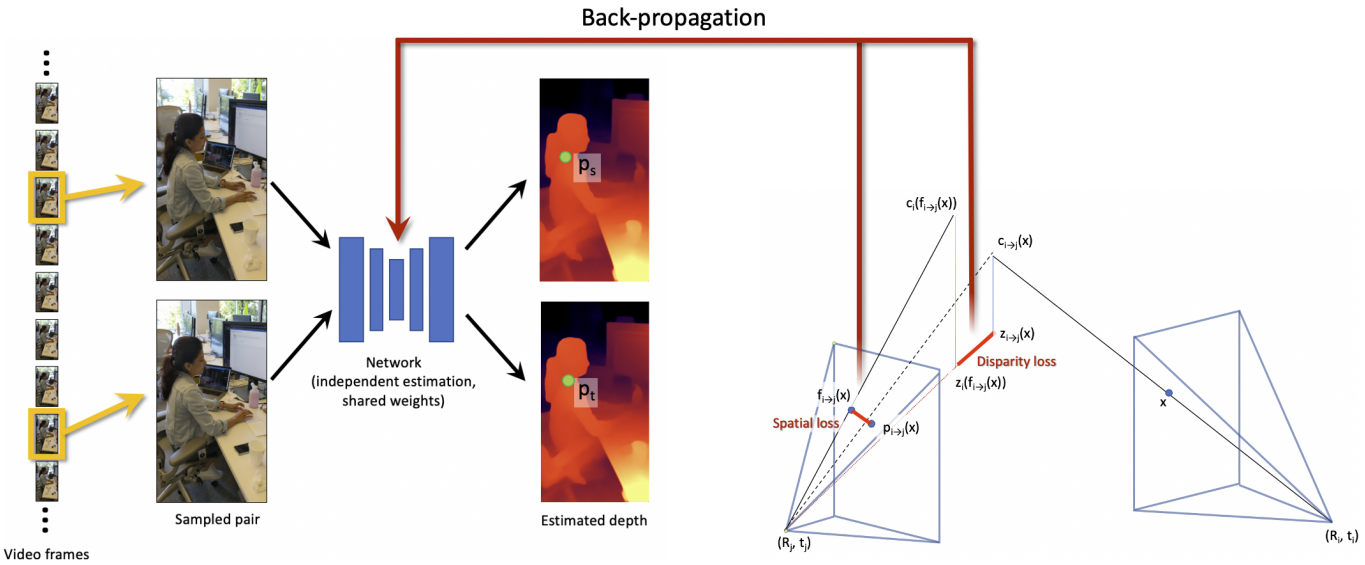


Fig. 2. Method overview. With a monocular video as input, we sample a pair of (potentially distant) frames and estimate the depth using a pre-trained, single-image depth estimation model to obtain initial depth maps. From the pair of images, we establish correspondences using optical flow with forward-backward consistency check. We then use these correspondences and the camera poses to extract geometric constraints in 3D. We decompose the 3D geometric constraints into two losses: 1) spatial loss and 2) disparity loss and use them to fine-tune the weight of the depth estimation network via standard backpropagation. This test-time training enforces the network to minimize the geometric inconsistency error across multiple frames for this particular video. After the fine-tuning stage, our final depth estimation results from the video is computed from the fine-tuned model.

frame results in geometrically inconsistent and temporally flickering results over time.

Our idea is to combine the strengths of both types of methods. We leverage existing single-image depth estimation networks [Godard et al. 2019; Li et al. 2019; Ranftl et al. 2019] that have been trained to synthesize plausible (but not consistent) depth for general color images, and we fine-tune the network using the extracted geometric constraints from a video using traditional reconstruction methods. The network thus learns to produce geometrically consistent depth on a particular video.

Our method proceeds in two stages:

Pre-processing (Section 4): As a foundation for extracting geometric constraints among video frames, we first perform a traditional Structure-from-Motion (SfM) reconstruction pipeline using an off-the-shelf open-source software COLMAP [Schonberger and Frahm 2016]. To improve pose estimation for videos with dynamic motion, we apply Mask R-CNN [He et al. 2017] to obtain people segmentation and remove these regions for more reliable keypoint extraction and matching, since people account for the majority of dynamic motion in our videos. This step provides us with accurate intrinsic and extrinsic camera parameters as well as a sparse point cloud reconstruction. We also estimate dense correspondence between pairs of frames using optical flow. The camera calibration and dense correspondence, together, enable us to formulate our geometric losses, as described below.

The second role of the SfM reconstruction is to provide us with the scale of the scene. Because our method works with monocular input, the reconstruction is ambiguous up to scale. The output of the learning-based depth estimation network is scale-invariant as well. Consequently, to limit the amount the network has to change, we adjust the scale of the SfM reconstruction so that it matches the learning-based method in a robust average sense.

Test-time Training (Section 5): In this stage, which comprises our primary contribution, we fine-tune a pre-trained depth estimation network so that it produces more geometrically consistent depth for a *particular* input video. In each iteration, we sample a pair of frames and estimate depth maps using the current network parameters (Figure 2). By comparing the dense correspondence with reprojections obtained using the current depth estimates, we can validate whether the depth maps are geometrically consistent. To this end, we evaluate two geometric losses, 1) spatial loss and 2) disparity loss and back-propagate the errors to update the network weights (which are shared across for all frames). Over time, iteratively sampling many frame pairs, the losses are driven down, and the network learns to estimate depth that is geometrically consistent for this video while retaining its ability to provide plausible regularization in less constrained parts.

The improvement is often dramatic, our final depth maps are geometrically consistent, temporally coherent across the entire video while accurately delineate clear occluding boundaries even for dynamically moving objects. With depth computed, we can have proper depth edge for occlusion effect and make the geometry of the real scene interact with the virtual objects. We show various compelling visual effects made possible by our video depth estimation in Section 6.5.

4 PRE-PROCESSING

Camera registration. We use the structure-from-motion and multi-view stereo reconstruction software COLMAP [Schonberger and Frahm 2016] to estimate for each frame i of the N video frames the

intrinsic camera parameters K_i, the extrinsic camera parameters (R_i, t_i), as well as a semi-dense depth map D_i^{MVS}. We set the values to zeros for pixels where the depth is not defined.

Because dynamic objects often cause errors in the reconstruction, we apply Mask R-CNN [He et al. 2017] to segment out people (the most common "dynamic objects" in our videos) in every frame independently, and suppress feature extraction in these areas (COLMAP provides this option). Since smartphone cameras are typically not distorted[2], we use the `SIMPLE_PINHOLE` camera model and solve for the shared camera intrinsics for all the frames, as this provides a faster and more robust reconstruction. We use the exhaustive matcher and enable guided matching.

Scale calibration. The scale of the SfM and the learning-based reconstructions typically do *not* match, because both methods are scale-invariant. This manifests in different value ranges of depth maps produced by both methods. To make the scales compatible with the geometric losses, we adjust the SfM scale, because we can simply do so by multiplying all camera translations by a factor.

Specifically, let D_i^{NN} be the initial depth map produced by the learning-based depth estimation method. We first compute the relative scale for image i as:

$$s_i = \underset{x}{\text{median}}\left\{ D_i^{NN}(x) \;/\; D_i^{MVS}(x) \;\middle|\; D_i^{MVS}(x) \neq 0 \right\}, \tag{1}$$

where $D(x)$ is the depth at pixel x.

We can then compute the global scale adjustment factor s as

$$s = \underset{i}{\text{mean}}\{s_i\}, \tag{2}$$

and update all the camera translations

$$\tilde{t_i} = s \cdot t_i. \tag{3}$$

Frame sampling. In the next step, we compute a dense optical flow for certain pairs of frames. This step would be prohibitively computationally expensive to perform for all $O(N^2)$ pairs of frames in the video. We, therefore, use a simple hierarchical scheme to prune the set of frame pairs down to $O(N)$.

The first level of the hierarchy contains all consecutive frame pairs,

$$S_0 = \left\{ \; (i, j) \;\middle|\; |i - j| = 1 \; \right\}. \tag{4}$$

Higher levels contain a progressively sparser sampling of frames,

$$S_l = \left\{ \; (i, j) \;\middle|\; |i - j| = 2^l, \; i \bmod 2^{l-1} = 0 \; \right\}. \tag{5}$$

The final set of sampled frames is the union of the pairs from all levels,

$$S = \bigcup_{0 \leq l \leq \lfloor \log_2(N-1) \rfloor} S_l. \tag{6}$$

Optical flow estimation. For all frame pairs (i, j) in S we need to compute a dense optical flow field $F_{i \to j}$. Because flow estimation works best when the frame pairs align as much as possible, we first align the (potentially distant) frames using a homography-warp (computed with a RANSAC-based fitting method [Szeliski 2010]) to eliminate dominant motion between the two frames (e.g., due to camera rotation). We then use FlowNet2 [Ilg et al. 2017] to compute the optical flow between the aligned frames. To account of moving objects and occlusion/dis-occlusion (as they do not satisfy the geometric contraints or are unreliable), we apply a forward-backward consistency check and remove pixels that have forward-backward errors larger than 1 pixel, producing a binary map $M_{i \to j}$. Furthermore, we observe that the flow estimation results are not reliable for frame pairs with little *overlap*. We thus exclude any frame pairs where $|M_{i \to j}|$ is less than 20% of the image area from consideration.

[2]Our test sequences (Section 6.1) are captured with a fisheye camera, and we remove the distortion through rectification.

5 TEST-TIME TRAINING ON INPUT VIDEO

Now we are ready to describe our test-time training procedure, i.e., how we coerce the depth network through fine-tuning it with a geometric consistency loss to producing more consistent depth for a particular input video. We first describe our geometric loss, and then the overall optimization procedure.

Geometric loss. For a given frame pair $(i, j) \in S$, the optical flow field $F_{i \to j}$ describes which pixel pairs show the same scene point. We can use the flow to test the geometric consistency of our current depth estimates: if the flow is correct and a flow-displaced point $f_{i \to j}(x)$ is identical to the depth-reprojected point $p_{i \to j}(x)$ (both terms defined below), then the depth *must* be consistent.

The idea of our method is that we can turn this into a geometric loss $\mathcal{L}_{i \to j}$ and back-propagate any consistency errors through the network, so as to coerce it into to producing depth that is *more* consistent than before. $\mathcal{L}_{i \to j}$ comprises two terms, an image-space loss $\mathcal{L}_{i \to j}^{spatial}$, and a disparity loss $\mathcal{L}_{i \to j}^{disparity}$. To define them, we first discuss some notation.

Let x be a 2D pixel coordinate in frame i. The flow-displaced point is simply

$$f_{i \to j}(x) = x + F_{i \to j}(x). \tag{7}$$

To compute the depth-reprojected point $p_{i \to j}(x)$, we first lift the 2D coordinate to a 3D point $c_i(x)$ in frame i's camera coordinate system, using the camera intrinsics K_i as well as the current depth estimate D_i,

$$c_i(x) = D_i(x)\, K_i^{-1} \tilde{x}, \tag{8}$$

where $\tilde{x}$ is the homogeneous augmentation of x. We then further project the point to the other frame j's camera coordinate system,

$$c_{i \to j}(x) = R_j^{\mathsf{T}} \Big(R_i\, c_i(x) + \tilde{t_i} - \tilde{t_j} \Big), \tag{9}$$

and finally convert it back to a pixel position in frame j,

$$p_{i \to j}(x) = \pi\big(K_j\, c_{i \to j}(x) \big), \tag{10}$$

where $\pi\big([x, y, z]^{\mathsf{T}}\big) = [\frac{x}{z}, \frac{y}{z}]^{\mathsf{T}}$.

With this notation, the image-space loss for a pixel can be easily defined:

$$\mathcal{L}_{i \to j}^{spatial}(x) = \left\| p_{i \to j}(x) - f_{i \to j}(x) \right\|_2, \tag{11}$$

which penalizes the image-space distance between the flow-displaced and the depth-reprojected point.

The disparity loss, similarly, penalizes the disparity distance in camera coordinate system:

$$\mathcal{L}_{i \to j}^{disparity}(x) = u_i \left| z_{i \to j}^{-1}(x) - z_j^{-1}\big(f_{i \to j}(x)\big) \right|, \tag{12}$$

where u_i is frame i's focal length, and z_i and $z_{i\to j}$ are the scalar z-component from c_i and $c_{i\to j}$, respectively.

The overall loss for the pair is then simply a combination of both losses for all pixels where the flow is valid,

$$\mathcal{L}_{i\to j} = \frac{1}{|M_{i\to j}|} \sum_{x\in M_{i\to j}} \mathcal{L}^{spatial}_{i\to j}(x) + \lambda \mathcal{L}^{disparity}_{i\to j}(x), \qquad (13)$$

where $\lambda = 0.1$ is a balancing coefficient.

Discussion. While the second term in Equation 11 (flow mapping) can handle dynamic motion, the first term (depth reprojection) assumes a *static* scene. How can this still result in an accurate depth estimation? There are two cases: (1) Consistent motion (e.g., a moving car) can sometimes be aligned with the epipolar geometry and cause our method, like most others, to estimate the wrong depth. (2) Consistent motion that is *not* epipolar-aligned or inconsistent motion (e.g., a waving hand) causes conflicting constraints; empirically, our test-time training is tolerant to these conflicting constraints and produces accurate results (as seen in many examples in this submission and accompanying materials.)

Optimization. Using the geometric loss between i-th and j-th frames $\mathcal{L}_{i\to j}$, we fine-tune the network weights using standard backpropagation. Initializing the network parameters using a pre-trained depth estimation model allows us to transfer the knowledge for producing plausible depth maps on images that are challenging for traditional geometry-based reconstructon systems. We fine-tune the network using a fixed number of epochs (20 epochs for all our experiments). In practice, we find that with this simple fine-tuning step the network training does not overfit the data in the sense that it does not lose its ability to synthesize plausible depth in unconstrained or weakly constrained parts of the scene[3]. We also observe that the training handles a certain amount of erroneous supervision (e.g., when the correspondences are incorrectly established).

Implementation details. We have experimented with several monocular depth estimation architectures and pre-trained weights [Godard et al. 2019; Li et al. 2019; Ranftl et al. 2019]. If not otherwise noted, results in the paper and accompanying materials use Li et al.'s network [2019] (single-image model). We use the other networks in evaluations as noted there. Given an input video, an epoch is defined by one pass over all frame pairs in $\mathcal{S}$. In all of our experiments, we fine-tune the network for 20 epochs with a batch size of 4 and a learning rate of 0.0004 using ADAM optimizer [Kingma and Ba 2015]. The time for test-time training varies for videos of different lengths. For a video of 244 frames, training on 4 NVIDIA Tesla M40 GPUs takes 40 min.

6 RESULTS AND EVALUATION

In this section, we first describe the experimental setup (Section 6.1). We then present quantitative comparison with the state-of-the-art depth estimation methods (Section 6.2). We conduct an extensive ablation study to validate the importance of our design choices and their contributions to the results (Section 6.3). Finally, we show qualitative results of our depth estimation and their applications to new advanced video-based visual effects (Section 6.5).

[3]We note that there are more advanced regularization techniques for transfer learning [Kirkpatrick et al. 2017; Li et al. 2018]. These can be applied to further improve the performance of our method.

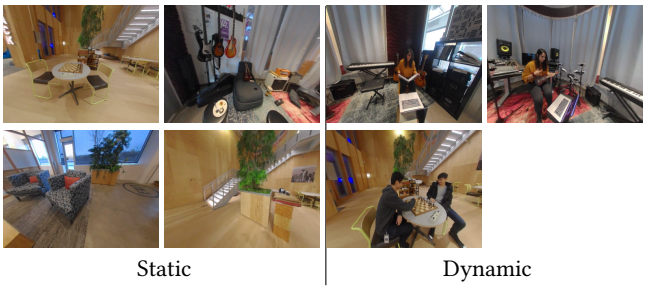


Fig. 3. Example frames from our test set that includes four static sequences and three dynamic ones. The dynamic videos contain gentle amount of seated motion like playing ukulele and body motion while playing chess, flipping the notes while singing, etc. These videos resemble casual video capture scenario where the hand-held camera is shaky and frames contain motion blur.

6.1 Experimental Setup

Dataset. Many datasets have been constructed for evaluating depth reconstruction. However, these existing datasets are either for synthetic [Butler et al. 2012; Mayer et al. 2016a] , specific domains (e.g., driving scenes) [Geiger et al. 2013], single images [Chen et al. 2016; Li et al. 2019; Li and Snavely 2018], or videos (or multiple images) of static scenes [Schops et al. 2017; Silberman et al. 2012; Sturm et al. 2012a]. Consequently, we capture custom stereo video datasets for evaluation. Our test set consists of both static and dynamic scenes with a gentle amount of object motion (see Fig. 3 for samples. We capture the videos with stereo fisheye QooCam cameras.[4] The handheld camera rig provides a handy way to capture stereo videos, but it is highly distorted in the periphery due to the fisheye lenses. We, therefore, rectify and crop the center region using the Qoocam Studio[5] and obtain videos of resolution 1920×1440 pixels. The lengths of the captured video range from 119 to 359 frames. Our new video dataset is available on the accompanying website for evaluating future video-based depth estimation.

For completeness, we also provide quantitative comparisons with the state-of-the-art depth estimation models on three publicly available datasets: (1) the TUM dataset [Sturm et al. 2012b] (using the 3D Object Reconstruction category), (2) the ScanNet dataset [Dai et al. 2017] (using the testing split provided by [Teed and Deng 2020]), and (3) the KITTI 2015 dataset [Geiger et al. 2012] (using the Eigen split [Eigen et al. 2014]).

Evaluation metrics. To evaluate and compare the quality of the estimated depth from a monocular video on our custom stereo video dataset, we use the following three different metrics.

Photometric error E_p: We use photometric error to quantify the accuracy of the recovered depth. All the methods estimate the depth from the *left* video stream. Using the estimated depth, we then reproject the pixels from the left video stream to the right one and compute the photometric error as mean squared error of the RGB differences. As the depth map can only be estimated up to a

[4]https://www.kandaovr.com/qoocam/
[5]https://www.kandaovr.com/download/

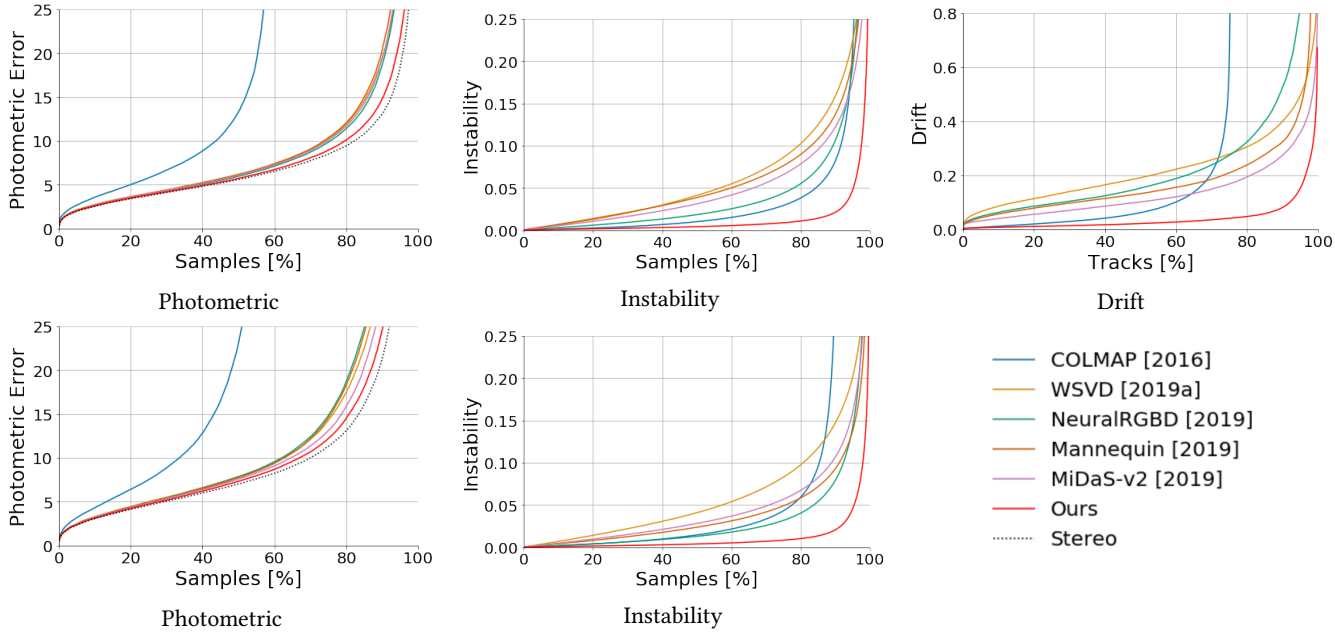


Fig. 4. Quantitative comparison with the state-of-the-art. We plot the error of all reconstructed pixels, sorted by error. Note, that COLMAP drops some pixels from the reconstruction. Hence, its curve in the left column stops short of 100%; the second and third column evaluate on tracks, which tend to be in textured areas where COLMAP has a higher level of completeness. Top: static sequences; bottom: Dynamic sequences.

Table 1. Quantitative comparisons with the state-of-the-art depth estimation algorithms.

	Static			Dynamic	
	E_s (%) ↓	E_d (%) ↓	E_p ↓	E_s (%) ↓	E_p ↓
WSVD [2019a]	4.13	19.12	11.90	4.10	17.46
NeuralRGBD [2019]	1.86	15.25	11.33	1.30	18.62
Mannequin [2019]	3.88	13.22	12.05	2.38	18.16
MiDaS-v2 [2019]	3.14	10.14	11.74	2.83	15.76
COLMAP [2016]	1.02	6.19	-	1.47	-
Ours	**0.44**	**2.12**	**10.09**	**0.40**	**14.44**

scale ambiguity, we need to align the estimated depth maps to the stereo disparity. Specifically, we compute the stereo disparity by taking the horizontal components from the estimated flow on the stereo pair (using Flownet2 [Ilg et al. 2017]). For each video frame, we then compute the *scale* and *shift* alignment to the computed stereo disparity using RANSAC-based linear regression. We can obtain the global (video-level) scale and shift parameters by taking the mean of the scales/shifts for all the frames.

Instability E_s: We measure instability of the estimated depth maps over time in a video as follows. We first extract a sparse set of reliable tracks from the input monocular video using a standard KLT tracker. We then convert the 2D tracks to 3D tracks, using the camera poses and calibrated depths to unproject 2D tracks to 3D. For a perfectly stable reconstruction, each 3D track should collapse to a single 3D point. We thus can quantify the instability by computing the Euclidean distances of the 3D points for each pair of consecutive frames.

Drift E_d: In many cases, while 3Dtracks described above may appear somewhat stable for consecutive frames, the errors could be accumulated and cause *drift* over time. To measure the amount of drift for a particular 3D track, we compute the maximum eigenvalue of the covariance matrix formed by the 3D track. Intuitively, this measures how *spread* the 3D points is across time.

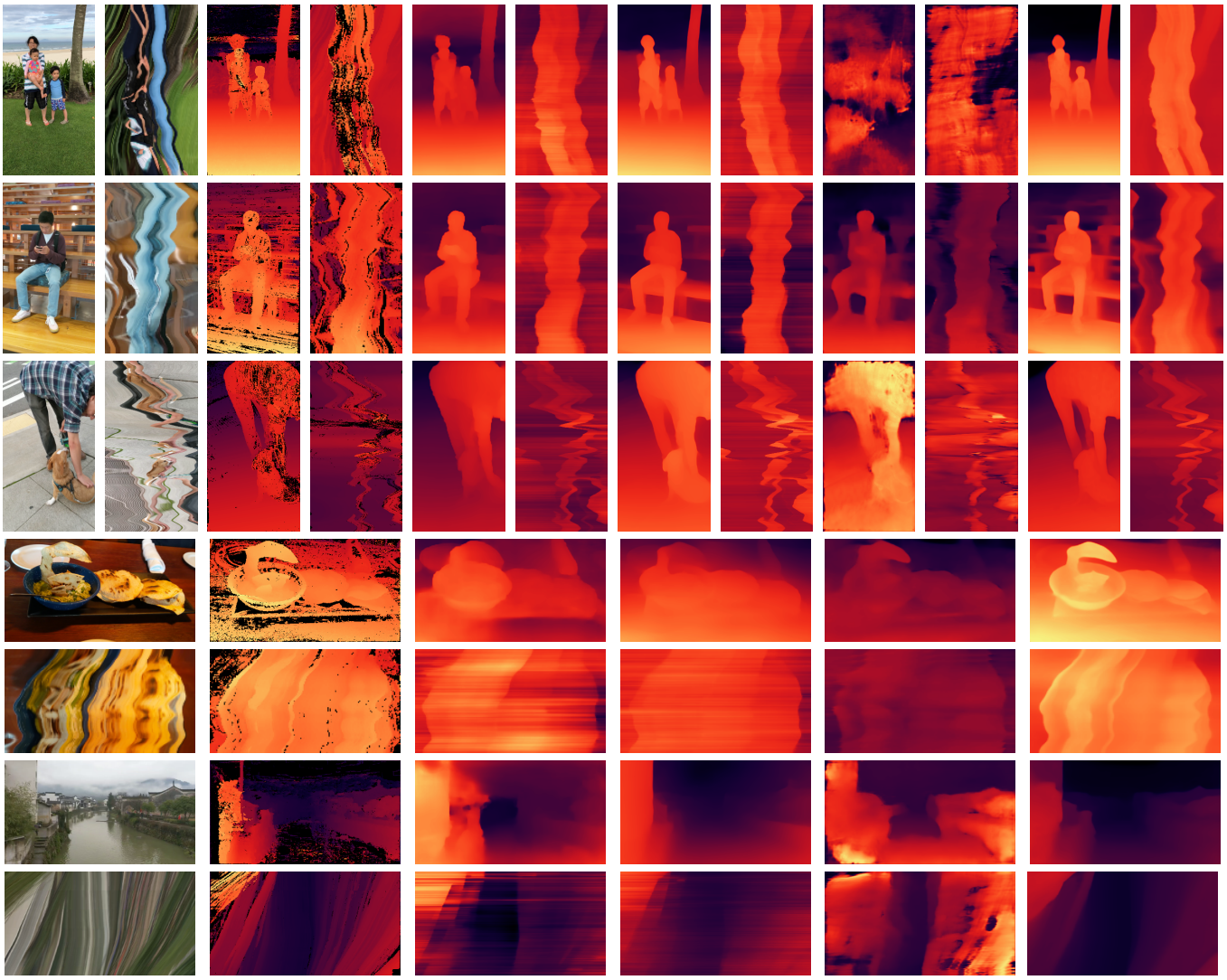
For static sequences, we evaluate the estimated depth using all three metrics. For dynamic sequences, we evaluate only on photometric error and instability, as the drift metric does not account for dynamically moving objects in the scene.

6.2 Comparative Evaluation

Compared methods. We compare our results with state-of-the-art depth estimation algorithms from three main categories.

- **Traditional multi-view stereo system**: COLMAP [Schonberger and Frahm 2016].
- **Single-image depth estimation**: Mannequin Challenge [Li et al. 2019] and MiDaS-v2 [Ranftl et al. 2019].
- **Video-based depth estimation**: WSVD [Wang et al. 2019a] (two frames) and NeuralRGBD [Liu et al. 2019] (multiple frames).

Quantitative comparison. Fig. 4 shows the plot of the photometric error, instability, and drift metrics against completeness. In all three metrics, our method compares favorably against previously published algorithms. Our results particularly shine when evaluated on the instability and the drift metrics, highlighting the *consistency* of our results. Table 1 further reports the summary of the results for different methods.



(a) Input (b) COLMAP (c)Mannequin (d) MiDaS-v2 (e) NeuralRGBD (f) Ours

Fig. 5. Visual comparisons with the state-of-the-arts. Our method produces depth, geometrically consistent, and flicker-free depth estimation from casually captured videos by a hand-held cellphone camera. The first image in each pair is a sample frame, while the second is a scanline slice through the spatio-temporal volume (either color video or video depth).

Visual comparison. We present in Fig. 5 the qualitative comparison of different depth estimation methods. The traditional multi-view stereo method produces accurate depths at highly textured regions, where reliable matches can be established. These depth maps contain large holes (black pixels), as shown in Fig. 5b. The learning-based single-image depth estimation approaches [Li et al. 2019; Ranftl et al. 2019] produce dense, plausible depth maps for each individual video frame. However, flickering depths over time cause geometrically inconsistent depth reconstructions. Video-based methods such as NeuralRGBD alleviate the temporal flicker, yet suffer from drift due to the limited temporal window used for depth estimation. We refer the readers to the video results in the supplementary material.

Table 2. Ablation study. The quantitative evaluation highlights the importance of our method design choices.

	$E_s(\%) \downarrow$	$E_d(\%) \downarrow$	$E_p \downarrow$
Ours w/o scale calibration	0.93	3.37	**9.99**
Ours w/o disparity loss	0.76	3.30	**9.99**
Ours w/o overlap test	0.51	2.49	13.20
Ours	**0.44**	**2.12**	10.08

6.3 Ablation Study

We conduct an ablation study to validate the effectiveness of several design choices in our approaches. We first study the effect of losses

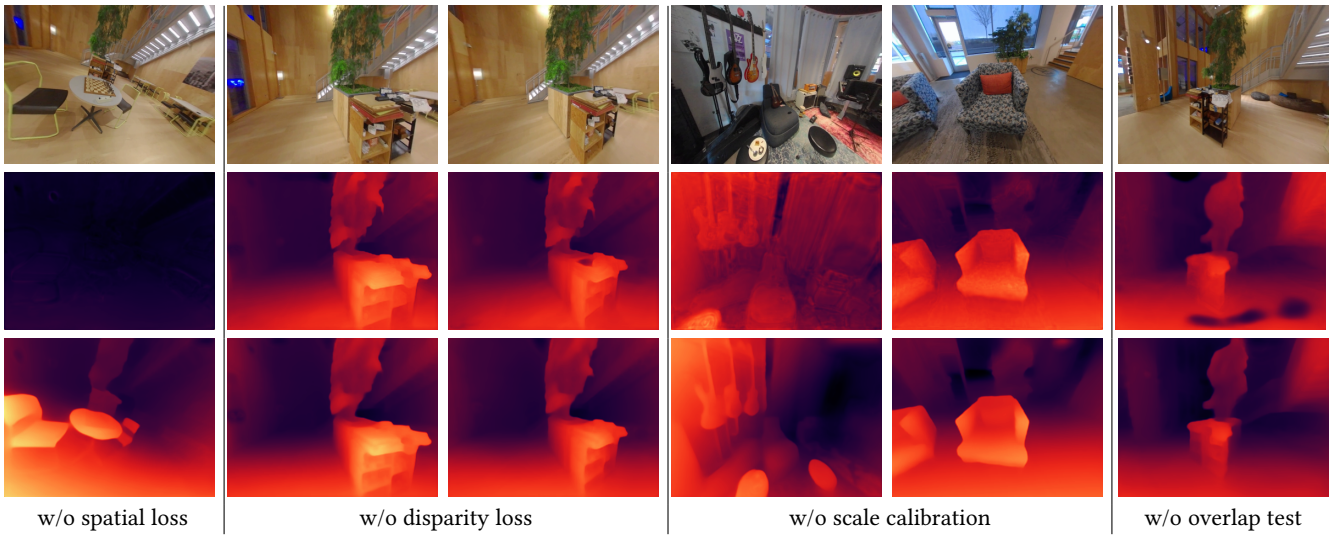


Fig. 6. Contribution of our design choices to the results. (*Top*): Sample frame from input videos. (*Middle*): corresponding ablation results. Bottom: result with full pipeline. Without spatial loss, there is no constraint for what the depth should be. We end up losing all the structure and it fails. Without disparity loss, depth can get sharper but also more flicker. Without scale calibration, we often observe degraded depth with blurrier depth discontinuities. Without overlap test, erroneous flow causes wrong depth.

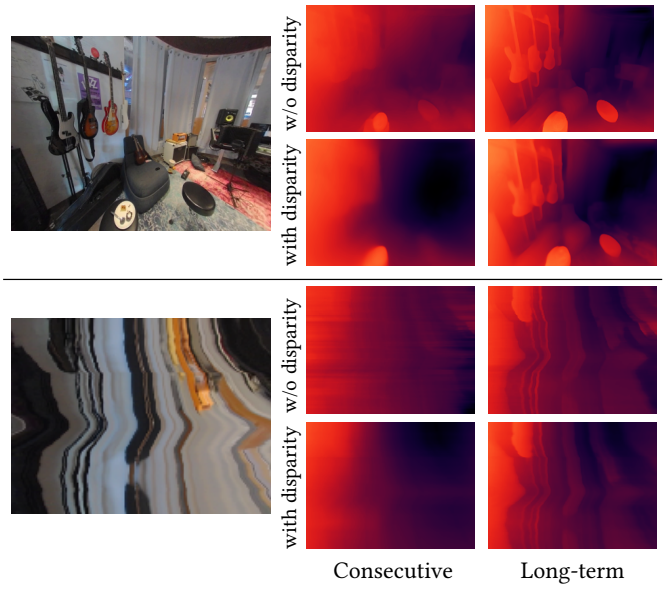


Fig. 7. Analysis of the effects of using long-term temporal constraints and the disparity loss. Please see the supplementary for video comparisons.

and the importance of different steps in the pipeline, including scale calibration and overlap test. We summarize these results in Table 2. Fig. 6 visualizes the contributions of various components.

We observe that using long-term constraints help improve the stability of the estimated depth over time. As the disparity loss also helps reduce temporal flickering, we further investigate the effects of both design choices in Fig. 7. Our results show that including constraints from long-term frame pairs leads to sharper and temporally more stable results. In contrast, while adding disparity loss reduces temporal flickers, it produces blurry results when using only consecutive frame pairs.

6.4 Quantitative Comparisons on Public Benchmarks

We provide quantitative results on three publicly available benchmark datasets for evaluating the performance of our depth estimation. In all of the evaluation settings, we resize the input images so that the longest image dimension to 384. We finetune the monocular depth estimation network for 20 epochs (the same evaluation setting used in the stereo video dataset).

TUM-RGBD dataset. We evaluate our method on the 11 scenes in the "3D Object Reconstruction" category in the TUM-RGBD dataset [Sturm et al. 2012b]. For evaluation, we subsample the videos every 5 frames and obtain sequences ranging from 195 to 593 frames. Here, we use the ground truth camera pose provided by the dataset. We then fine-tune the single-image model from Li et al. [2019] on the subsampled frames. To compute the error metrics, we align the predicted depth map to the ground truth using per-image median scaling. We report the errors in the disparity (inverse depth) space as it does not require clipping any depth ranges.

Table 4 reports the quantitative comparisons with single-frame methods [Li et al. 2019; Ranftl et al. 2019] and multi-frame methods [Liu et al. 2019; Wang et al. 2019a]. Our approach performs favorably against prior methods with a large margin in all evaluation metrics. In particular, our proposed test-time training significantly improves the performance over the baseline model from Li et al. [2019].

ScanNet dataset. Following the evaluation protocol of Teed and Deng [2020], we evaluate our method on the 2,000 sampled frames from the 90 test scenes in the ScanNet dataset [Dai et al. 2017]. We finetune the MiDaS-v2 model [Ranftl et al. 2019] on each testing sequence with a learning rate of 10^{-5} and $\lambda = 10^{-5}$. Following Teed

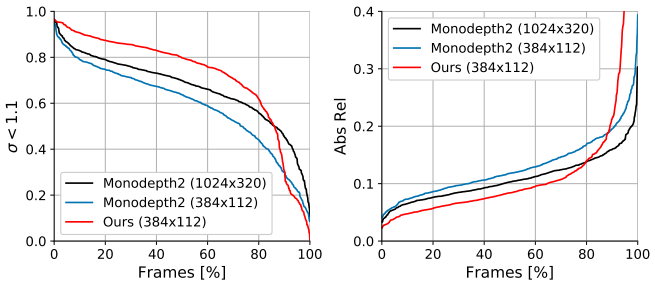


Fig. 8. Quantitative comparison before and after fine-tuning Monodepth2 on KITTI. We plot the each metric over all the test frames sorted by their values. After fine-tuning, we have more outliers due to extreme dynamic motion or failure in camera pose estimation, but achieve improved results for more than 80% of the frames.

and Deng [2020], we apply per-image median scaling to align the predicted depth to the ground truth depth map (using the range of [0.5, 8] meters). We then evaluate all the metrics in the depth space over the regions where the ground truth depth is within the range of 0.1 to 10 meters.

Table 3 shows the quantitative comparisons with several other multi-frame based depth estimation methods [Tang and Tan 2019; Teed and Deng 2020; Ummenhofer et al. 2017] and the baseline single-image model [Ranftl et al. 2019]. Our method achieves competitive performance with the state-of-the-art algorithms, performing slightly inferior to the DeepV2D method that is trained on the ScanNet training set.

KITTI dataset. We evaluate our method on the KITTI dataset [Geiger et al. 2012] using the Eigen test split [Eigen et al. 2014] for comparison with prior monocular depth estimation methods. We estimate the camera poses for each test sequence using COLMAP [Schonberger and Frahm 2016]. We observe that the FlowNet2 model (pre-trained on the Flying Chairs [Dosovitskiy et al. 2015] and Flying Things 3D [Mayer et al. 2016b] datasets) performs poorly in the KITTI dataset [Geiger et al. 2012]. Consequently, we use the FlowNet2 model finetuned on a combination of the KITTI2012 [Geiger et al. 2012] and KITTI2015 [Menze et al. 2015] training sets, FlowNet2-ft-kitti, for extracting dense correspondence across frames. Due to the challenging large forward motion in the driving videos, the flow estimations between temporally distant frames are not accurate. We thus only sample pairs with more than 50% consistent flow matches. We use the Monodepth2 [Godard et al. 2019] as the base single-image depth estimation network. We apply our fine-tuning method at the resolution of 384×112 over each sequence with a learning rate of 4×10^{-5} and $\lambda = 1$. Following the standard protocol [Godard et al. 2017], we cap the depth to 80m and report the results using the per-image median ground truth scaling alignment.

Table 5 presents the quantitative comparisons with the state-of-the-art monocular depth estimation methods. Under this evaluation setting, the results appear to show that our method does not provide an overall improvement over the baseline model Monodepth2 [Godard et al. 2019]. To investigate this issue, we show in Figure 8 the sorted error (Abs Rel) and accuracy ($\sigma < 1.1$) metrics for all the testing frames. The results show that our method indeed improves the performance in more than 80% of the testing frames (even when compared with the model with a high-resolution outputs). However, as COLMAP produce erroneous pose estimates in sequences with large dynamic objects in the scene, our fine-tuning method inevitably results in depth estimation with very large errors. Our method also has difficulty in handling significant dynamic scene motion. As a result, our method does not achieve clear improvement when the results are averaged over all the testing frames. Please see the supplementary material for video result comparison.

Table 3. Quantitative comparison on the ScanNet dataset [Dai et al. 2017] using the test split provided by Tang and Tan [2019].

	Error metric ↓				
	Abs Rel	Sq Rel	RMSE	RMSE log	Sc Inv
DeMoN [2017]	0.231	0.520	0.761	0.289	0.284
BA-Net [2019]	0.161	0.092	0.346	0.214	0.184
DeepV2D (NYU) [2020]	0.080	0.018	0.223	0.109	0.105
DeepV2D (ScanNet) [2020]	**0.057**	**0.010**	**0.168**	**0.080**	**0.077**
MiDaS-v2 [2019]	0.208	0.318	0.742	0.246	0.239
Ours	0.073	0.037	0.217	0.105	0.103

6.5 Video-based Visual Effects

Consistent video depth estimation enables interesting video-based special effects. Fig. 9 showcases samples of these effects. Full video results can be found in the supplementary material.

6.6 Limitations

There are several limitations and drawbacks of the proposed video depth estimation method.

Poses Our method currently relies on COLMAP [Schonberger and Frahm 2016] to estimate the camera pose from a monocular video. In challenging scenarios, e.g., limited camera translation and motion blur, however, COLMAP may not be able produce reliable sparse reconstruction and camera pose estimation. Large pose errors have a strong degrading effect on our results. This limits the applicability of our method on such videos. Integrating learning-based pose estimation (e.g., as in [Liu et al. 2019; Teed and Deng 2020]) with our approach is an interesting future direction.

Dynamic motion Our method supports videos containing moderate object motion. It breaks for extreme object motion.

Flow We rely on FlowNet2 [Ilg et al. 2017] to establish geometric constraints. Unreliable flow is filtered through forward-backward consistency checks, but it might be by chance erroneous in a consistent way. In this case our method will fail to produce correct depth. We tried using sparse flow (subsampling dense flow on a regular grid), but it did not perform well.

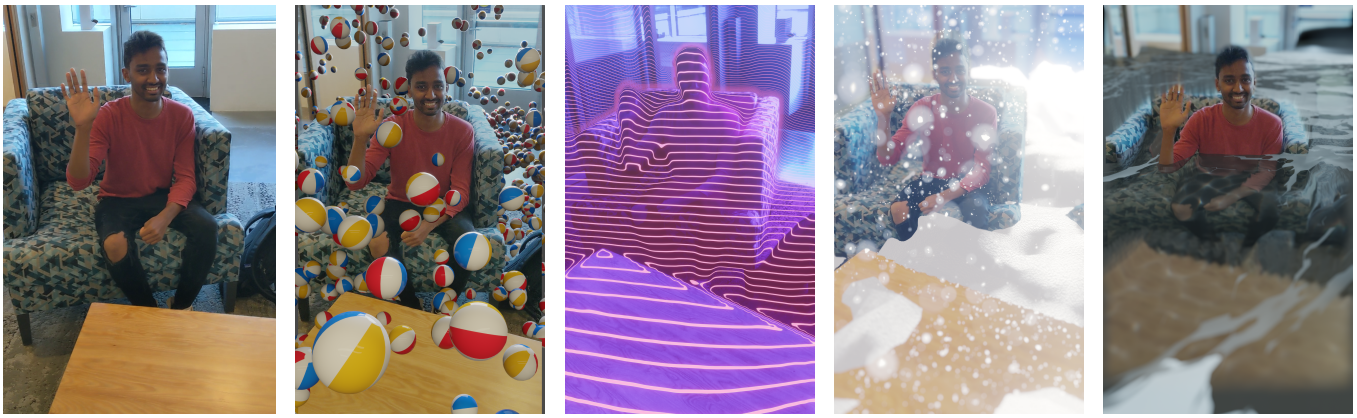
Speed As we extract geometric constraints using all the frames in a video, we do not support online processing. For example, our test-time training step takes about 40 minutes for a video of 244 frames and 708 sampled flow pairs. Developing online and fast variants in the future will be important for practical applications.

Table 4. Quantitative comparison on the TUM-RGBD dataset (3D Object Reconstruction category) [Sturm et al. 2012b] in the disparity space. We report the averaged results over 11 video sequences.

		Error metric ↓				Accuracy metric ↑		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Single-frame	Mannequin [2019]	0.306	0.101	0.244	0.385	0.569	0.772	0.885
	MiDaS-v2 [2019]	0.220	0.061	0.187	0.292	0.665	0.861	0.945
Multi-frame	WSVD [2019a]	0.281	0.083	0.228	0.365	0.551	0.794	0.905
	NeuralRGBD [2019]	0.615	0.365	0.392	0.661	0.361	0.571	0.710
	Ours	**0.144**	**0.036**	**0.144**	**0.211**	**0.785**	**0.934**	**0.979**

Table 5. Quantitative comparisons with existing methods on the KITTI benchmark dataset using the Eigne split. (*Top*): methods that produce full resolution (1024×320) depth maps. (*Bottom*): methods that produce low-resolution (384×112) depth maps. Note that for fair comparison, we align the depth results from all the compared methods with per-image median ground truth scaling. Therefore, our reported numbers for Monodepth2 (1024×320) [2019] differ slightly from those in their paper where they use a constant scale for alignment.

	Error metric ↓				Accuracy metric ↑		
	Abs Rel	Sq Rel	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Zhou [2017]	0.183	1.595	6.709	0.270	0.734	0.902	0.959
GeoNet [2018]	0.149	1.060	5.567	0.226	0.796	0.935	0.975
DF-Net [2018]	0.150	1.124	5.507	0.223	0.806	0.933	0.973
Struct2depth [2019]	0.109	0.825	4.750	0.187	0.874	0.958	**0.983**
GLNet [2019b]	**0.099**	**0.796**	4.743	**0.186**	0.884	0.955	0.979
Monodepth2 (1024×320) [2019]	0.108	0.806	**4.606**	0.187	**0.887**	**0.962**	0.981
Monodepth2 (384×112) [2019]	0.128	1.040	5.216	0.207	0.849	0.951	0.978
Ours (384×112)	0.130	2.086	4.876	0.205	0.878	0.946	0.970



Input video | Bouncing balls | Disco | Snow | Water

Fig. 9. Our consistent depth estimation enables a wide range of fully-automated video-based visual effects. We refer the readers to the supplementary video.

7 CONCLUSIONS

We have presented a simple yet effective method for estimating *consistent* depth from a monocular video. Our idea is to leverage geometric constraints extracted using conventional multi-view reconstruction methods and use them to fine-tune a pre-trained single-image depth estimation network. Using our test-time fine-tuning strategy, our network learns to produce geometrically consistent depth estimates across entire video. We conduct extensive quantitative and qualitative evaluation. Our results show that our method compares favorably against several state-of-the-art depth estimation algorithms. Our consistent video depth estimation enables compelling video-based visual effects.

ACKNOWLEDGMENTS

We would like to thank Patricio Gonzales Vivo, Dionisio Blanco, and Ocean Quigley for creating the artistic effects in the accompanying video. We also thank True Price for his practical and insightful advice on reconstruction and Ayush Saraf for his suggestions in engineering.

REFERENCES

Lorenzo Andraghetti, Panteleimon Myriokefalitakis, Pier Luigi Dovesi, Belen Luque, Matteo Poggi, Alessandro Pieropan, and Stefano Mattoccia. 2019. Enhancing self-supervised monocular depth estimation with traditional visual odometry. In *International Conference on 3D Vision (3DV)*.

Amir Atapour-Abarghouei and Toby P Breckon. 2018. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. 2019. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. 2018. CodeSLAM—learning a compact, optimisable representation for dense visual SLAM. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. 2015. Blind video temporal consistency. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 196.

Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. 2012. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*.

Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. 2019. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. 2017. Coherent online video style transfer. In *International Conference on Computer Vision*.

Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. 2016. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Weifeng Chen, Shengyi Qian, and Jia Deng. 2019a. Learning single-image depth from videos using quality assessment networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. 2019b. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *International Conference on Computer Vision*.

Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. 2011. Unsupervised metric learning for face identification in TV video. In *International Conference on Computer Vision*.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qi Dai, Vaishakh Patil, Simon Hecker, Dengxin Dai, Luc Van Gool, and Konrad Schindler. 2019. Self-supervised Object Motion and Depth Estimation from Video. *arXiv preprint arXiv:1912.04250* (2019).

A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox. 2015. FlowNet: Learning Optical Flow with Convolutional Networks. In *IEEE International Conference on Computer Vision (ICCV)*. http://lmb.informatik.uni-freiburg.de/Publications/2015/DFIB15

David Eigen and Rob Fergus. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *International Conference on Computer Vision*.

David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. 2018. Deep ordinal regression network for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yasutaka Furukawa, Carlos Hernández, et al. 2015. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision* 9, 1-2 (2015), 1–148.

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

Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. 2017. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. 2019. Digging into self-supervised monocular depth estimation. In *International Conference on Computer Vision*.

Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. 2019. Depth from Videos in the Wild: Unsupervised Monocular Depth Learning from Unknown Cameras. In *International Conference on Computer Vision*.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, and Adrien Gaidon. 2019. PackNet-SfM: 3D Packing for Self-Supervised Monocular Depth Estimation. *arXiv preprint arXiv:1905.02693* (2019).

Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. 2018. Learning monocular depth by distilling cross-domain stereo networks. In *European Conference on Computer Vision (ECCV)*.

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. 2017. Mask R-CNN. In *International Conference on Computer Vision*.

Peter Hedman, Suhib Alsisan, Richard Szeliski, and Johannes Kopf. 2017. Casual 3D photography. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 234.

Peter Hedman and Johannes Kopf. 2018. Instant 3d photography. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 101.

Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. 2018. Deep blending for free-viewpoint image-based rendering. In *ACM Transactions on Graphics (TOG)*.

Derek Hoiem, Alexei A Efros, and Martial Hebert. 2005. Geometric context from a single image. In *International Conference on Computer Vision*.

Aleksander Holynski and Johannes Kopf. 2018. Fast depth densification for occlusion-aware augmented reality. In *ACM Transactions on Graphics (TOG)*. ACM, 194.

Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. 2017. Real-time neural style transfer for videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. 2018. Deepmvs: Learning multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. 2019. DPSNet: end-to-end deep plane sweep stereo. In *International Conference on Learning Representations (ICLR)*.

Vidit Jain and Erik Learned-Miller. 2011. Online domain adaptation of a pre-trained cascade of classifiers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. 2011. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 7 (2011), 1409–1422.

Kevin Karsch, Ce Liu, and Sing Bing Kang. 2014. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 11 (2014), 2144–2158.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 13 (2017), 3521–3526.

Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. 2019. Normal Assisted Stereo Depth Estimation. *arXiv preprint arXiv:1911.10444* (2019).

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. 2018. Learning blind video temporal consistency. In *European Conference on Computer Vision (ECCV)*.

Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. 2016. Deeper depth prediction with fully convolutional residual networks. In *International Conference on 3D Vision (3DV)*.

Manuel Lang, Oliver Wang, Tunc Aydin, Aljoscha Smolic, and Markus Gross. 2012. Practical temporal consistency for image-based graphics applications. *ACM Transactions on Graphics (TOG)* 31, 4 (2012), 1–8.

Xuhong Li, Yves Grandvalet, and Franck Davoine. 2018. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning (ICML)*.

Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. 2019. Learning the Depths of Moving People by Watching Frozen People. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhengqi Li and Noah Snavely. 2018. Megadepth: Learning single-view depth prediction from internet photos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G Narasimhan, and Jan Kautz. 2019. Neural RGB (r) D Sensing: Depth and Uncertainty From a Video Camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. 2015. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 10 (2015), 2024–2039.

Reza Mahjourian, Martin Wicke, and Anelia Angelova. 2018. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. 2016a. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. 2016b. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. http://lmb.informatik.uni-freiburg.de/Publications/2016/MIFDB16 arXiv:1512.02134.

Moritz Menze, Christian Heipke, and Andreas Geiger. 2015. Joint 3D Estimation of Vehicles and Scene Flow. In *ISPRS Workshop on Image Sequence Analysis (ISA)*.

Vaishakh Patil, Wouter Van Gansbeke, Dengxin Dai, and Luc Van Gool. 2020. Don't Forget The Past: Recurrent Depth Estimation from Monocular Video. *arXiv preprint arXiv:2001.02613* (2020).

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

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2019. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *arXiv:1907.01341* (2019).

Rene Ranftl, Vibhav Vineet, Qifeng Chen, and Vladlen Koltun. 2016. Dense monocular depth estimation in complex dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. 2019. Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

David A Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. 2008. Incremental learning for robust visual tracking. *International Journal of Computer Vision* 77, 1-3 (2008), 125–141.

Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. 2016. Artistic style transfer for videos. In *German Conference on Pattern Recognition*.

Ashutosh Saxena, Min Sun, and Andrew Y Ng. 2008. Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 5 (2008), 824–840.

Johannes L Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yunxiao Shi, Jing Zhu, Yi Fang, Kuochin Lien, and Junli Gu. 2019. Self-Supervised Learning of Depth and Ego-motion with Differentiable Bundle Adjustment. *arXiv preprint arXiv:1909.13163* (2019).

Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 2020. 3D Photography using Context-aware Layered Depth Inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. 2012a. A benchmark for the evaluation of RGB-D SLAM systems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 573–580.

J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. 2012b. A Benchmark for the Evaluation of RGB-D SLAM Systems. In *International Conference on Intelligent Robot Systems (IROS). Oct. 2012*.

Richard Szeliski. 2010. *Computer Vision: Algorithms and Applications* (1st ed.). Springer-Verlag, Berlin, Heidelberg.

Chengzhou Tang and Ping Tan. 2019. BA-Net: Dense bundle adjustment network. In *International Conference on Learning Representations (ICLR)*.

Kevin Tang, Vignesh Ramanathan, Li Fei-Fei, and Daphne Koller. 2012. Shifting weights: Adapting object detectors from image to video. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zachary Teed and Jia Deng. 2020. DeepV2D: Video to Depth with Differentiable Structure from Motion. In *International Conference on Learning Representations (ICLR)*.

Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. 2017. Demon: Depth and motion network for learning monocular stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Julien Valentin, Adarsh Kowdle, Jonathan T Barron, Neal Wadhwa, Max Dzitsiuk, Michael Schoenberg, Vivek Verma, Ambrus Csaszar, Eric Turner, Ivan Dryanovski, et al. 2018. Depth from motion for smartphone AR. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–19.

Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. 2017. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804* (2017).

Neal Wadhwa, Rahul Garg, David E Jacobs, Bryan E Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T Barron, Yael Pritch, and Marc Levoy. 2018. Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 64.

Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. 2019a. Web Stereo Video Supervision for Depth Prediction from Dynamic Scenes. In *International Conference on 3D Vision (3DV)*.

Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. 2018b. Learning depth from monocular videos using direct methods. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rui Wang, Stephen M Pizer, and Jan-Michael Frahm. 2019b. Recurrent Neural Network for (Un-) supervised Learning of Monocular Video Visual Odometry and Depth. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018a. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. 2019. Self-Supervised Monocular Depth Hints. In *International Conference on Computer Vision*.

Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. 2018. Lego: Learning edge with geometry all at once by watching videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*.

Zhichao Yin and Jianping Shi. 2018. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. 2019b. Exploiting temporal consistency for real-time video depth estimation. In *International Conference on Computer Vision*.

Shun Zhang, Jia-Bin Huang, Jongwoo Lim, Yihong Gong, Jinjun Wang, Narendra Ahuja, and Ming-Hsuan Yang. 2019a. Tracking Persons-of-Interest via Unsupervised Representation Adaptation. *International Journal of Computer Vision* (2019).

Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. 2018. Deeptam: Deep tracking and mapping. In *European Conference on Computer Vision (ECCV)*.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. 2017. Unsupervised learning of depth and ego-motion from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuliang Zou, Zelun Luo, and Jia-Bin Huang. 2018. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *European Conference on Computer Vision (ECCV)*.