

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.

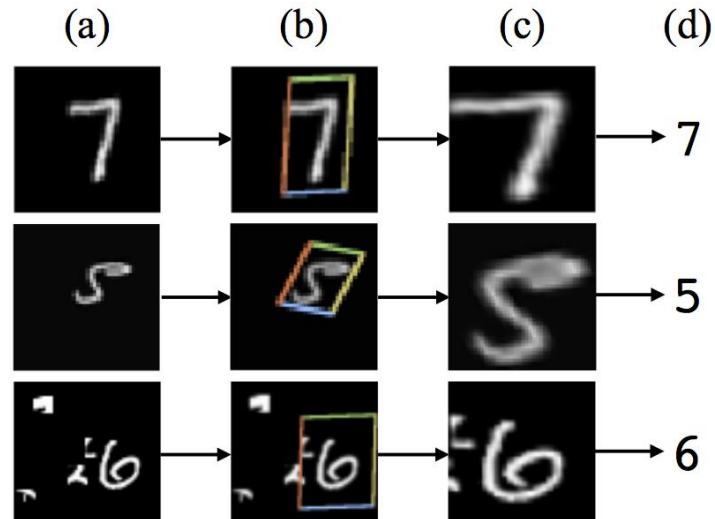
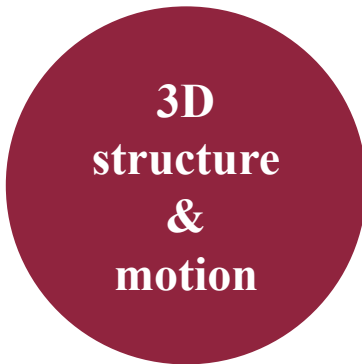


Figure 2: Differentiable image warping.

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!*



- 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

classification

Fully connected layer

Camera Motion

Object Motion

③ Mask K objects

Pair of Frames
384x128x6

3 × 3 convolution layer

MOTION NETWORK

classification

Object Masks

① Extract depth information

STRUCTURE NETWORK

Single Frame
384x128x3

s = 1

s = 2

s = 3

s = 4

s = 5

s = 6

s = 7

s = 8

s = 9

s = 10

s = 11

s = 12

s = 13

s = 14

s = 15

s = 16

s = 17

s = 18

s = 19

s = 20

s = 21

s = 22

s = 23

s = 24

s = 25

s = 26

s = 27

s = 28

s = 29

s = 30

s = 31

s = 32

s = 33

s = 34

s = 35

s = 36

s = 37

s = 38

s = 39

s = 40

s = 41

s = 42

s = 43

s = 44

s = 45

s = 46

s = 47

s = 48

s = 49

s = 50

s = 51

s = 52

s = 53

s = 54

s = 55

s = 56

s = 57

s = 58

s = 59

s = 60

s = 61

s = 62

s = 63

s = 64

s = 65

s = 66

s = 67

s = 68

s = 69

s = 70

s = 71

s = 72

s = 73

s = 74

s = 75

s = 76

s = 77

s = 78

s = 79

s = 80

s = 81

s = 82

s = 83

s = 84

s = 85

s = 86

s = 87

s = 88

s = 89

s = 90

s = 91

s = 92

s = 93

s = 94

s = 95

s = 96

s = 97

s = 98

s = 99

s = 100

s = 101

s = 102

s = 103

s = 104

s = 105

s = 106

s = 107

s = 108

s = 109

s = 110

s = 111

s = 112

s = 113

s = 114

s = 115

s = 116

s = 117

s = 118

s = 119

s = 120

s = 121

s = 122

s = 123

s = 124

s = 125

s = 126

s = 127

s = 128

s = 129

s = 130

s = 131

s = 132

s = 133

s = 134

s = 135

s = 136

s = 137

s = 138

s = 139

s = 140

s = 141

s = 142

s = 143

s = 144

s = 145

s = 146

s = 147

s = 148

s = 149

s = 150

s = 151

s = 152

s = 153

s = 154

s = 155

s = 156

s = 157

s = 158

s = 159

s = 160

s = 161

s = 162

s = 163

s = 164

s = 165

s = 166

s = 167

s = 168

s = 169

s = 170

s = 171

s = 172

s = 173

s = 174

s = 175

s = 176

s = 177

s = 178

s = 179

s = 180

s = 181

s = 182

s = 183

s = 184

s = 185

s = 186

s = 187

s = 188

s = 189

s = 190

s = 191

s = 192

s = 193

s = 194

s = 195

s = 196

s = 197

s = 198

s = 199

s = 200

s = 201

s = 202

s = 203

s = 204

s = 205

s = 206

s = 207

s = 208

s = 209

s = 210

s = 211

s = 212

s = 213

s = 214

s = 215

s = 216

s = 217

s = 218

s = 219

s = 220

s = 221

s = 222

s = 223

s = 224

s = 225

s = 226

s = 227

s = 228

s = 229

s = 230

s = 231

s = 232

s = 233

s = 234

s = 235

s = 236

s = 237

s = 238

s = 239

s = 240

s = 241

s = 242

s = 243

s = 244

s = 245

s = 246

s = 247

s = 248

s = 249

s = 250

s = 251

s = 252

s = 253

s = 254

s = 255

s = 256

s = 257

s = 258

s = 259

s = 260

s = 261

s = 262

s = 263

s = 264

s = 265

s = 266

s = 267

s = 268

s = 269

s = 270

s = 271

s = 272

s = 273

s = 274

s = 275

s = 276

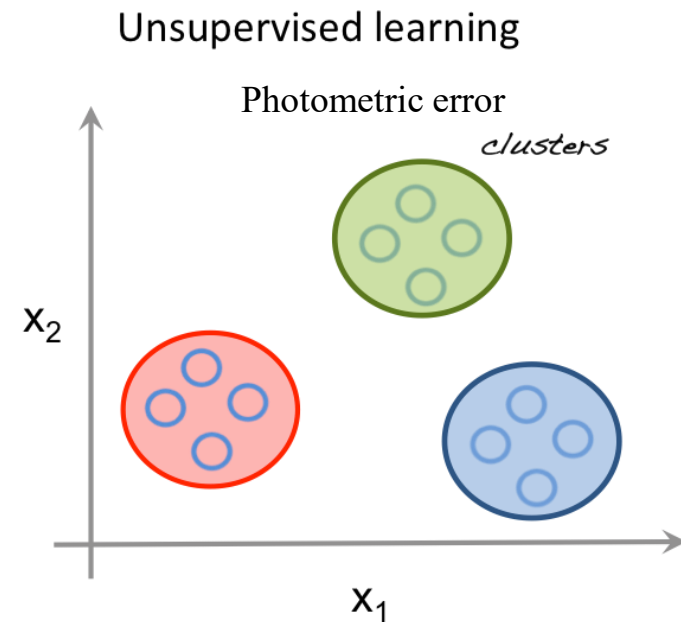
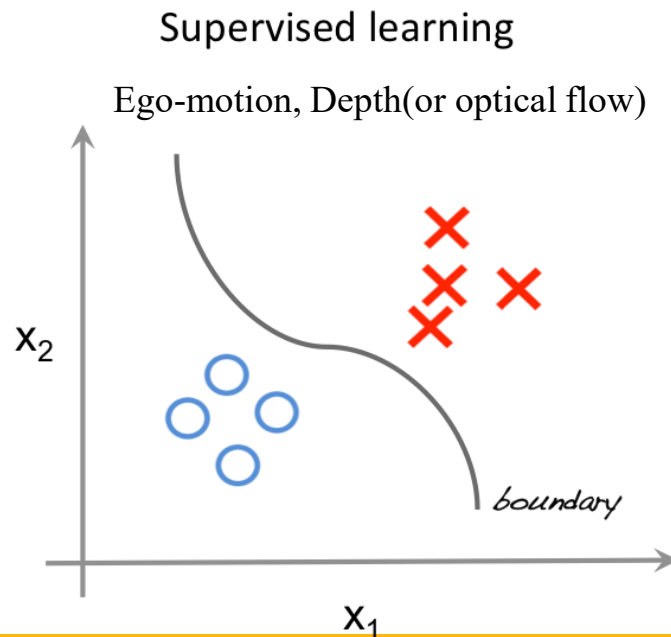
s = 277

s = 278

s = 279

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)



Problem Setting & result

Definition 1. Prediction Problem

Given frames $I_t, I_{\{t+1\}} \in \mathbf{R}^{w \times h}$, 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 $m_t^k \in [0, 1]^{w \times h}, k \in 1, \dots, K$

(sequence)



Predicted Motion Masks

Ground Truth Mask

Predicted Flow

Ground Truth Flow



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.