Consistent Video Depth Estimation

Mohamed Nasser

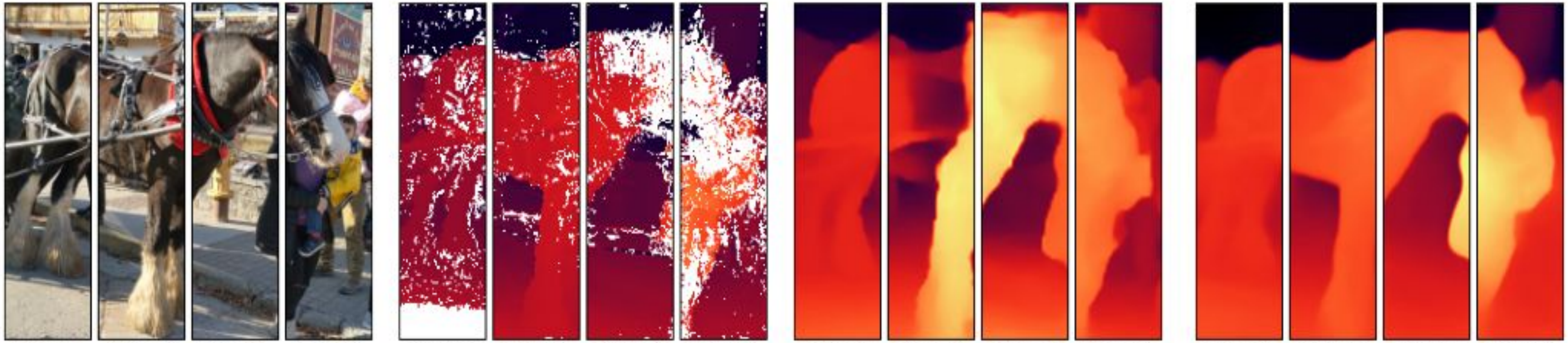
Introduction

algorithm for reconstructing dense, geometrically consistent depth for all pixels in a monocular video.

use a learning-based prior, Conventional structure-from-motion reconstruction to establish geometric constraints on pixels in the video.

able to handle challenging hand-held captured input videos with a dynamic motion.

enables several applications such as scene reconstruction and advanced video-based visual effects.



Frame 1 Frame 2 Frame 3 Frame 4
(a) Input video

Frame 1 Frame 2 Frame 3 Frame 4
(b) COLMAP depth

Frame 1 Frame 2 Frame 3 Frame 4
(c) Mannequin Challenge depth

Frame 1 Frame 2 Frame 3 Frame 4
(d) Our result

1.Pre-processing

Structure-from-Motion (SfM) :COLMAP

apply Mask R-CNN to segment out people To improve pose estimation for videos with dynamic motion ,
SfM : to provide us with the scale of the scene. Because our method works with monocular input, the reconstruction is ambiguous up to scale.

Scale calibration:scale of the SfM and the learning-based reconstructions typically do not match, because both methods are scale-invariant.

Scale calibration:

first compute the relative scale for image i as:

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

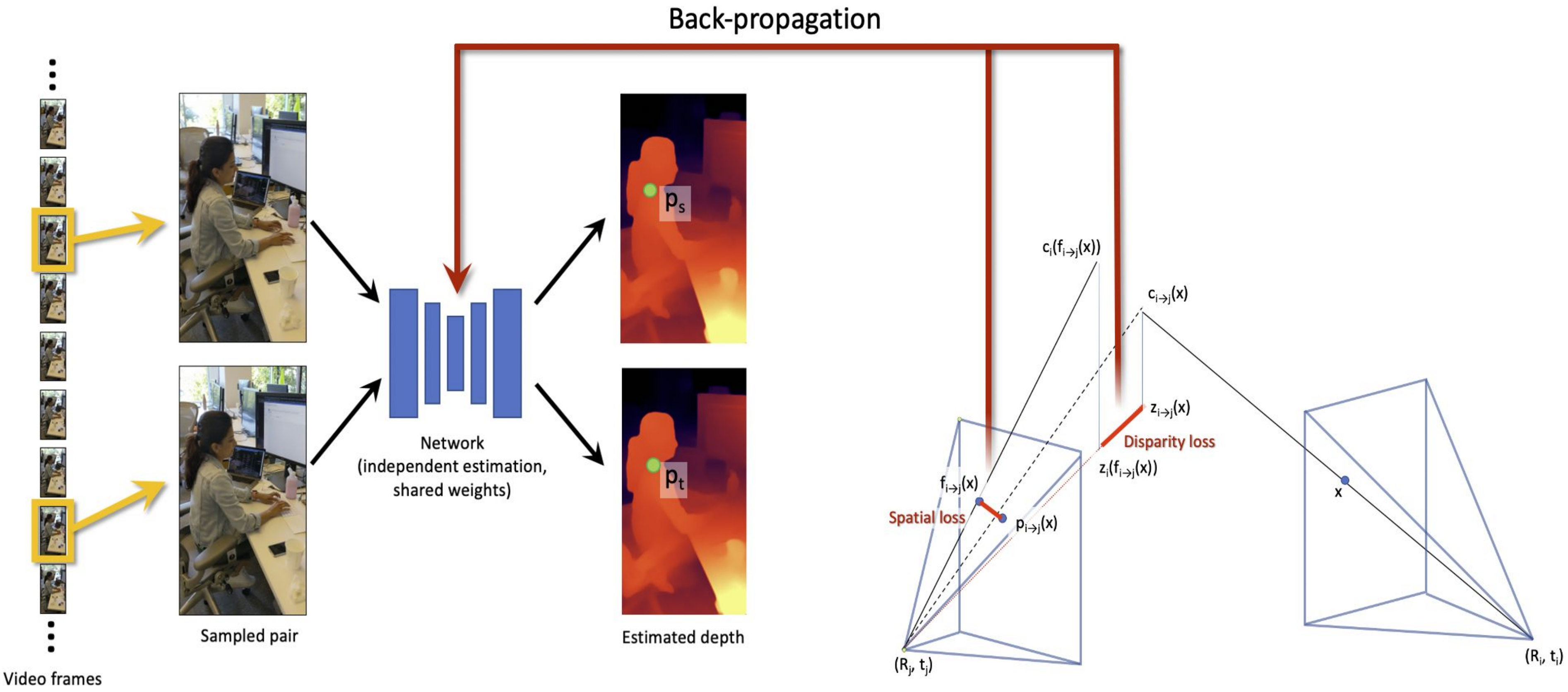
compute the global scale adjustment factor s as

$$s = \underset{i}{\operatorname{mean}} \{ s_i \}, \qquad (2)$$

update all the camera translations

$$\tilde{t}_i = s \cdot t_i. \qquad (3)$$

2.TEST-TIME TRAINING ON INPUT VIDEO



Geometric loss.

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

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

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

$$c_{i\to j}(x) = R_j^{\top}\left(R_i\, c_i(x) + \tilde{t}_i - \tilde{t}_j\right),$$

$$p_{i\to j}(x) = \pi\left(K_j\, c_{i\to j}(x)\right),$$

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

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

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

RESULTS AND EVALUATION

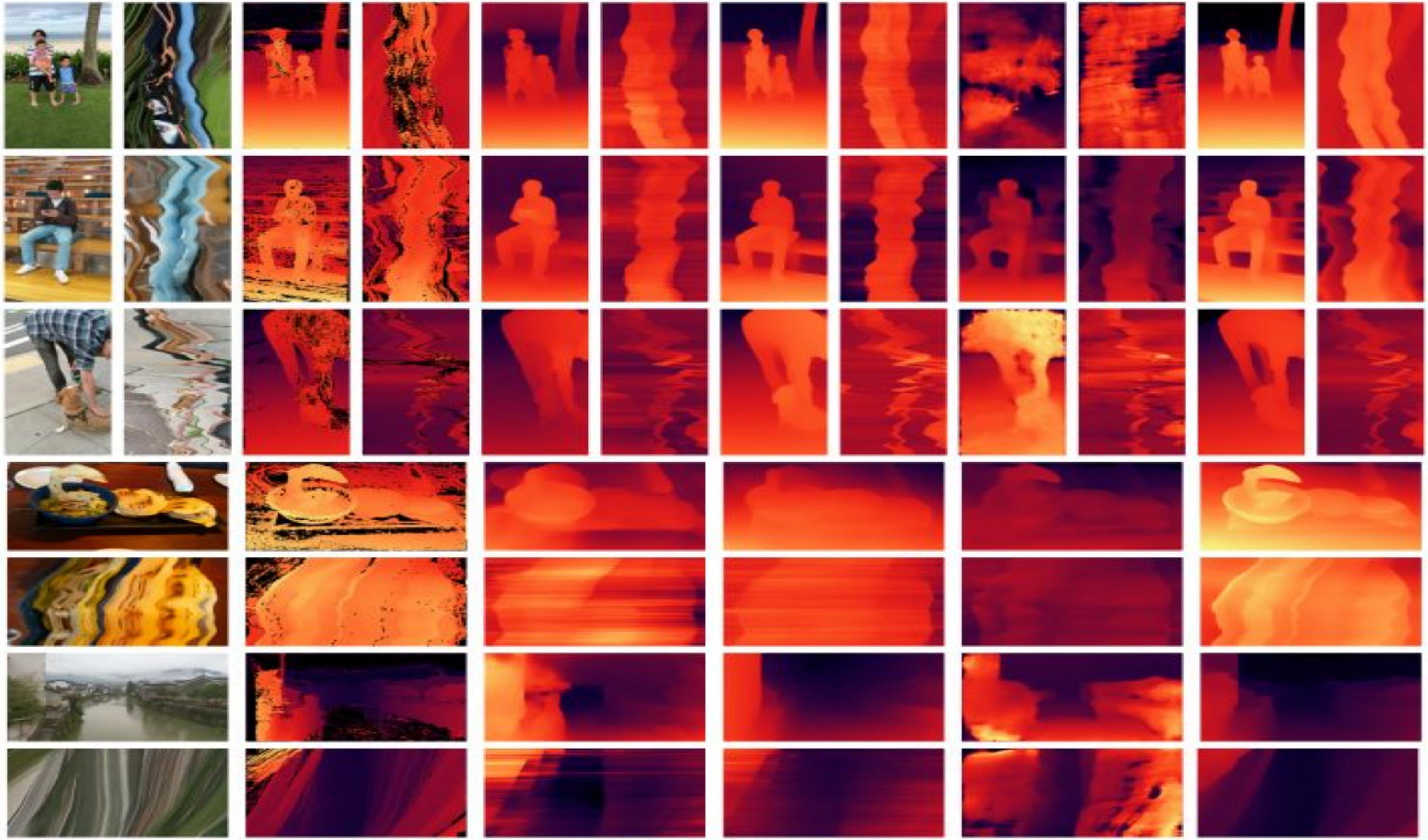
custom stereo video datasets for evaluation.

(1) the TUM dataset (2) the ScanNet dataset (3) the KITTI 2015 datasets

Evaluation

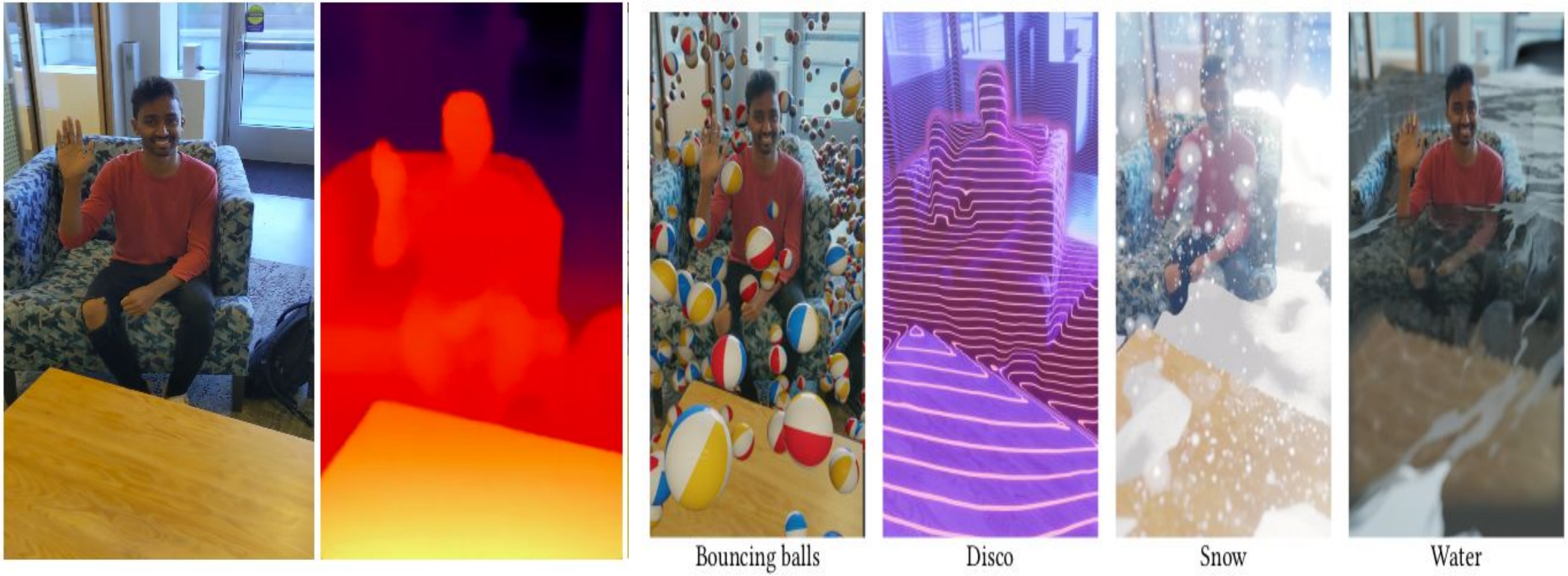
Evaluation metrics.

	Static			Dynamic	
	E_s (%) $\downarrow$	E_d (%) $\downarrow$	$E_p \downarrow$	E_s (%) $\downarrow$	$E_p \downarrow$
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**



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

Consistent video depth estimation enables interesting video-based special effects.



Limitations

- Colmap : to estimate the camera pose from a monocular video
- Dynamic motion : the method supports videos containing moderate object motion. It breaks for extreme object motion.
- Speed : As they extract geometric constraints using all the frames in a video, theydo not support online processing.