Unsupervised Learning of Monocular Depth and Ego-Motion using Conditional *Patch*GANs

Madhu Vankadari,
Swagat Kumar, Anima Majumder & Kaushik Das

TCS Research, Bangalore, India.

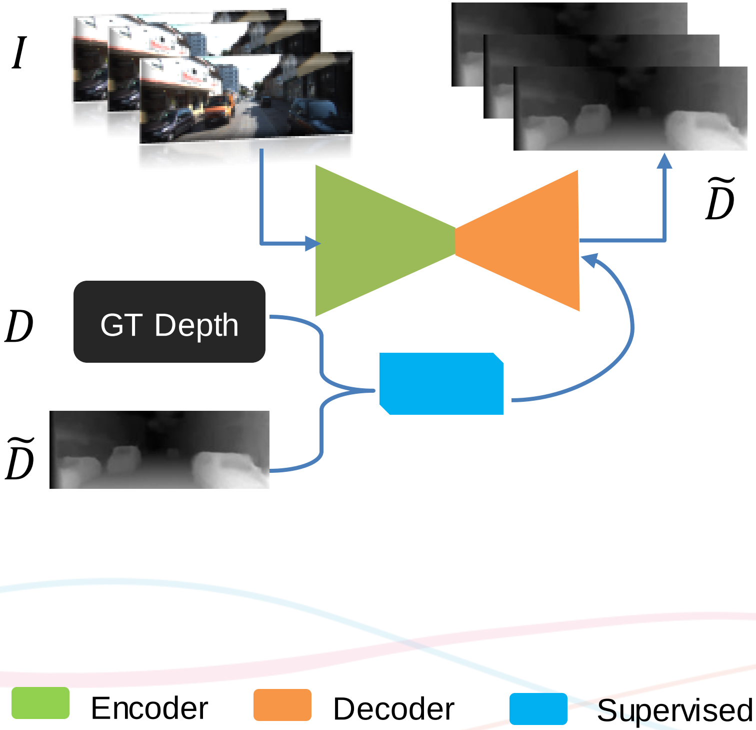
Why do we need Depth and Ego-Motion ?

- Robot Navigation (Autonomous driving, Drones, etc.)
- 3D Reconstruction
- Manipulator Grasping
- Computer Graphics

Outline

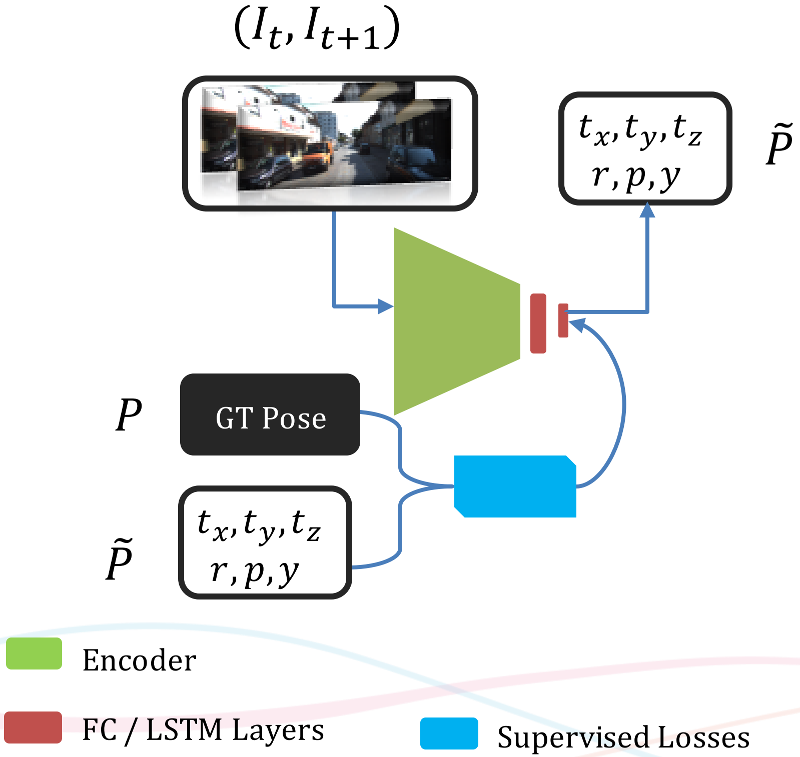
- Supervised Learning approaches
- Unsupervised Learning approaches
- Proposed method
- Quantitative and Qualitative comparison
- Conclusions and Future Scope

Supervised Methods for Depth Estimation



- CNN Encoder-Decoder regression by Egien et al., 2014
- CRFs with CNN as post-processing by Li et al. 2015.
- Robust loss functions using scene priors by Laina et al., 2017
- Regression losses to classification losses with depth discretization by Cao et al., 2018

Ego-Motion Estimation using Deep Learning



- Learning Visual odometry using a CNN by konda. K et al., 2015
- Flow-Net variant as CNN encoder followed by two LSTM layers for 6 DOF pose by Wang et al., 2017

SOTA for Unsupervised Depth and Ego Motion

- Stereo-Monocular Methods

  Geometry to the rescue by Garg et al., 2016

  Monodepth by Godard et al., 2017

  UnDEMoN by Babu et al., 2018

  Depth – Feat VO by Zhan et al., 2018

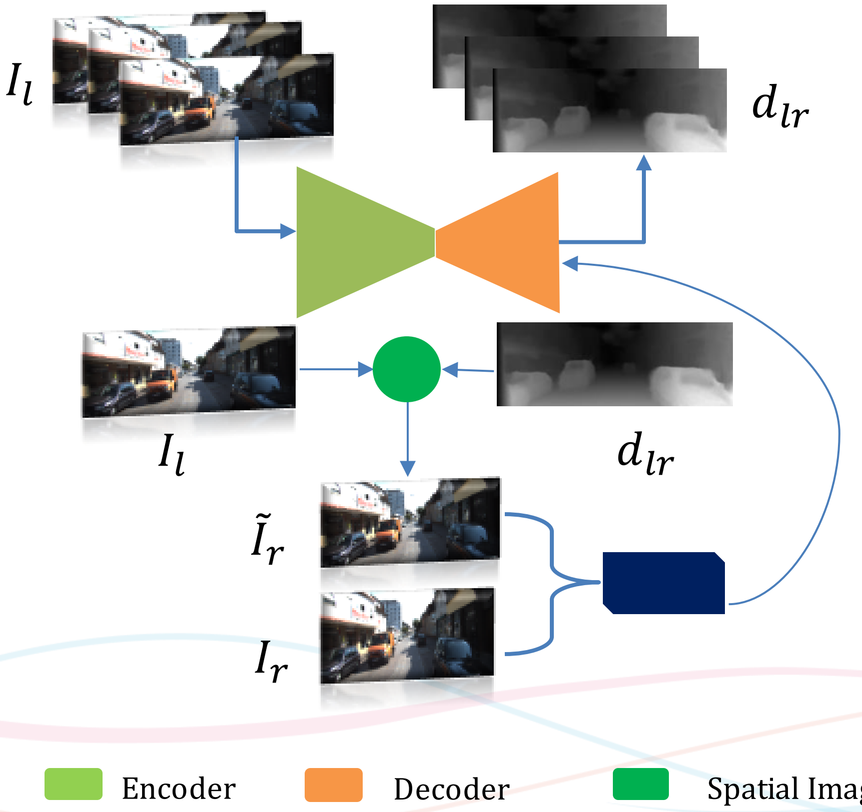
  MonoGAN by Aleotti el al., 2018

- Monocular only

  SfMLearner by Zhou et al., 2017

  Vid2Depth by Mahjourian et al., 2018

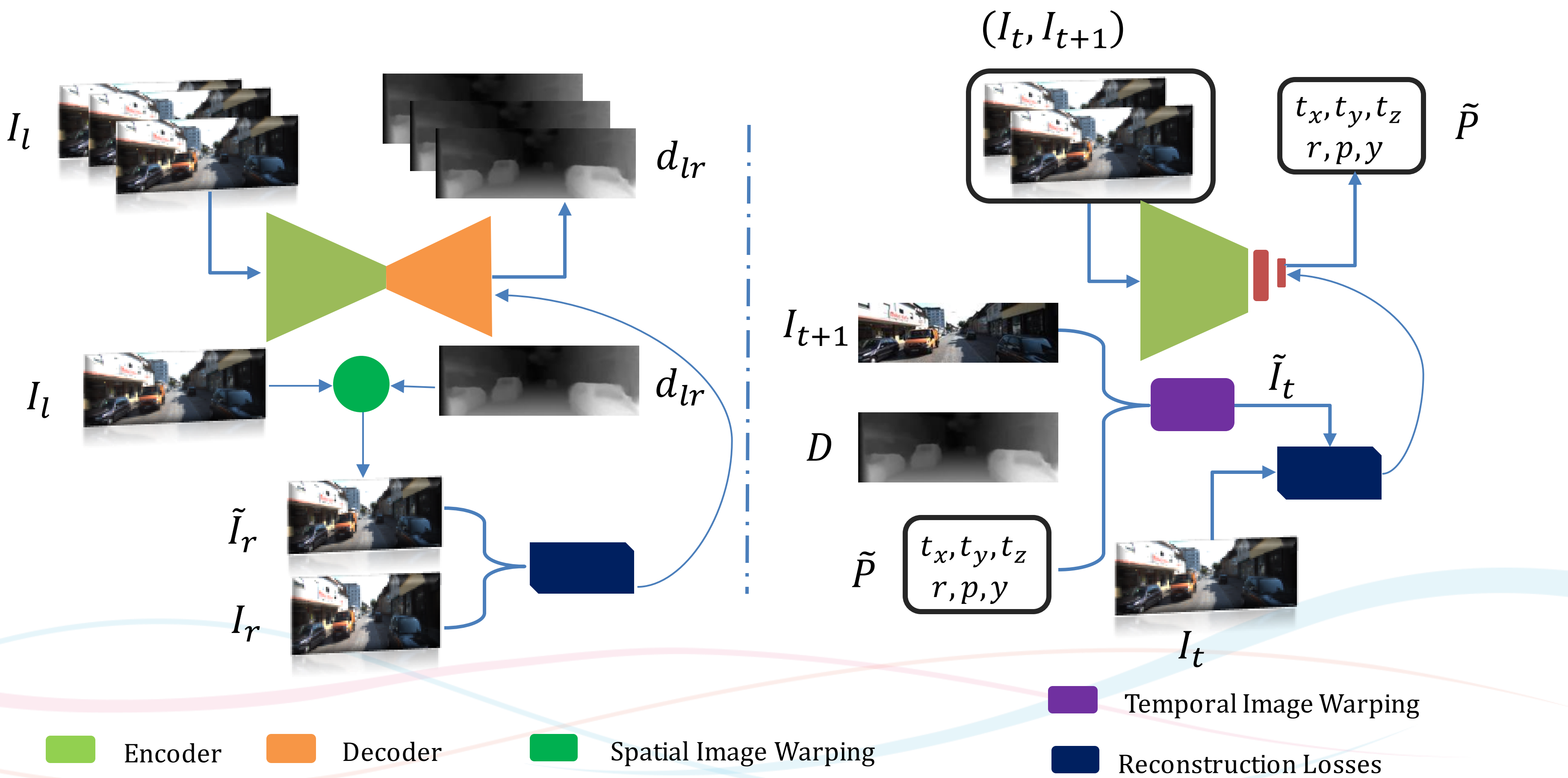
Unsupervised Disparity / Depth Learning



- Predict disparity and reconstruct the opposite stereo image from the input by Garg et al., 2016

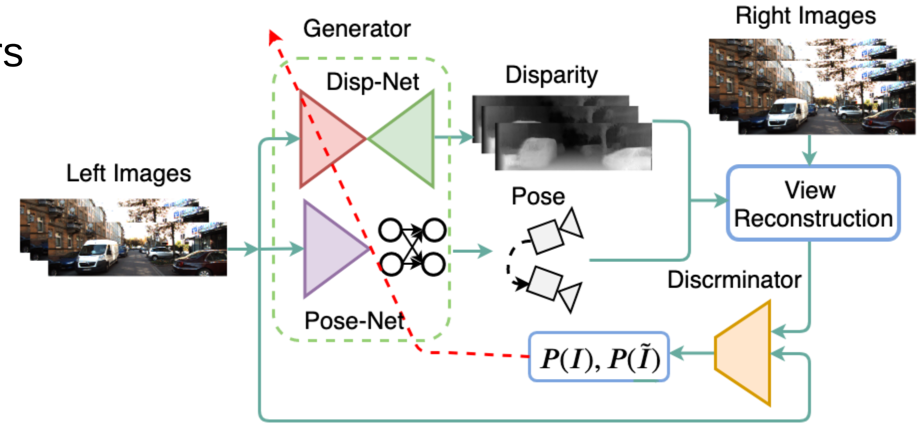
- Left-right and right-left disparity prediction and enforcing consistency between them by Godard et al., 2017

Depth and Ego-Motion together

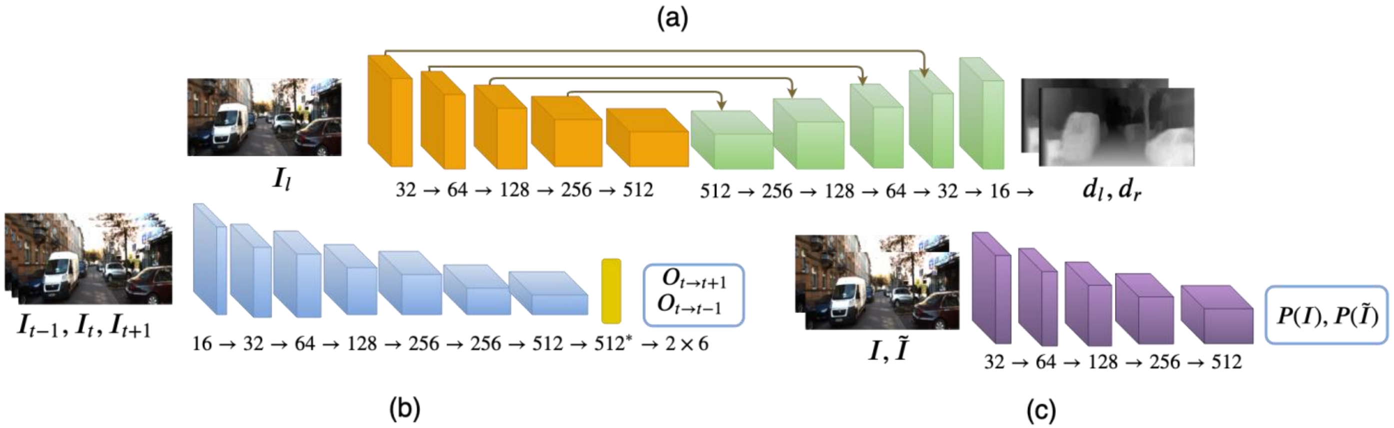


Proposed Method

- Total approach as GAN paradigm
- Depth and Pose Networks are Generators **conditioned** on **input** RGB Image(s)
- Image reconstruction using predicted disparity and pose
- Patch-Based Image discriminator (Patch GAN)
- Total loss is a weighted combination of reconstruction and adversarial losses



Network Architecture



a.) Disp-Net, inspired from SfMLearner by Zhou et al., but with a lesser number of conv-layers.

b.) Pose-Net (Conv Encoder followed by fully connected layers)

c.) Patch based Image discriminator

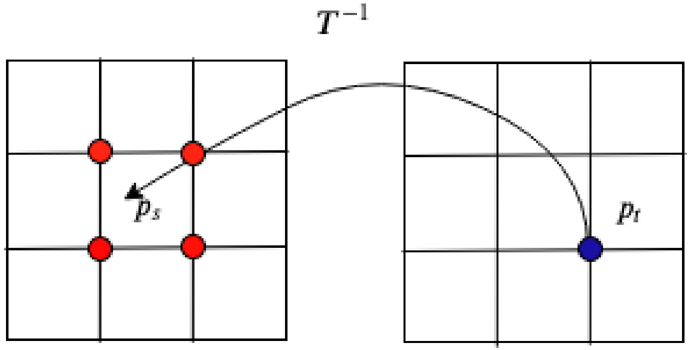
Total No of trainable parameters 19M. The Disp-Net has only 8M, and Pose-Net has 6M

Spatial / Temporal Image Warping

Spatial Image Warping :

The transformation **T** is the **disparity**,

The **source image** is **left for right** and vice versa



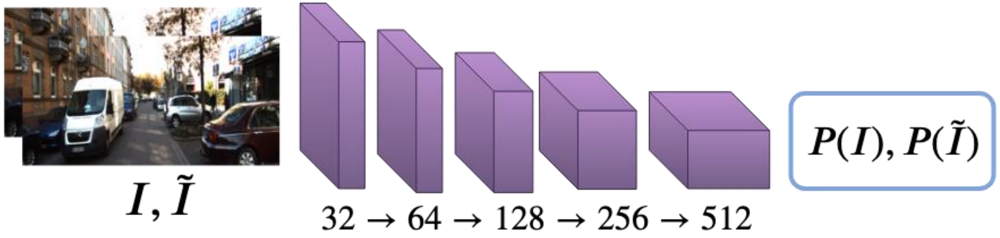
Temporal Image Warping

The transformation **T** is the **ego-motion** estimated

The **source images** will temporally aligned images

The transformation happens in **Camera-coordinate system**

Patch GAN



- Completely **convolutional** so computationally simple
- Facilitates to evaluate the image locally as **patches**
- Patch sizes can be **varied**
- Ablation studies are performed to fix the **size** of the patch

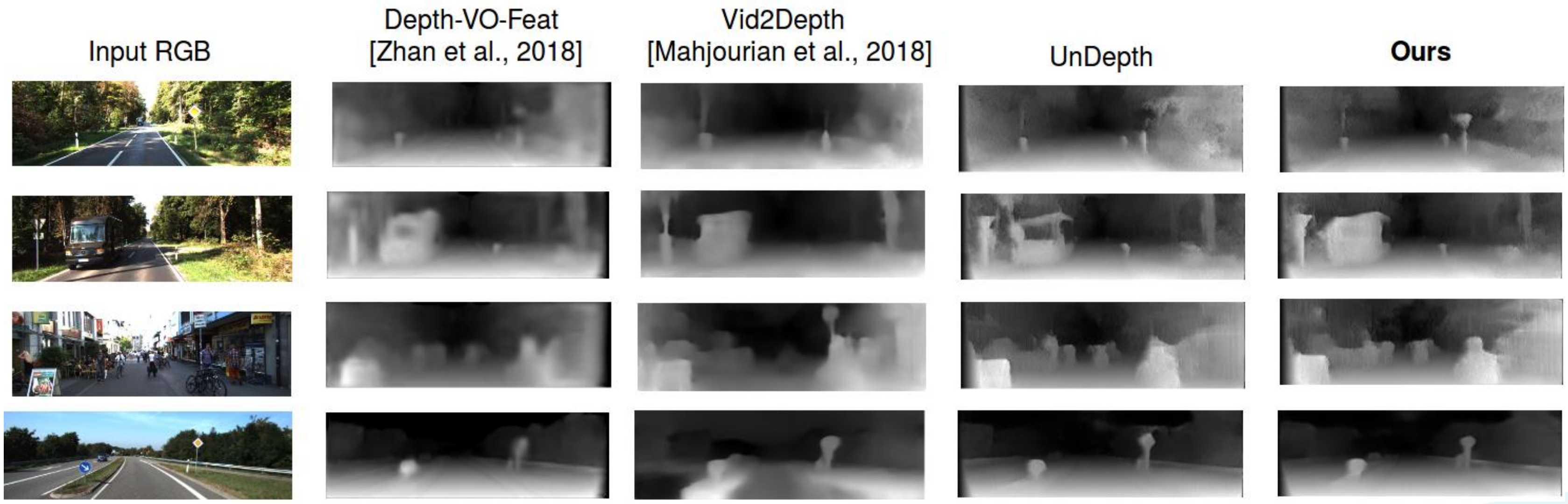
Loss Functions

- Image reconstruction losses both spatial and temporal domains
  - Pixel-wise mean squared error (MSE)
  - Structural Similarity Index (SSIM)
- Edge aware disparity smoothness for both left-right and right-left disparities
- Left-Right consistency Loss
- Adversarial losses

Dataset and Training

- KITTI Outdoor Driving Dataset
- Total different **61** sequences with **42382** images of **1242x375** resolution
- Eigen et al split: **32** scenes with **22600** images for training, **697** from **29** scenes for testing. (evaluation is done with **LIDAR** data)
- KITTI Stereo split: **33** scenes with **29000** for training, **200** from **28** scenes for testing (evaluation is done with **GT depth** data given with the dataset)
- Trained for **0.24M** iterations on GTX 1080 for **23** hours and the total script is written in Tensorflow

Depth Results



Depth Results

Method	Abs.Rel	Sq.Rel	RMSE	Log RMSE	Δ<1.25	Δ < 1.25^2	Δ < 1.25^3
MonoGAN	0.119	1.239	5.998	0.212	0.846	0.940	0.976
Proposed	**0.110**	**1.044**	**5.535**	**0.200**	**0.849**	**0.944**	**0.979**

Method	Abs.Rel	Sq.Rel	RMSE	Log RMSE	Δ<1.25	Δ < 1.25^2	Δ < 1.25^3
UnDEMoN	0.139	1.174	5.59	0.239	0.812	0.930	0.968
Proposed	**0.1269**	**0.9982**	**5.309**	**0.226**	**0.827**	**0.934**	**0.971**

Pose Results

- Absolute trajectory error (ATE) as Zhou et al., 2017

Seq	UnDEMoN (t)	**Proposed (t)**
00	0.0644	**0.0593**
04	0.0974	**0.0713**
05	0.0696	**0.0651**
07	0.0742	**0.0666**

Conclusion

- Proposed method predict depth without actually using GT-Depth
- Also, predicts ego-motion
- Reconstruction losses and adversarial losses are used for training
- Able to get 5.2% improvement over state-of-the art
- Small Depth-Net which is able to produce around 30 fps on a GTX 1080 machine

Future Work

- Night Vision Depth and Ego-Motion Estimation
- Initial Results

Thank You