Deep Classification Network for Monocular Depth Estimation

Azeez Oluwafemi 1,2, Yang Zou 2, B.V.K. Vijaya Kumar 2
1.InstaDeep, 2. Carnegie Mellon University

Carnegie Mellon University

InstaDeep™

ABSTRACT

Monocular Depth Estimation is usually treated as a supervised and regression problem when it actually is very similar to semantic segmentation task since they both are fundamentally pixel-level classification tasks. We applied depth increments that increases with depth in discretizing depth values and then applied Deeplab v2 [1] and the result was higher accuracy. We were able to achieve a state-of-the-art result on the KITTI dataset [2] and outperformed existing architecture by an 8% margin

RESULT

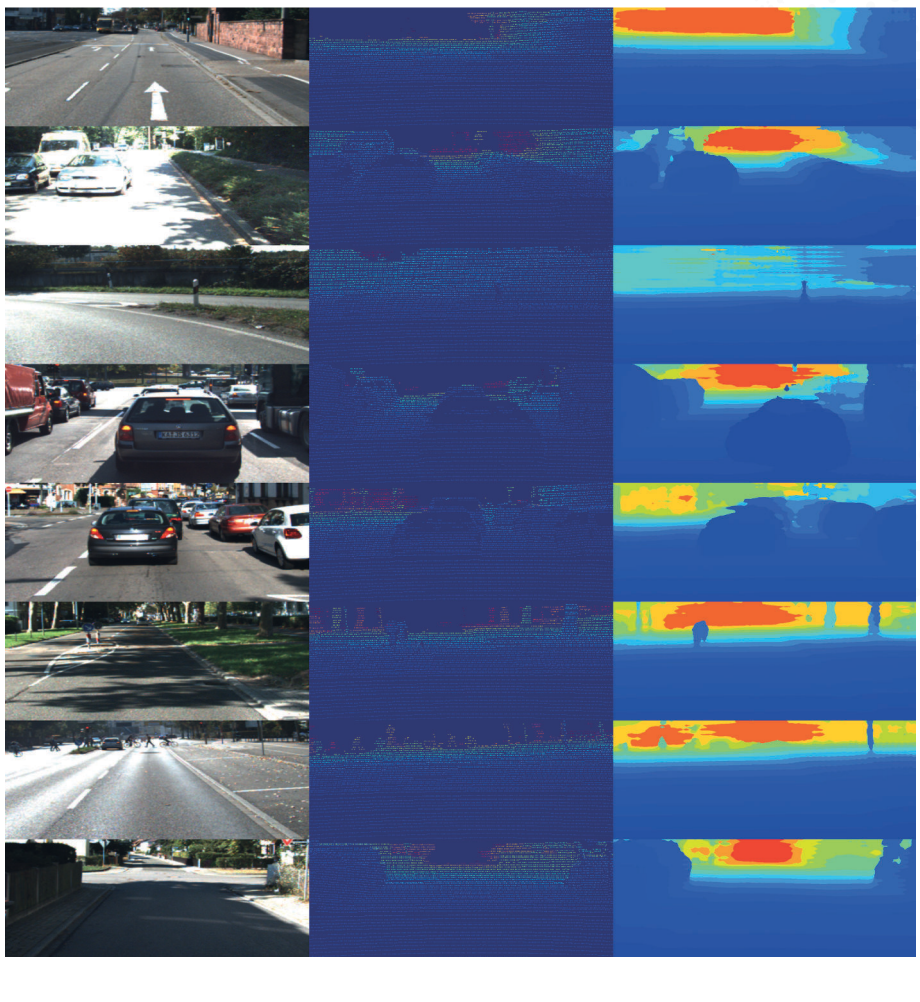

FIGURE 1: Depth Predictions on KITTI}. The Image on the left column, ground truth in the middle and our model prediction on the right. The color bar below shows the range of depths. Blue is nearer.

METHODS

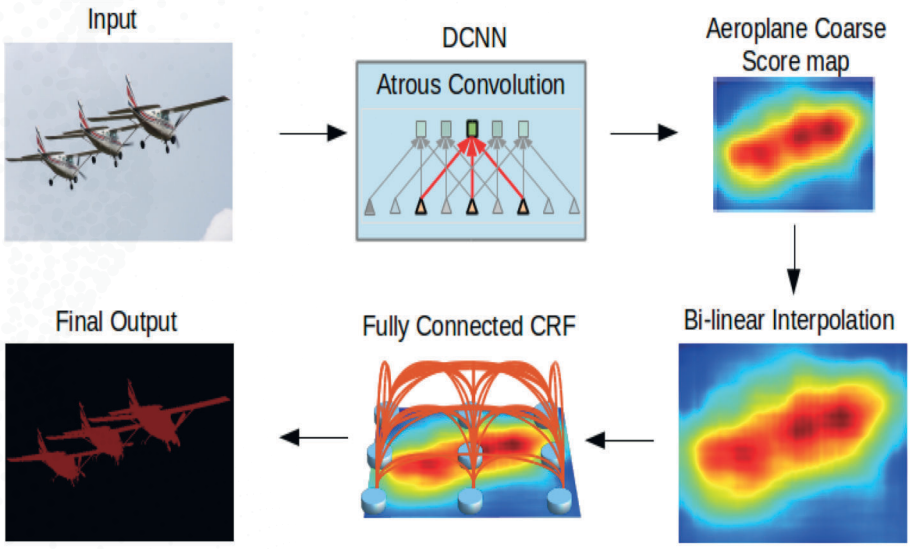

FIGURE 2: Deeplab [1] . An input image is passed through a Deep Convolutional Neural Network(DCNN) such as ResNet101, using atrous convolution to reduce downsampling. The score map output is then interpolated for upsampling to original image resolution. Low-level details are finally incorporated with the final result through a pre-processing step of fully connected conditional random field (CRF)

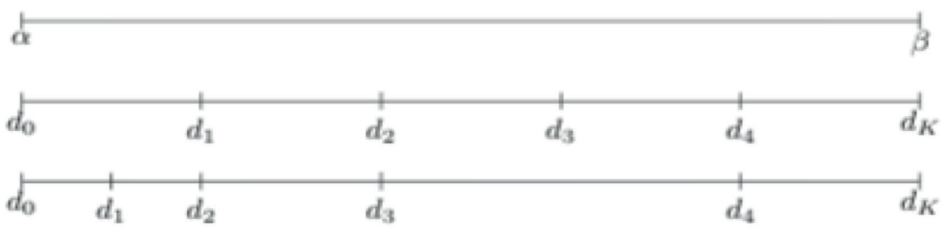

FIGURE 3: UD (middle) and SID (bottom) to discretize depth interval [a, b] into sub intervals di where i = 0, 1,2,...K and K = number of class. UD is Uniform Discretization, SID is spatially increasing interval discretization

$$d_i = \exp(\log\alpha + \frac{\log\beta/\alpha * i}{K}), i = 0, ..., K \quad (1)$$

where di in d0, d1,..., dK are depth interval boundaries.

BENCHMARK METRICS

Absolute Relative Error,

$$absRel = \frac{1}{T}\sum_p \frac{|d_p - \hat{d}_p|}{d_p} \quad (2)$$

Root Mean Square Error,

$$RMSE = \sqrt{\frac{1}{T}\sum_p (d_p - \hat{d}_p)^2} \quad (3)$$

log scale invariant Root Mean Square Error ,

$$RMSE_{log} = \frac{1}{T}\sum_p (\log\hat{d}_p - \log d_p + \alpha(\hat{d}_p, d_p))^2 \quad (4)$$

Accuracy under a threshold

$$\delta < th = max(\frac{\hat{d}_p}{d_p}, \frac{d_p}{\hat{d}_p}) \quad (5)$$

BENCHMARK PERFORMANCE

	Higher is better			Lower is better		
Method	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Abs Rel	RMSE	$RMSE_{log}$
Make3D	0.601	0.820	0.926	0.280	8.734	0.361
Eigen	0.692	0.899	0.967	0.190	7.156	0.270
Liu LRC (CS + K)	0.861	.949	0.976	0.114	4.935	0.206
Kuznietsov	0.862	0.960	0.986	.113	4.621	0.189
DORN (VGG)	0.915	0.980	0.993	0.081	3.056	0.132
DORN (ResNet)	**0.932**	0.984	0.994	**0.072**	2.727	**0.120**
Ours(ResNet)	0.796	**0.985**	**1.000**	0.075	**2.499**	0.156

Table 1: Performance on KITTI. K is KITTI, CS is Cityscapes. 1.25, 1.252 and 1.253 are pre-defined thresholds for Accuracy under threshold metric (d)

FURTHER RESEARCH

In the future, we would investigate domain gaps in monocular depth maps estimation and apply class balanced self training to attempt to reduce the gap.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4):834–848, 2018.

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. The International Journal of Robotics Research, 32(11):1231–1237, 2013.

Conclusion

We've been able to demonstrate how a depth estimation task can be formulated as a semantic segmentation problem since the two tasks are fundamentally just pixel-level classification tasks at the core. Simply discretizing the depth map and treating the binned depths as classes and applying a state-of-the-art semantic segmentation network can produce a result that outperforms existing results.