Consistent Video Depth Estimation

Andrey Gusev

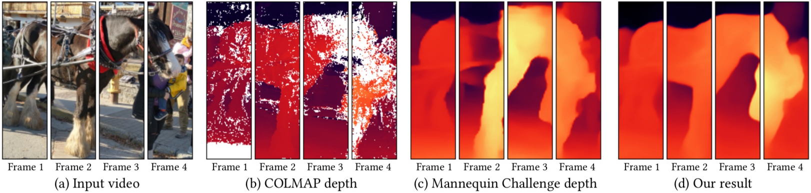
Higher School of Economics

aagusev_2@edu.hse.ru

September 30, 2020

Overview

Introduction



Problems

- Poorly textured areas
- Repetitive patterns,
- Higher noise level
- Shake and motion blur
- Dynamic objects

Approaches

- *Supervised monocular depth estimation*. Single image processing, requires either ground truth or syntetical depth maps.

Approaches

- *Supervised monocular depth estimation*. Single image processing, requires either ground truth or syntetical depth maps.
- *Multi-view reconstruction*. These multi-view stereo techniques assume a static scene.

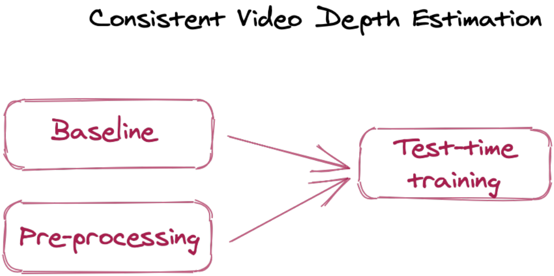
Approaches

- *Supervised monocular depth estimation*. Single image processing, requires either ground truth or syntetical depth maps.
- *Multi-view reconstruction*. These multi-view stereo techniques assume a static scene.
- *Temporal consistency* introduces a “temporal consistency loss”.

Approaches

- *Supervised monocular depth estimation*. Single image processing, requires either ground truth or syntetical depth maps.
- *Multi-view reconstruction*. These multi-view stereo techniques assume a static scene.
- *Temporal consistency* introduces a "temporal consistency loss".

Geometric loss



Geometric loss

- People detection through Mask R-CNN
- Each frame $\Rightarrow$ COLMAP [2016]

Pre-processing

- People detection through Mask R-CNN
- Each frame $\Rightarrow$ COLMAP [2016]
- Scale calibration

$$s_i = \operatorname*{median}_x \left\{ D_i^{NN}(x) / D_j^{MVS}(x) \middle| D_j^{MVS}(x) \neq 0 \right\}$$
$$s = \operatorname*{mean}_i s_i$$
$$\tilde{t}_i = s \cdot t_i$$

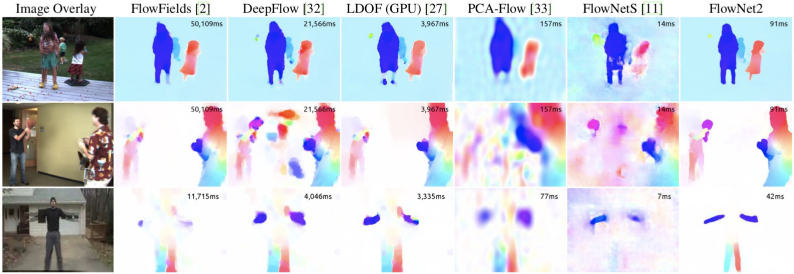
Pre-processing

- People detection through Mask R-CNN
- Each frame $\Rightarrow$ COLMAP [2016]
- Scale calibration
- Frame sampling $\mathcal{O}(N)$

$$S_0 = \{(i,j) : |i-j| = 1\}$$
$$S_k = \left\{(i,j) : |i-j| = 2^k,\ i \bmod 2^{k-1} = 0\right\}$$
$$S = \bigcup_{0 \le k \le \lfloor \log_2(N-1) \rfloor} S_k$$

Geometric loss

- People detection through Mask R-CNN
- Each frame $\Rightarrow$ COLMAP [2016]
- Scale calibration
- Frame sampling $\mathcal{O}(N)$
- Optical flow estimation

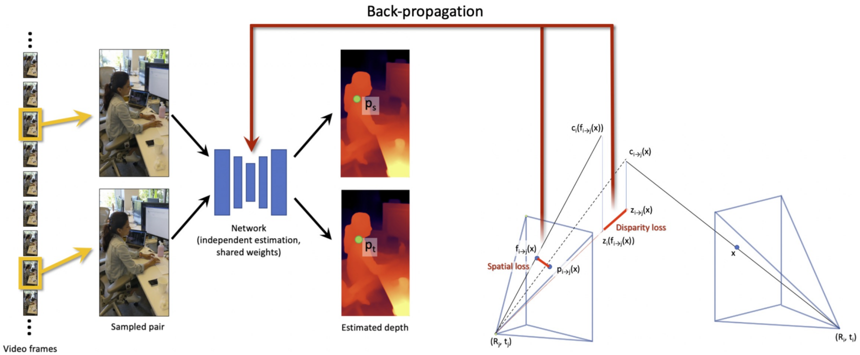


FlowNet2 [Ilg et al. 2017]

Test-time training

Fine-tune a pre-trained depth estimation network so that it produces more geometrically consistent depth for a *particular* input video.

Geometric loss



Spatial loss

Let x be a 2D pixel coordinate in frame i.

$$
\begin{aligned}
f_{i\to j}(x) &= x + F_{i\to j}(x) \\
c_i(x) &= D_i(x)K_i^{-1}(x) \\
c_{i\to j}(x) &= R_j^{\mathsf{T}}\left(R_i c_i(x) + \tilde{t}_i - \tilde{t}_j\right) \\
p_{i\to j}(x) &= \pi\left(K_j c_{i\to j}(x)\right), \text{ where } \pi([x,y,z]^{\mathsf{T}}) = \left[\frac{x}{z}, \frac{y}{z}\right]^{\mathsf{T}}
\end{aligned}
$$

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

Disparity loss

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

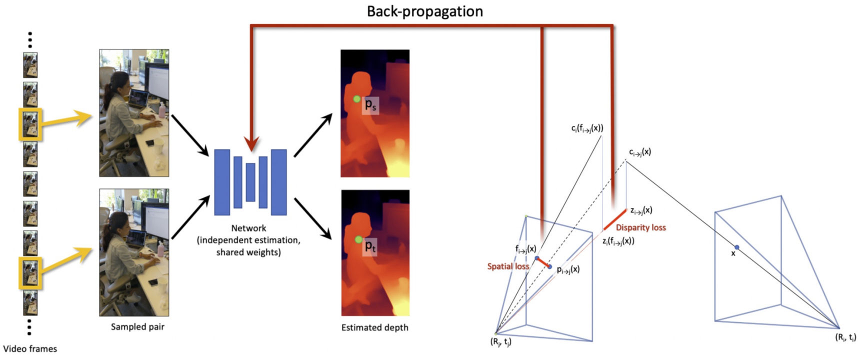
where u_i is frame i's focal length, and z_i and $z_{i\to j}$ are the scalar z-component from c_i and $c_{i\to j}$.

Geometric loss

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

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

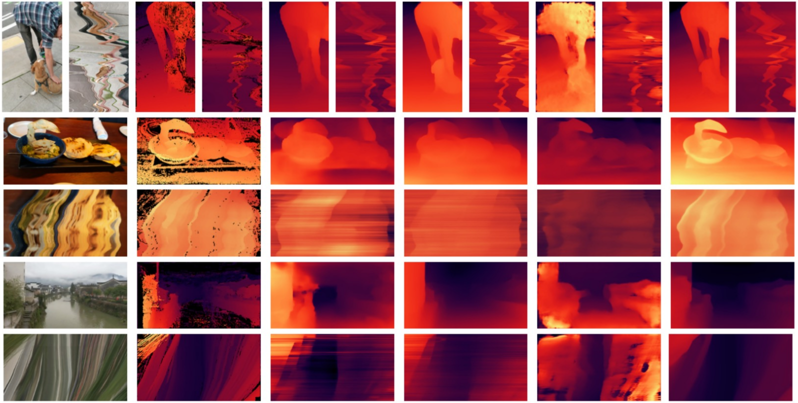
Geometric loss



Geometric loss

w/o spatial loss | w/o disparity loss | w/o scale calibration | w/o overlap test

Examples



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

Examples

(Switch to the tab with videos)

References

Luo et al. (2020)
Consistent Video Depth Estimation

Schönberger et al. (2016)
Structure-from-Motion Revisited

Casser et al. (2019)
Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos

Questions

1. What problems do reconstruction systems have to deal with?
2. What is the primary contribution of this article? What is the main feature of the system that ensures geometrical consistency of depth estimation?
3. How to compute disparity loss?