SfM-Net: Learning of Structure and Motion from Video

120170157 안권환

Dept. of Electronic Engineering
Sogang University

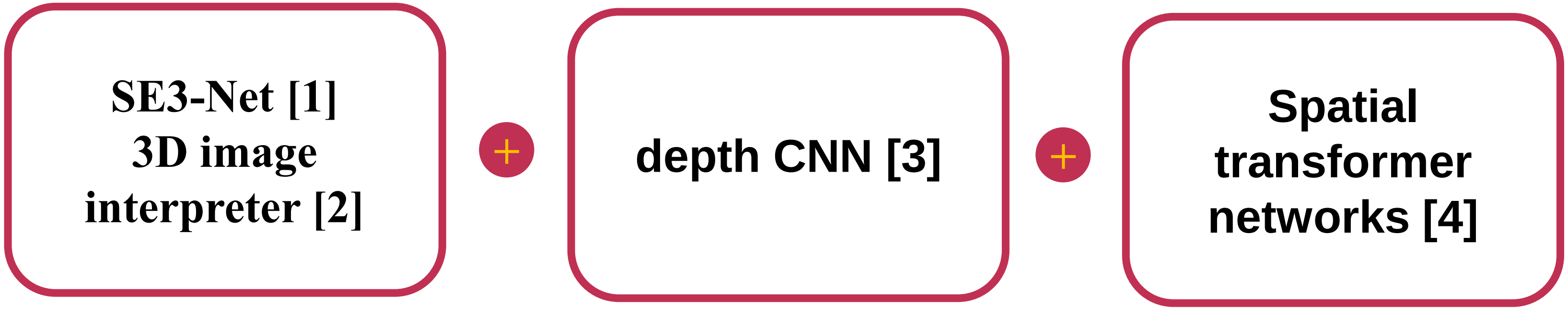
Outline

- Introduction
- What is SfM-Net?
- Preliminaries
- Intuition & Contributions
- Network Architecture
- Training
- Problem Setting & result
- References

What is SfM-Net?

- ***SfM-Net*** =

3D rotation and translations + Single image depth map + Image masking



SE3-Net [1]
3D image interpreter [2]

+

depth CNN [3]

+

Spatial transformer networks [4]

- 3D rotation and translations
  - use an actuation force from a robot
  - an input point cloud to forecast a set of 3D rigid object motions
- Single image depth map
  - Using only single image, extract pixel depth.
- Differentiable image warping

Preliminaries

- Structure from motion (SfM): ***SLAM!***
  - 2차원 정보와 로컬 모션 신호를 결합해서 3차원 구조를 추정하는 방법
  - Point cloud: A set of voxels
- Differentiable image warping
  - learn invariance to translation, scale, rotation and more generic warping

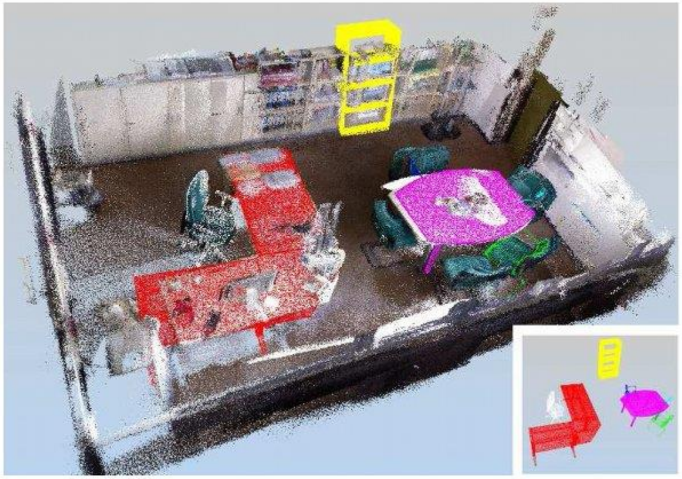


Figure 1: Point cloud.

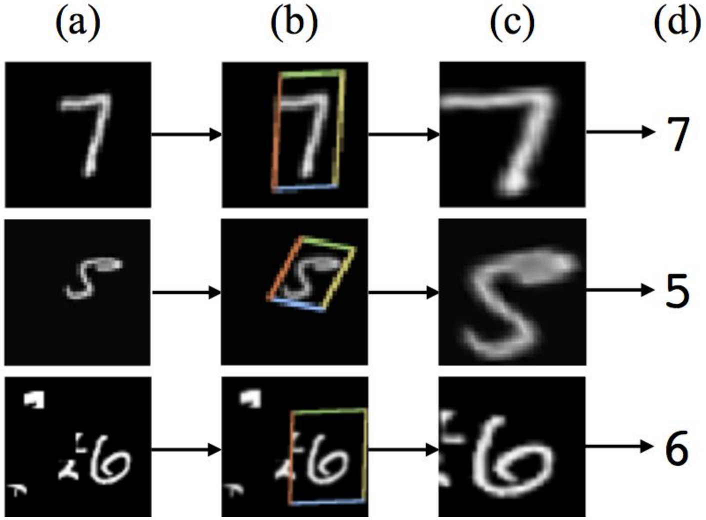
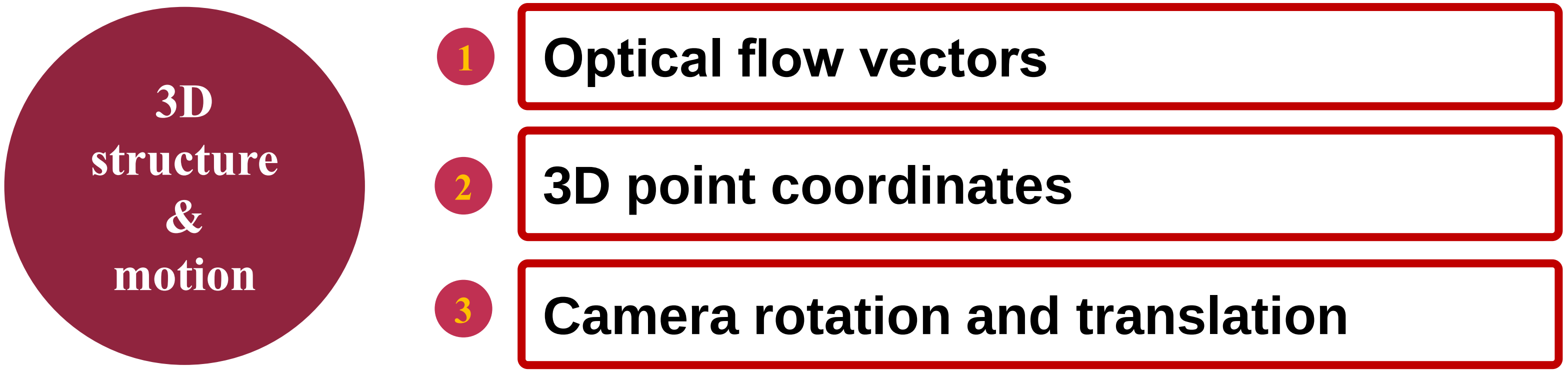


Figure 2: Differentiable image warping.

Intuition & Contributions

- Contributions
  - The model can be trained with various degrees of supervision
  - Supervised by ego-motion (camera motion)
  - Supervised by depth (e.g., as provided by RGBD sensors).

- ***No Direct!***



3D structure & motion

1 Optical flow vectors

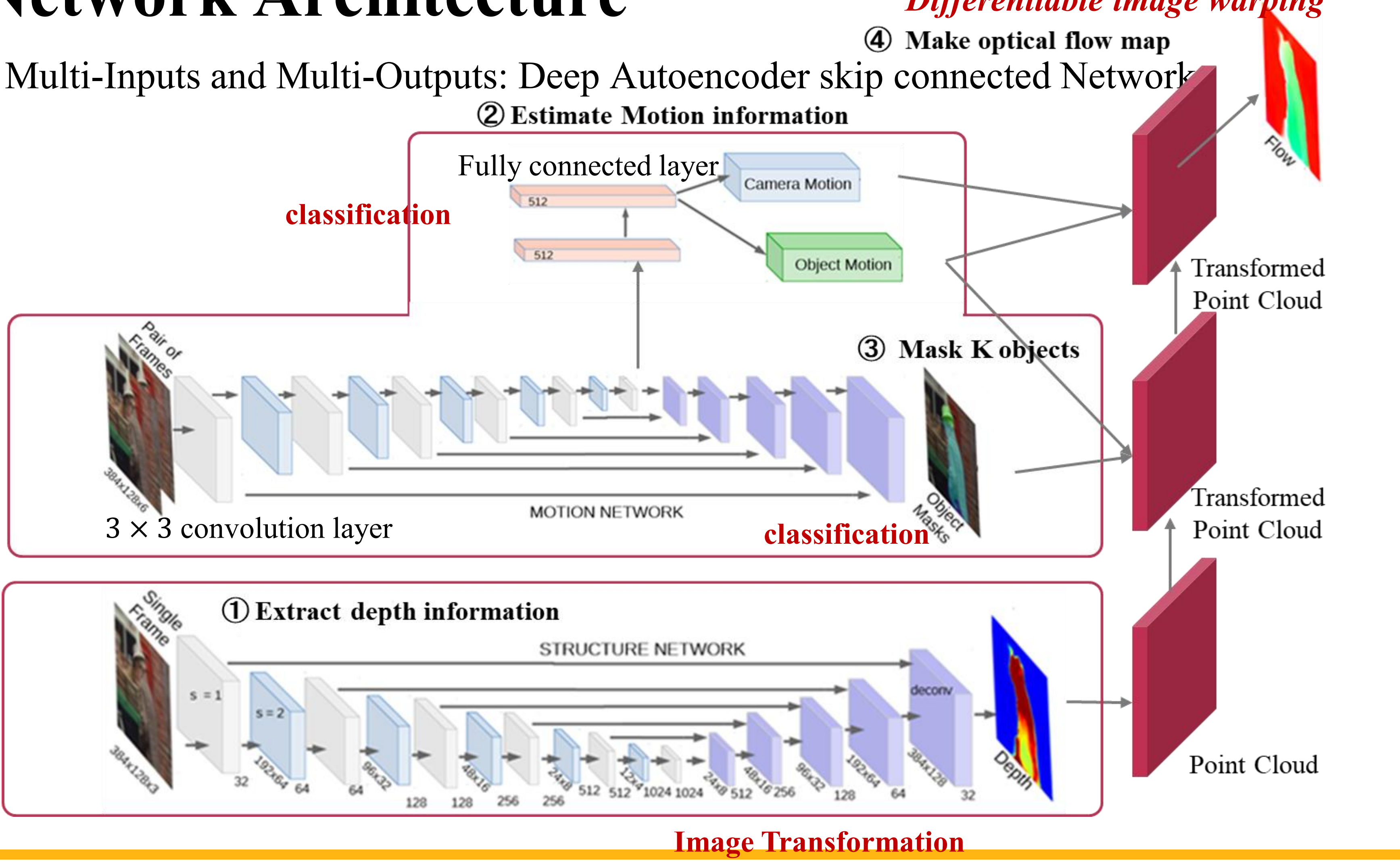
2 3D point coordinates

3 Camera rotation and translation

  - *여러가지를 할 수 있는 하나의 network!*
  - *어려운 것을 풀기 위해서, 하나하나 씩*

Network Architecture

- Multi-Inputs and Multi-Outputs: Deep Autoencoder skip connected Network



Differentiable image warping

④ Make optical flow map

② Estimate Motion information

Fully connected layer

classification

③ Mask K objects

3 × 3 convolution layer

classification

① Extract depth information

Transformed Point Cloud

Transformed Point Cloud

Point Cloud

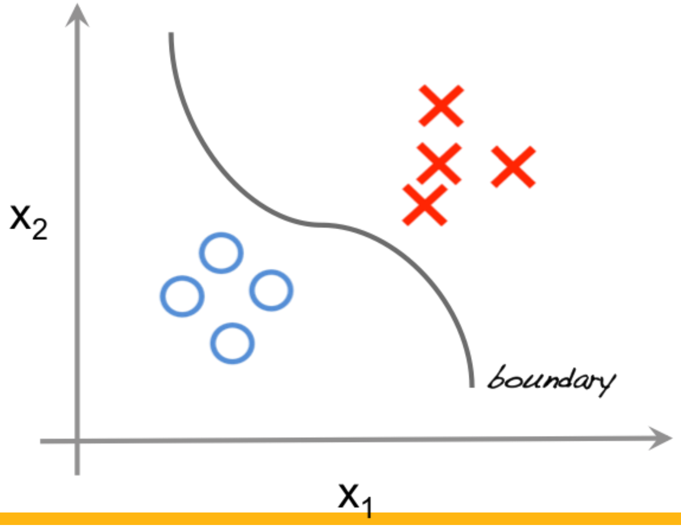
Image Transformation

Training

- Supervised learning / Unsupervised
  - Supervised
    - supervised by ego-motion (camera motion)
    - supervised by depth (e.g., as provided by RGBD sensors)
  - self-supervised by the reprojection photometric error (completely unsupervised)

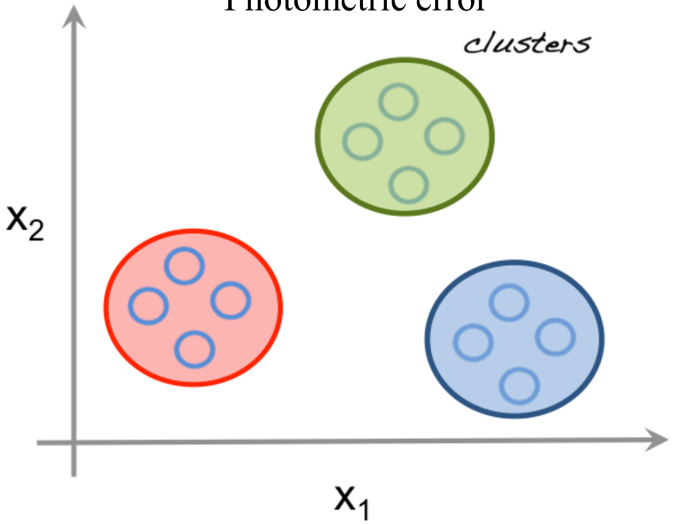
Supervised learning

Ego-motion, Depth(or optical flow)



Unsupervised learning

Photometric error

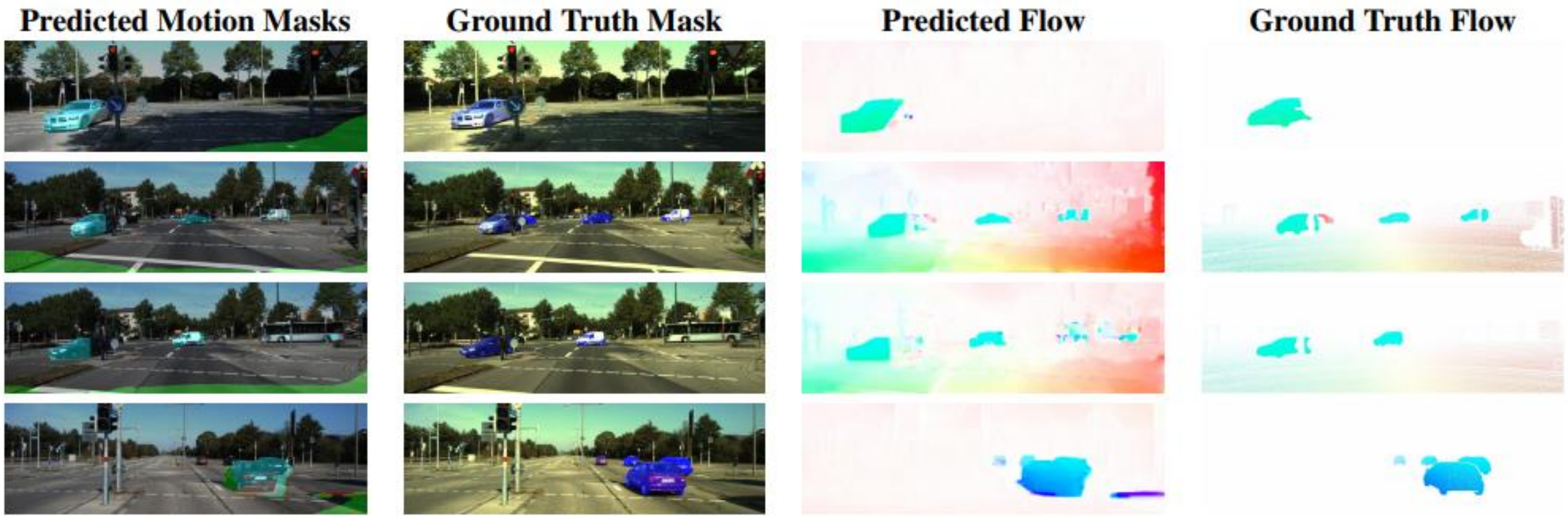


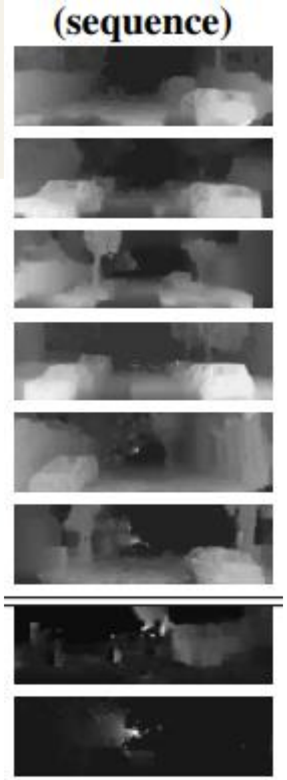
Problem Setting & result

Definition 1. Prediction Problem

Given frames $I_t, \quad I_{\{t+1\}} \in \boldsymbol{R}^{\{w\times h\}}$, Predict

1. Frame depth $d_t \in [0, \infty)^{w\times h}$
2. Camera rotation and translation $\{R_t^c, t_t^c\} \in SE3$
3. A set of K motion masks $m_t^k \in [0,1]^{w\times h}$, $k \in 1, \cdots, K$





References

[1] A. Byravan and D. Fox. SE3-Nets: Learning rigid body motion using deep neural networks. CoRR, abs/1606.02378, 2016.

[2] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3D interpreter network. In ECCV, 2016.

[3] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In ECCV, 2016.

[4] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In NIPS, 2015.