Unsupervised Learning of Depth and Ego-Motion from Video

Tinghui Zhou\textsuperscript{1}, Matthew Brown\textsuperscript{2}, Noah Snavely\textsuperscript{2}, David G.Lowe\textsuperscript{2}

UC Berkeley\textsuperscript{1}, Google\textsuperscript{2}

Presenter: Zhang Menghe
Jan. 18\textsuperscript{th}, 2019
Background

- Humans can easily perceive 3D from 2D

Image from Cityscape Dataset
Background

- Humans can easily perceive 3D from 2D

Image from Cityscape Dataset
Background

- Humans can easily perceive 3D from 2D

Image from Cityscape Dataset
Humans can easily perceive 3D from 2D
Projection kills the 3\textsuperscript{rd} Dimension

A specific object shape in the 2D plane could be caused by multiple different 3D objects

[Sinha & Adelson, 1993]
Mimic humans’ approach

Learn 3D from a large number of 2D views without any ground-truth 3D labels.
Related Work

Structure from Motion

- Estimate 3D structures from 2D image sequences that may be coupled with local motion signals.
- Rely on accurate image correspondence
  - Bad performance with low texture, complex geometry/photometry, thin structures, occlusions
Related Work

Warping-based View Synthesis

- Synthesize the appearance of the scene seen from novel camera viewpoints.

- DeepStereo
  - End-to-end learning construct view by transforming the input based on depth or flow.
  - The underlying geometry is represented by quantized depth planes.
**Related Work**

**Warping-based View Synthesis**

- Synthesize the appearance of the scene seen from novel camera viewpoints.

- Warping-based method:
  - Estimate the underlying 3d geometry explicitly or establish pixel correspondence among input views.
  - Synthesize the novel views by compositing image patches from input views.
  - Forced to learn intermediate predictions of geometry and/or correspondence.

Render a new view at C from existing images at \( V_1 \) and \( V_2 \)  
[DeapStereo] Flynn et al.2015
Related Work

Single Image 3D Prediction

“Blocks world” Larry Roberts

[Multi-Scale DN] Eigen et al.

1963 2005 2014 2017

[Photo Pop-up] Hoiem et al. [Make 3D] Saxena et al.

[Multi-Scale DN] Eigen et al.

A stereopsis based auto-encoder setup: Part 1 is encoder CNN that maps image to depth map. Part 2 is decoder synthesize a backward warp image. Part 3 simple loss to match reconstructed output with encoder input

Main Idea

Build a consistently well-performed geometric view synthesis system

- Get Intermediate Prediction & Camera Poses
- Build ENTIRE View Synthesis Pipeline
  - The inference procedure of a convolutional neural network.
  - The network is forced to learn about intermediate tasks of depth and camera pose estimation
Main Idea

Unlabeled Video Clips

Single-view depth

Target view → Depth CNN → [gray image]

Relative pose

Nearby views → Pose CNN → [3D representation]

Joint Training Framework

- Jointly train a single-view depth CNN and a camera pose estimation CNN from unlabeled video sequences.
- Jointly train, Independently use.
- Totally unsupervised.
Approaches

View Synthesis

- Learn from Video clips
- Geometric-based view synthesis

If we know the 3D model and camera viewpoints of video frames, we can synthesis video frames by projection.
Approaches

View Synthesis as Supervision

- Use this task as supervision.
- Learn both 3D and pose estimation.
Approaches

View Synthesis as Supervision

- Use this task as supervision.
- Learn both 3D and pose estimation
- 3D representation: Depth Map, Voxels, Layers
Approaches

View Synthesis as Supervision

- **Depth CNN:***
  - Input: Single frame at time $t$, $I_t$
  - Output: Per-pixel depth map $\hat{D}_t$

- **Pose CNN:***
  - Input: Target View($I_t$) and the nearby/source views($I_{t-1}$, $I_{t+1}$)
  - Output: Relative camera poses($\hat{T}_{t \rightarrow t-1}, \hat{T}_{t \rightarrow t+1}$)
Approaches

View Synthesis as Supervision

Photometric error as objective

\[ L_{vs} = \sum_{s \in \{\text{nearby frames}\}} \sum_{p} |I_t(p) - \hat{I}_s(p)| \]

How to solve?

Parameters

- Input: \(<I_1, \ldots, I_n>\) as training frames, where \(I_t\) is the target view others are source view \(I_s\).
- \(\hat{I}_s\) is the source view \(I_s\) warped to target coordinate frame (discuss later).
- \(p\) indexes over pixel coordinates.
Approaches

Differentiable depth image rendering

Reconstruct target view $I_t$ by sampling pixels from a source view $I_s$

Parameters

- $K$ intrinsic camera matrix
- $\hat{D}_t$ predicted depth map
- $\hat{T}_{t\rightarrow s}$ relative pose between $p_t$ and $p_s$

Illustration of the differentiable image warping process.

$$p_s \sim K\hat{T}_{t\rightarrow s} \hat{D}_t(p_t)K^{-1} p_t$$
Approaches

Differentiable Pixel Sampling

Differentiable bilinear sampling mechanism to sample the continuous \( p_s \)

Illustration of the differentiable image warping process.

\[
\hat{I}_s(p_t) = \text{Bilinear interpolation } I_s(p^{tl}_s, p^{tl}_s, p^{br}_s, p^{tr}_s)
\]
Approaches

Modeling the model limitation

1) Dynamic objects
   Target view  Nearby view  Explainability mask

2) visibility/occlusion
   Target view  Nearby view  Explainability mask

- Explainability prediction network that output a per-pixel soft mask $\hat{E}_s$

\[ L_{vs} = \sum_{<I_1...I_n> \in s} \sum_p \hat{E}_s(p) |I_t(p) - \hat{i}_s(p)| \]
Approaches

Overcoming Gradient Locality

- Multi-scale and smoothness loss.
- Allow gradients to be derived from larger spatial regions directly.

\[ L_{final} = \sum_l L_{vs} + \lambda_{SL_{smooth}} + \lambda_e L_{reg}(\hat{E}_s) \]
Approaches

Network Architecture

Depth Network = DispNet + multi-scale side predictions

- DispNet: kernel size 3 for all layers except the first (7,7,5,5) size layers.
Approaches

Network Architecture

Pose Net & Explainability networks:
- Share the first 5 feature encoding layers
- Branch out to predict 6-DOF relative pose and multi-scale explainability masks
- Kernel size 3 for all the layers except for the first 2 and last 2 layers with (7,5,5,7) respectively.
Experiment

Datasets

► Cityscapes
  ▪ Large, semantic, instance-wise, dense pixel annotations of 30 classes
  ▪ 5000 images with high quality annotations, 20,000 images with coarse annotations, 50 different cities

► KITTI
  ▪ Smaller

► Make3D
  ▪ Range Image Data
Experiment

Results for depth map-KITTI

Compared with other supervised training results.

Comparable without using any ground-truth depth or pose labels.
Experiment

Results for depth map-KITTI

- Compared with other supervised training results.

Comparable without using any ground-truth depth or pose labels.
**Experiment**

**Results for depth map**

<table>
<thead>
<tr>
<th>Method</th>
<th>Dataset</th>
<th>Supervision</th>
<th>Error metric</th>
<th>Accuracy metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train set mean</td>
<td>K</td>
<td>✓</td>
<td>0.403</td>
<td>0.593</td>
</tr>
<tr>
<td>Eigen et al. [7] Coarse</td>
<td>K</td>
<td>✓</td>
<td>0.214</td>
<td>0.673</td>
</tr>
<tr>
<td>Eigen et al. [7] Fine</td>
<td>K</td>
<td>✓</td>
<td>0.203</td>
<td>0.702</td>
</tr>
<tr>
<td>Liu et al. [32]</td>
<td>K</td>
<td>✓</td>
<td>0.202</td>
<td>0.678</td>
</tr>
<tr>
<td>Godard et al. [16]</td>
<td>K</td>
<td>✓</td>
<td>0.148</td>
<td>0.803</td>
</tr>
<tr>
<td>Godard et al. [16]</td>
<td>CS + K</td>
<td>✓</td>
<td>0.124</td>
<td>0.847</td>
</tr>
<tr>
<td>Ours (w/o explainability)</td>
<td>K</td>
<td>✓</td>
<td>0.221</td>
<td>0.676</td>
</tr>
<tr>
<td>Ours</td>
<td>K</td>
<td>✓</td>
<td>0.208</td>
<td>0.678</td>
</tr>
<tr>
<td>Ours</td>
<td>CS</td>
<td>✓</td>
<td>0.267</td>
<td>0.577</td>
</tr>
<tr>
<td>Ours</td>
<td>CS + K</td>
<td>✓</td>
<td>0.198</td>
<td>0.718</td>
</tr>
<tr>
<td>Garg et al. [14] cap 50m</td>
<td>K</td>
<td>✓</td>
<td>0.169</td>
<td>0.740</td>
</tr>
<tr>
<td>Ours (w/o explainability) cap 50m</td>
<td>K</td>
<td>✓</td>
<td>0.208</td>
<td>0.695</td>
</tr>
<tr>
<td>Ours cap 50m</td>
<td>K</td>
<td>✓</td>
<td>0.201</td>
<td>0.696</td>
</tr>
<tr>
<td>Ours cap 50m</td>
<td>CS</td>
<td>✓</td>
<td>0.260</td>
<td>0.590</td>
</tr>
<tr>
<td>Ours cap 50m</td>
<td>CS + K</td>
<td>✓</td>
<td>0.190</td>
<td>0.735</td>
</tr>
</tbody>
</table>

Single-view depth results on the KITTI dataset
Experiment

Results for depth map - KITTI Finetune

Comparison of single-view depth predictions on the KITTI dataset by initial Cityscapes model and the final model (pre-trained on Cityscapes and then fine-tuned on KITTI)
Experiment

Results for depth map-Make3D

- Evaluate cross-dataset generalization ability.
- Not seen during training
- Still capable to capture the global scene layout reasonably well.
Experiment

Results for Pose Estimation

<table>
<thead>
<tr>
<th>Method</th>
<th>Seq. 09</th>
<th>Seq. 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORB-SLAM (full)</td>
<td>0.014 ± 0.008</td>
<td>0.012 ± 0.011</td>
</tr>
<tr>
<td>ORB-SLAM (short)</td>
<td>0.064 ± 0.141</td>
<td>0.064 ± 0.130</td>
</tr>
<tr>
<td>Mean Odom.</td>
<td>0.032 ± 0.026</td>
<td>0.028 ± 0.023</td>
</tr>
<tr>
<td>Ours</td>
<td>0.021 ± 0.017</td>
<td>0.020 ± 0.015</td>
</tr>
</tbody>
</table>

- **ORB-SLAM (full):** Recovers odometry using all frames of the driving sequence (3 times more data)

- **ORB-SLAM (short):** Runs on 5-frame snippets

- When side-rotation is small, this network outperforms ORB-SLAM (short) and comparably to ORB-SLAM (full)
Conclusion

- An end-to-end unsupervised learning pipeline.
- Geometric consistency for learning 3D from unlabeled videos.
- “Meta-” supervision: supervise how data behave
Future Work

- Explicitly estimate scene dynamics and occlusions.
  - Direct modeling of scene dynamically.
- Address the situation with no camera intrinsic
- More complicated way to represent 3D scene instead of depth map
- Investigate in more detail the representation learned by this system.
Thanks