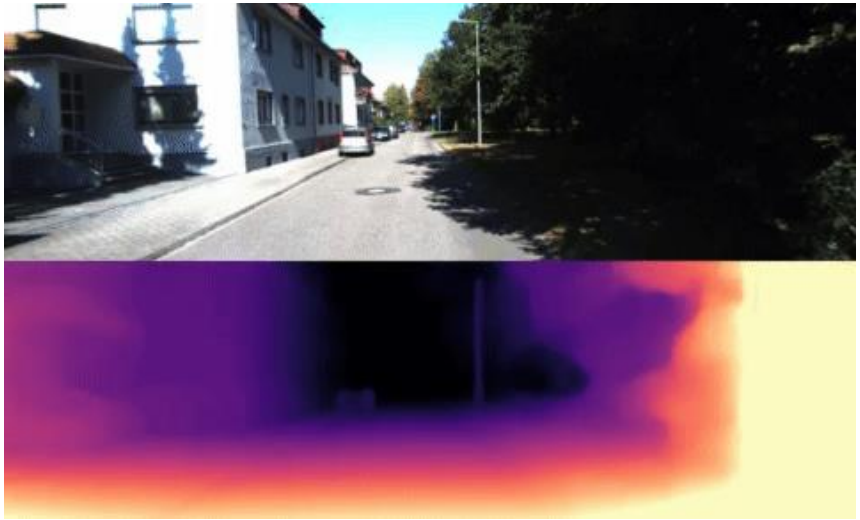
# Segmentation-Guided Monocular Depth Estimation

**Digging Into Self-Supervised Monocular Depth Estimation (ICCV 2019)**

Tesla Driver
Hakan Sivuk 21601899
https://github.com/hakansivuk/SegDepth

# Outline

- Introduction
- Related Work
- Proposed Approach
- Results
- Conclusion

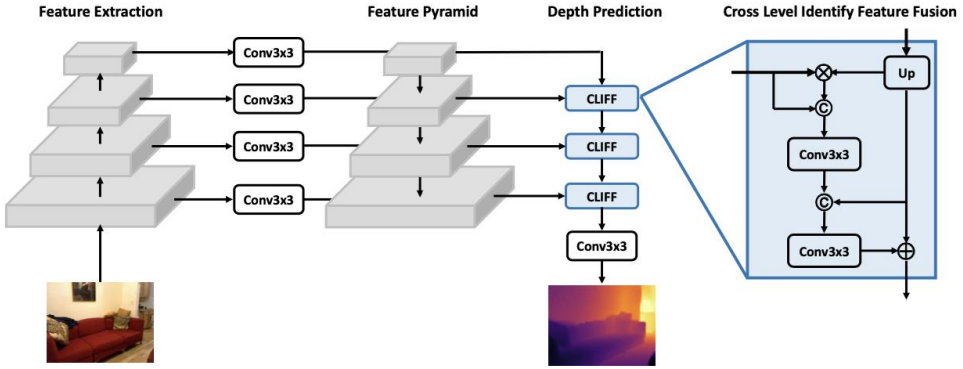


- Supervised Depth Estimation
- Self-Supervised Depth Estimation
- The close relationship between segmentation and depth estimation

# Related Work

- Self-Supervised Depth Estimation
- Segmentation with Depth Values
- Joint Learning
- Depth Estimation with Segmentation

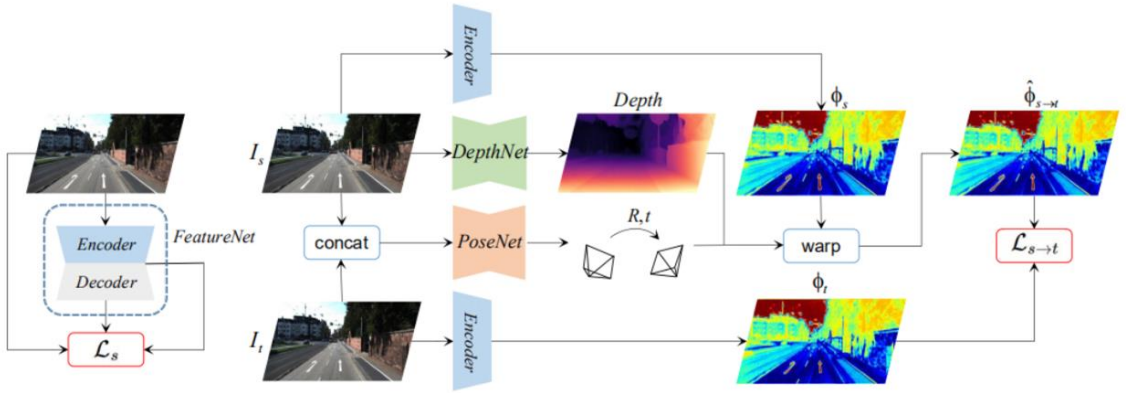
# CLIFFNet for Monocular Depth Estimation with Hierarchical Embedding Loss (ECCV 2020)




- **CLIFF Module**
- **Hierarchical Embedding Loss**
- NYU-Depth V2, Cityscapes
- RMSE (log), Abs Rel, Sq Rel, Pn accuracy for n=1,2,3

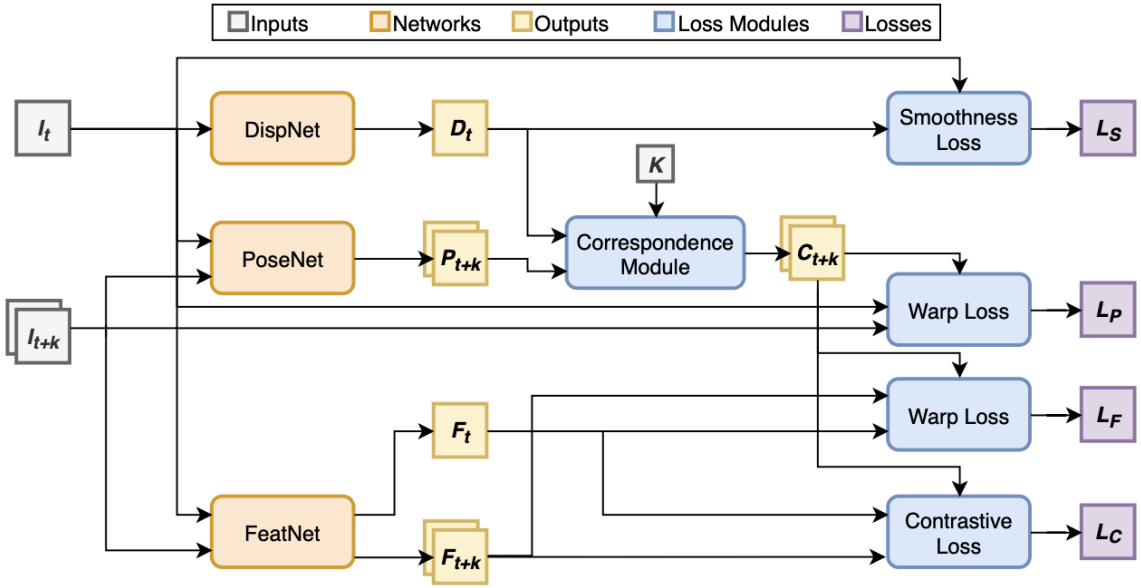
| Methods | L1 | Grad | SI | DA | MT | HEG-Rn | HEG-Im | HEG-R | HEG-S |
|---|---|---|---|---|---|---|---|---|---|
| RMS | 0.529 | 0.513 | 0.520 | 0.511 | 0.530 | 0.517 | 0.523 | 0.497 | **0.493** |
| Abs Rel | 0.135 | 0.132 | 0.134 | 0.130 | 0.134 | 0.134 | 0.132 | 0.129 | **0.128** |
| P1 | 0.817 | 0.830 | 0.820 | 0.835 | 0.815 | 0.815 | 0.829 | 0.841 | **0.844** |
| P2 | 0.961 | 0.964 | 0.963 | 0.964 | 0.960 | 0.961 | 0.963 | **0.963** | **0.964** |

# Feature-Metric Loss for Self-Supervised Learning of Depth and Egomotion (ECCV 2020)



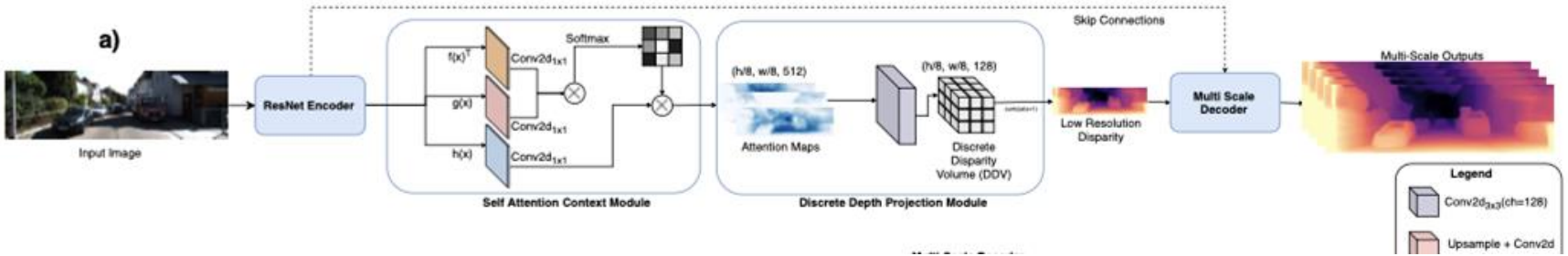

- **FeatureNet**
- **Feature-metric loss, discriminative loss, convergent loss,** photometric loss
- Minimum per pixel reprojection error
- KITTI 2015
- RMSE (log), Abs Rel, Sq Rel, Pn accuracy for n=1,2,3

# DeFeat-Net: General Monocular Depth via Simultaneous Unsupervised Representation Learning (CVPR 2020)



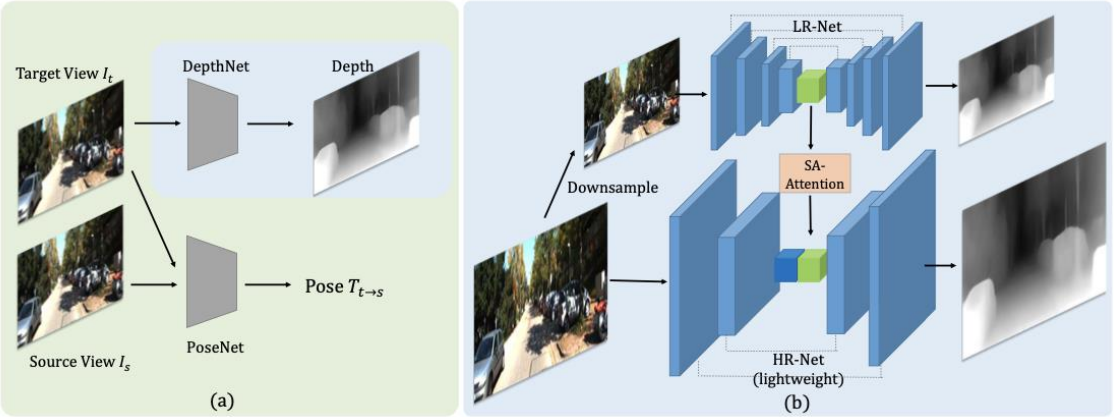

- **FeatNet**
- Minimum per pixel reprojection error, binary auto-masking*
- Smoothness loss, photometric and feature warp loss, contrastive loss
- KITTI 2015, Robot-Car Seasons
- RMSE (log), Abs Rel, Sq Rel, Pn accuracy for n=1,2,3

# Self-supervised Monocular Trained Depth Estimation using Self-attention and Discrete Disparity Volume(CVPR 2020)



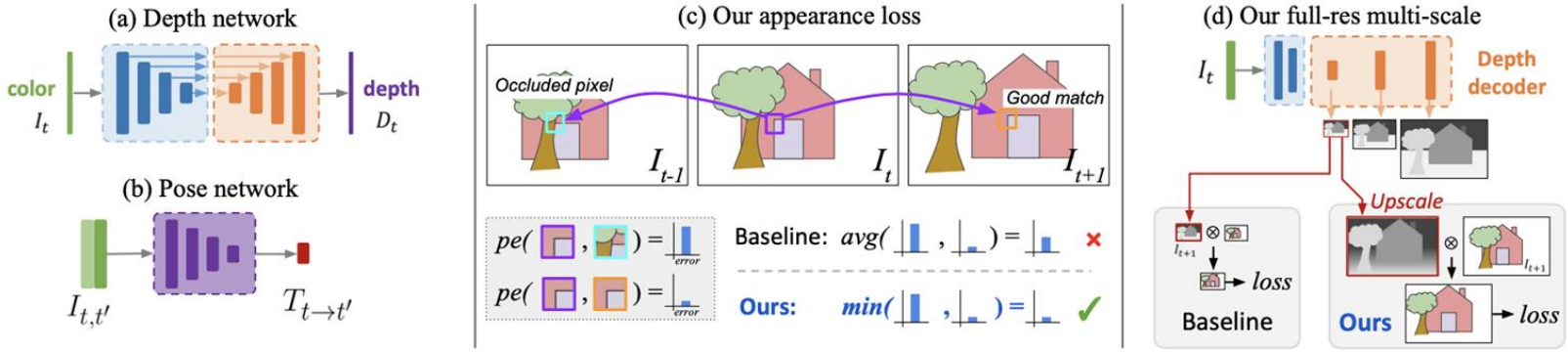

- **Self attention, discrete disparity prediction**
- Minimum per pixel reprojection error, binary auto-masking*
- Edge-aware smoothness loss, photometric loss
- KITTI 2015, Make3D
- RMSE (log), Abs Rel, Sq Rel, Pn accuracy for n=1,2,3

# Unsupervised High-Resolution Depth Learning From Videos With Dual Networks




- **LR-Net, HR-Net, SA_Attention**
- Minimum per pixel reprojection error, binary auto-masking*
- Smoothness loss, photometric loss, feature reconstruction loss
- KITTI 2015, Make3D
- RMSE (log), Abs Rel, Sq Rel, Pn accuracy for n=1,2,3

# Main Paper: Digging Into Self-Supervised Monocular Depth Estimation (ICCV 2019)




- **A minimum reprojection loss, binary-auto masking, full resolution multi scale sampling**
- Edge-aware smoothness loss, photometric loss
- KITTI 2015
- RMSE (log), Abs Rel, Sq Rel, Pn accuracy for n=1,2,3

# Select, Supplement and Focus for RGB-D Saliency Detection (CVPR 2020)

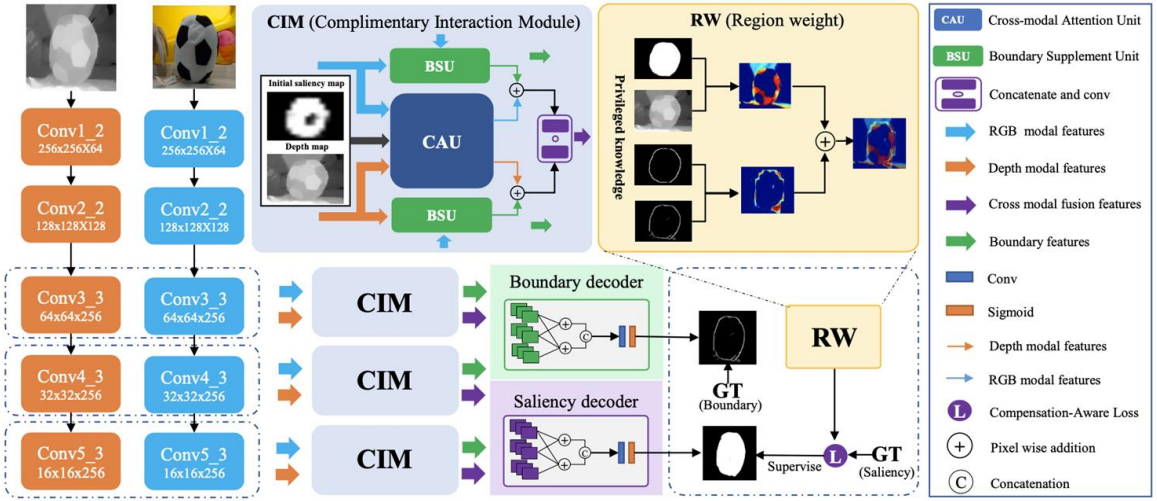
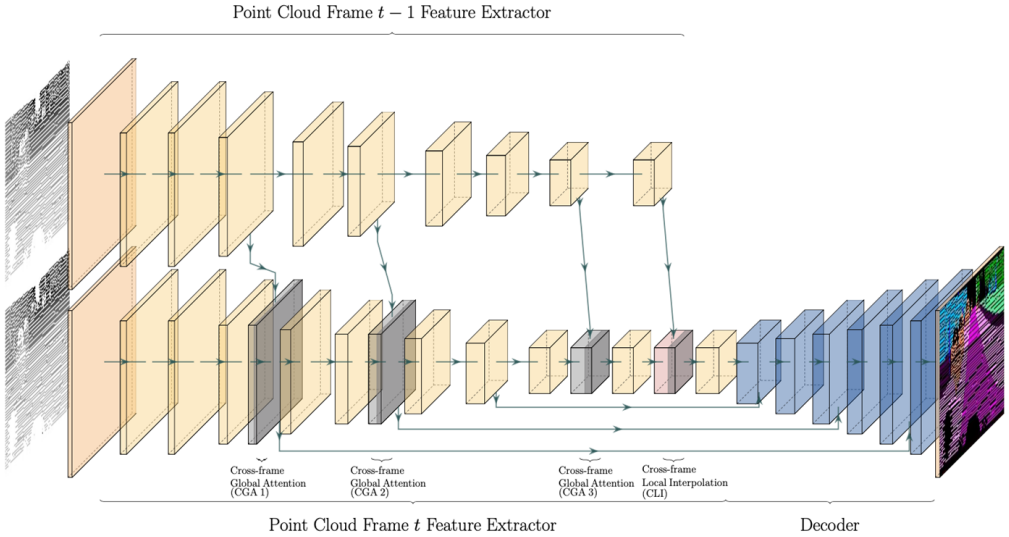



Figure 3. The overall architecture of our proposed network.

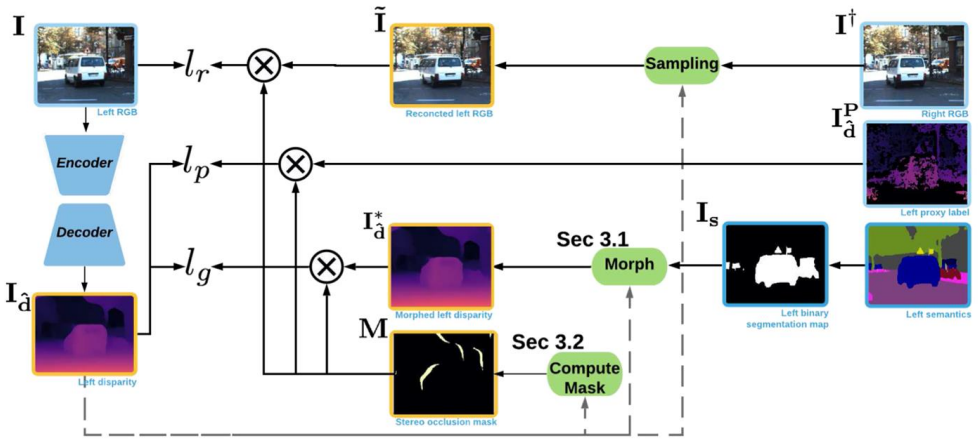
- Binary cross entropy loss, compensation-aware loss.
- CIM(Complementary Interaction Module) which uses the attention module and operates boundary enhancement

# SpSequenceNet: Semantic Segmentation Network on 4D Point Clouds (CVPR 2020)



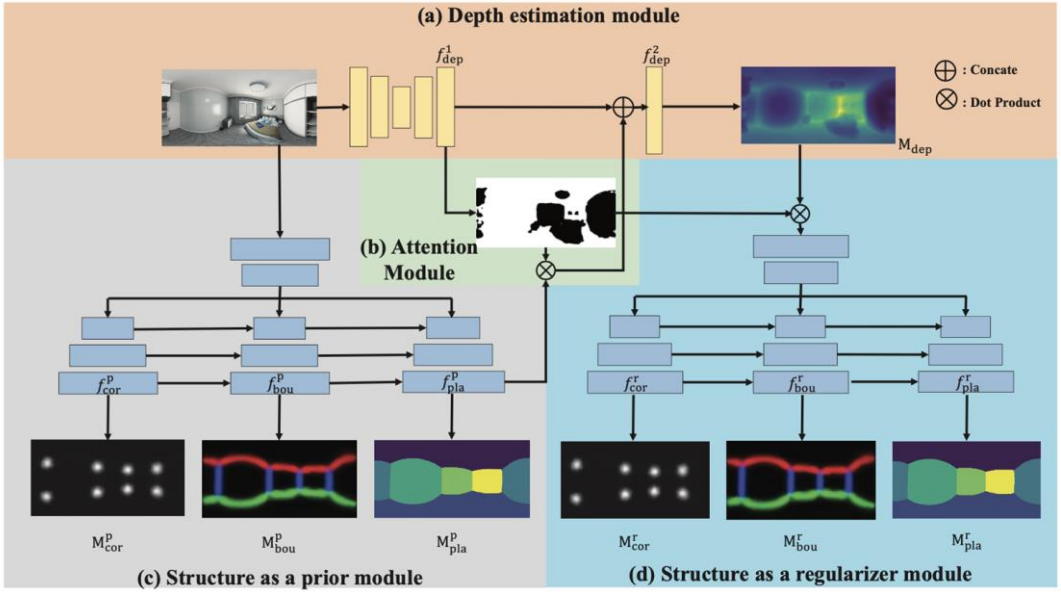

- Point clouds
- Cross-frame global attention, cross-frame local interpolation.
- Cross entropy loss

# The Edge of Depth: Explicit Constraint Between Segmentation and Depth (CVPR 2020)



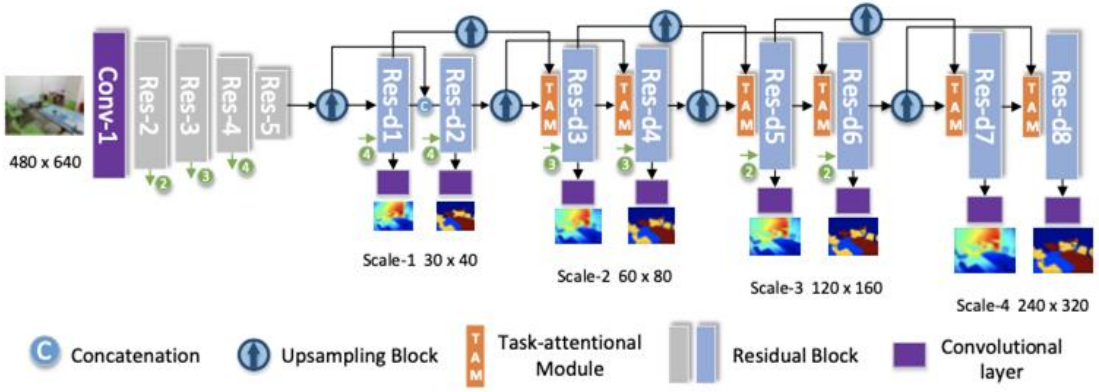

- Edge maps through 2d gradients from depth ground truth data and segmentation masks
- Photometric loss, morph loss, stereo matching proxy loss

# Geometric Structure Based and Regularized Depth Estimation From 360-Indoor Imagery (CVPR 2020)



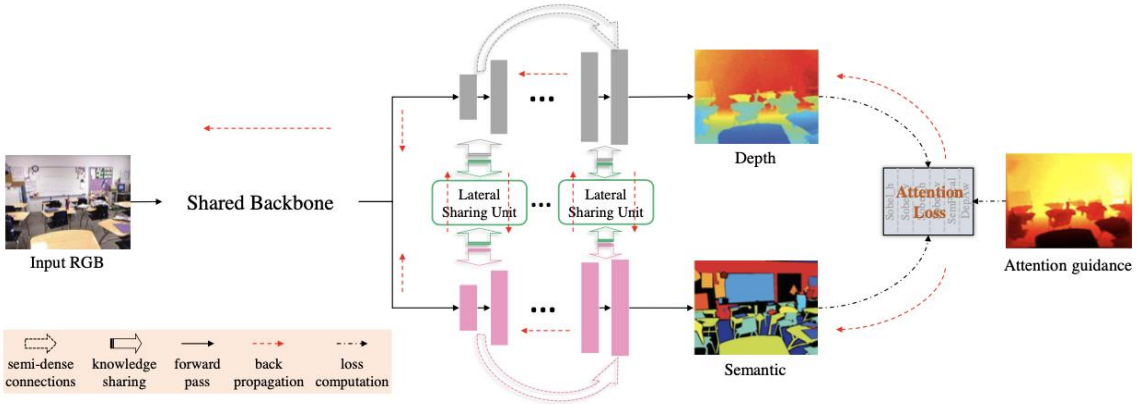

- Structure as prior and regularizer
- Attention module
- Refinement loss, photometric loss
- Make3D

# Joint Task-Recursive Learning for Semantic Segmentation and Depth Estimation (ECCV 2018)




- Joint learning
- Attention module
- Cross entropy loss, photometric loss

# Look Deeper Into Depth: Monocular Depth Estimation with Semantic Booster and Attention-Driven Loss (ECCV 2018)




- Attention based loss values
- NYU dataset

# Proposed Approach

- Dataset
- Loss functions
- Architecture
- Input/output dimensions
- Final loss

# Dataset

- Subset of KITTI Raw dataset (3949 training samples, 451 validation samples, 10 times less approximately)
- Eigen split
- Original resolutions are 1242x375 but during the training images are used as 640x192

# Conventional Reprojection Error

$$L_p = \sum_{t'} pe(I_t, I_{t' \to t}),$$

$$I_{t' \to t} = I_{t'} \left\langle proj(D_t, T_{t \to t'}, K) \right\rangle$$

$$pe(I_a, I_b) = \frac{\alpha}{2}(1 - \text{SSIM}(I_a, I_b)) + (1 - \alpha)\|I_a - I_b\|_1,$$

# Their Approach

$$L_p = \min_{t'} pe(I_t, I_{t' \to t}).$$

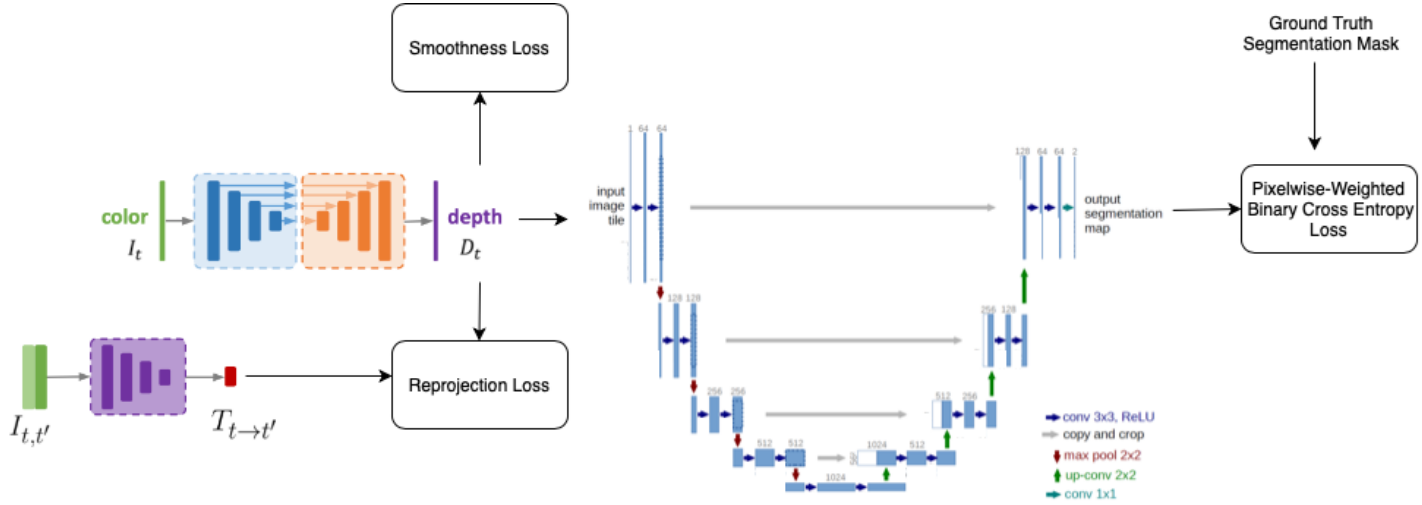
$$L_s = |\partial_x d_t^*| \, e^{-|\partial_x I_t|} + |\partial_y d_t^*| \, e^{-|\partial_y I_t|}$$

# U-Net Pixel-wise Weighted Cross Entropy Loss

$$w(\mathbf{x}) = w_c(\mathbf{x}) + w_0 \cdot \exp\left(-\frac{(d_1(\mathbf{x}) + d_2(\mathbf{x}))^2}{2\sigma^2}\right)$$

$$E = \sum_{\mathbf{x}\in\Omega} w(\mathbf{x}) \log(p_{\ell(\mathbf{x})}(\mathbf{x}))$$

# Final Network



# Final Loss

$$L = \mu L_p + \lambda L_s + \alpha L_b$$

$$\lambda = 1e-3, \alpha = 5e-4$$

# Training Procedure

- NVIDIA Geforce 1050 TI
- 7 hours appriximately

# Results

- Test data
- Evaluation metrics
- Comparison with the baseline network
- Comparison with state of the art networks on KITTI benchmark
- Failure cases

# Test Data

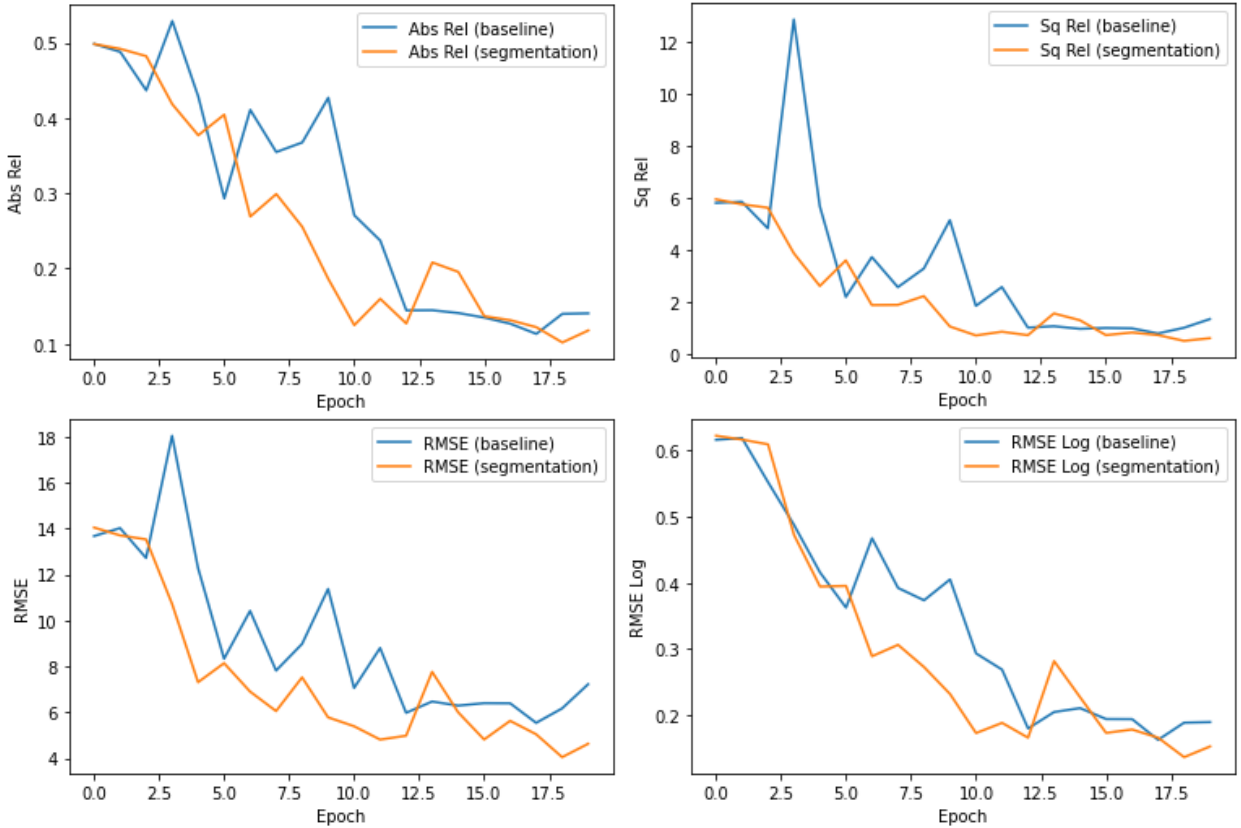
- KITTI Raw Eigen split test data (697 frames)

# Evaluation Metrics

$$\textbf{Abs Rel}: \frac{1}{|D|}\sum_{d\in D}|d^* - d|/d^* \quad \textbf{RMSE}: \sqrt{\frac{1}{|D|}\sum_{d\in D}||d^* - d||^2}$$

$$\textbf{Sq Rel}: \frac{1}{|D|}\sum_{d\in D}||d^* - d||^2/d^* \quad \textbf{RMSE log}: \sqrt{\frac{1}{|D|}\sum_{d\in D}||logd^* - logd||^2}$$

$$\delta_{\mathbf{t}}: \frac{1}{|D|}|\{d \in D|\, max(\frac{d^*}{d}, \frac{d}{d^*}) < 1.25^t\}| \times 100\%$$


Abs Rel (baseline)
Abs Rel (segmentation)
Abs Rel
Epoch
Sq Rel (baseline)
Sq Rel (segmentation)
Sq Rel
Epoch
RMSE (baseline)
RMSE (segmentation)
RMSE
Epoch
RMSE Log (baseline)
RMSE Log (segmentation)
RMSE Log
Epoch

| Method | Abs Rel | Sq Rel | RMSE | RMSE Log | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
|---|---|---|---|---|---|---|---|
| Zhou [Ref] | 0.183 | 1.595 | 6.709 | 0.270 | 0.734 | 0.902 | 0.959 |
| SfMLearner [Ref] | 0.208 | 1.768 | 6.958 | 0.283 | 0.678 | 0.885 | 0.957 |
| Vid2Depth [Ref] | 0.163 | 1.240 | 6.220 | 0.250 | 0.762 | 0.916 | 0.968 |
| DNC [Ref] | 0.182 | 1.481 | 6.501 | 0.267 | 0.725 | 0.906 | 0.963 |
| DualNet [Ref] | 0.121 | 0.837 | 4.945 | 0.197 | 0.853 | 0.955 | 0.982 |
| SuperDepth [Ref] | 0.116 | 1.055 | - | 0.209 | 0.853 | 0.948 | 0.977 |
| Yang | 0.182 | 1.481 | 6.501 | 0.267 | 0.725 | 0.906 | 0.963 |
| Mahjourian | 0.163 | 1.240 | 6.220 | 0.250 | 0.762 | 0.916 | 0.968 |
| GeoNet | 0.155 | 1.269 | 5.857 | 0.233 | 0.793 | 0.931 | 0.973 |
| DF-Net | 0.150 | 1.124 | 5.507 | 0.223 | 0.806 | 0.933 | 0.973 |
| LEGO | 0.162 | 1.352 | 6.276 | 0.252 | 0.783 | 0.921 | 0.969 |
| Struct2depth | 0.141 | 1.036 | 5.291 | 0.215 | 0.816 | 0.945 | 0.979 |
| Monodepth2 | 0.115 | 0.903 | 4.863 | 0.193 | 0.877 | 0.959 | 0.981 |
| Monodepth2 (Trained on subset) | 0.288 | 2.311 | 7.432 | 0.340 | 0.511 | 0.833 | 0.933 |
| Monodepth2 + Seg (Trained on subset) | 0.276 | 2.151 | 7.530 | 0.333 | 0.533 | 0.833 | 0.945 |

Table 1. Evaluation metric results for eigen split. Comparing to state of the art results, our network yields a poor performance. But, comparing the baseline network trained on the same subset, it gives better results.

# Conclusion

- The close relationship between segmentation and depth estimation
- Segmentation as a booster supervision signal
- More reliable segmentation masks are needed!
- More data is needed!

# References


[1] Gustavo Carneiro Adrian Johnston. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. *CVPR 2020*. 2

[2] Zhixiang Duan Chang Shu, Kun Yu and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. *ECCV 2020*. 2

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. 2018. 4

[4] Michael Firman Gabriel Brostow Clement Godard, Oisin Mac Aodha. 2

[5] Jianfei Cai Jianmin Zheng Jun Xiao Haiyong Jiang, Feilong Yan. End-to-end 3d point cloud instance segmentation without detection. *CVPR 2020*. 3

[6] Hao Wang Tzu-Yi HUNG Hanyu Shi, Guosheng Lin and Zhenhua Wang. Spsequencenet: Semantic segmentation network on 4d point clouds. *CVPR 2020*. 3

[7] Simon Hadfield Jaime Spencer, Richard Bowden. Defeatnet: General monocular depth via simultaneous unsupervised representation learning. *CVPR 2020*. 2

[8] Gabriel J. Brostow-Daniyar Turmukhambetov Jaime Watson, Michael Firman. Self-supervised monocular depth hints. *ICCV 2019*. 2

[9] Yibing Song Rynson Lau Jianbo Jiao, Ying Cao. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. *ECCV 2018*. 3

[10] Yuwang Wang Kaihuai Qin Junsheng Zhou and Wenjun Zeng. Unsupervised high-resolution depth learning from videos with dual networks. *ICCV 2019*. 2

[11] Oliver Wang Long Mai Zhe Lin Ke Xian, Jianming Zhang and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. *CVPR 2020*. 3

[12] Jia Zheng Jingyi Yu Junfei Zhang Rui Tang Shugong Xu Shenghua Gao Lei Jin, Yanyu Xu. Geometric structure based and regularized depth estimation from 360◦ indoor imagery. *CVPR 2020*. 3

[13] Shaoshuai Shi Shu Liu Chi-Wing Fu Jiaya Jia Li Jiang, Hengshuang Zhao. Pointgroup: Dual-set point grouping for 3d instance segmentation. *CVPR 2020*. 3

[14] Oliver Wang Zhe Lin Lijun Wang, Jianming Zhang and Huchuan Lu. Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. *CVPR 2020*. 2

[15] Yifan Wang Huchuan Lu Lijun Wang, Jianming Zhang and Xiang Ruan. Cliffnet for monocular depth estimation with hierarchical embedding loss. *ECCV 2020*. 2

[16] Yongri Piao Zhengkun Rong Huchuan Lu Miao Zhang, Weisong Ren. Select, supplement and focus for rgb-d saliency detection. *CVPR 2020*. 3

[17] Thomas Brox Olaf Ronneberger, Philipp Fischer. U-net: Convolutional networks for biomedical image segmentation. *MICCAI 2015*. 3

[18] Xiaoming Liu Shengjie Zhu, Garrick Brazil. The edge of depth: Explicit constraints between segmentation and depth. *CVPR 2020*. 3

[19] Chunyan Xu Zequn Jie Xiang Li Jian Yang Zhenyu Zhang, Zhen Cui. Joint task-recursive learning for semantic segmentation and depth estimation. *ECCV 2018*. 3