Deep RGB-D Canonical Correlation Analysis For Sparse Depth Completion

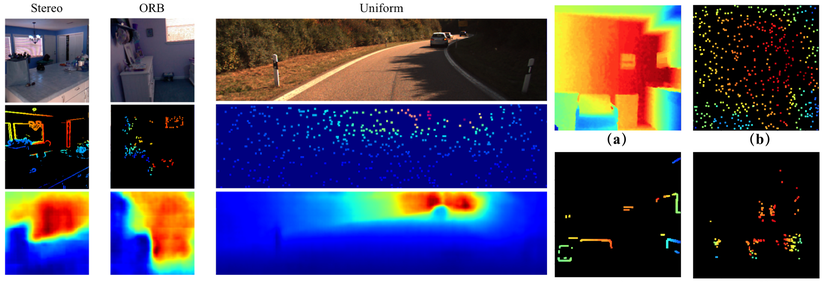
Yiqi Zhong*, Cho-Ying Wu*, Suya You, Ulrich Neumann (*Equal Contribution)

USC Viterbi School of Engineering

Task and Challenges

Task Introduction:

We work on depth completion using representations from both RGB images and sparse depth maps from different sensors.



The first, second, and third rows are images, sparse depths, and completed depths.

Sparsifiers example:
(a) Depth (b) Uniform
(c) Stereo (D) ORB

Previous methods work on **depth predictions** rather than **depth completion**, since they did not fully utilize the relationship between sparse depth and images.

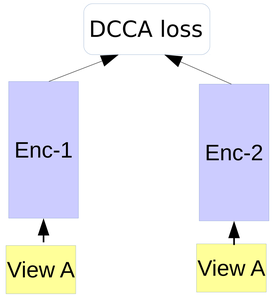
We use DCCA loss to ensure the semantically correlated features are captured between two modalities (depth and imagery).

(CCA)

$$(w_1^*, w_2^*) = \underset{w_1,w_2}{\operatorname{argmax}} \ \operatorname{corr}(w_1' X_1, w_2' X_2)$$

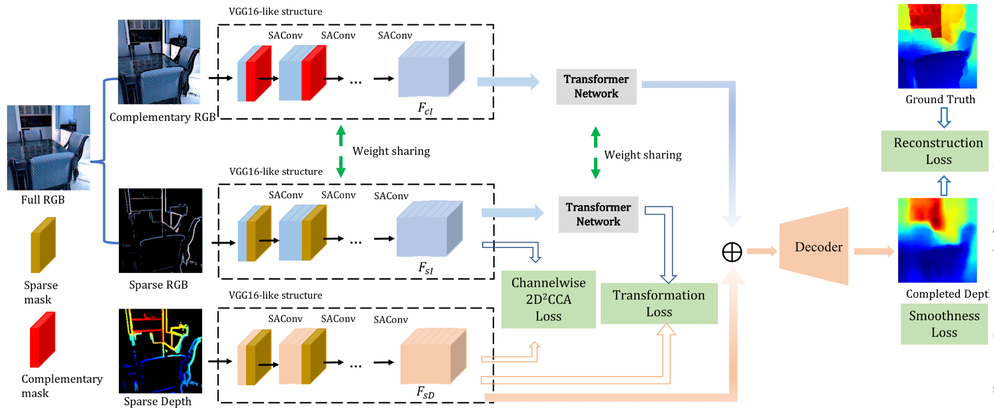
$$= \underset{w_1,w_2}{\operatorname{argmax}} \ \frac{w_1' \Sigma_{12} w_2}{\sqrt{w_1' \Sigma_{11} w_1 w_2' \Sigma_{22} w_2}}.$$

(DCCA)



Model and Objective

Proposed CFCNet



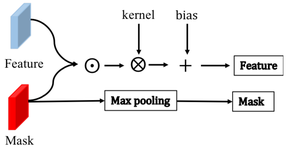
(1) Backbone encoder trained with 2D^2CCA loss using sparse RGB and sparse depth inputs. 2D^2CCA let the encoders capture *semantically correlated features.* Traditional CCA suffers from high-dimensional space with small sample size (SSS). Therefore, we calculate covariances in 2D matrix formulations.

$$\mathbf{corr}(F_{sD}, F_{sI}) = \left\| (\hat{\Sigma}_{sD}^{-\frac{1}{2}})(\hat{\Sigma}_{sD,sI})(\hat{\Sigma}_{sI}^{-\frac{1}{2}}) \right\|_{\mathbf{tr}}. \quad \hat{\Sigma}_{sD} = \frac{1}{C}\sum_{i=0}^{C-1} \left[F_{sD}^i - \mathbf{E}\left[F_{sD}\right]\right]\left[F_{sD}^i - \mathbf{E}\left[F_{sD}\right]\right]^T + r_I \mathbf{I}.$$

$$\hat{\Sigma}_{sD,sI} = \frac{1}{C}\sum_{i=0}^{C-1} \left[F_{sD}^i - \mathbf{E}\left[F_{sD}\right]\right]\left[F_{sI}^i - \mathbf{E}\left[F_{sI}\right]\right]^T,$$

Backprop

$$\frac{\partial \mathbf{corr}(F_{sD}, F_{sI})}{\partial F_{sI}} = \frac{1}{C}(2\nabla_{sDsD} F_{sD} + \nabla_{sDsI} F_{sI}), \quad \nabla_{sDsD} = -\tfrac{1}{2}\hat{\Sigma}_{sD}^{-\frac{1}{2}} U D U^T \hat{\Sigma}_{sD}^{-\frac{1}{2}} \quad \nabla_{sDsI} = \hat{\Sigma}_{sD}^{-\frac{1}{2}} U V^T \hat{\Sigma}_{sRGB}^{-\frac{1}{2}}$$



Sparsity-aware Attentional Convolutions (SAConv)

(2) Transformer network to ensure calibration between ranges of RGB and depth domains.

$$L_{trans} = \|\hat{F_{sD}} - \hat{F}_{sD}\|_2^2$$

(3) Recovered loss

$$L_{recon} = \|D_{gt} - \hat{D}\|_2^2$$

(4) Smooth

$$L_{smooth} = \|\nabla^2 \tilde{D}\|_1$$

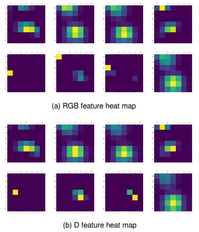
Results and Evaluations

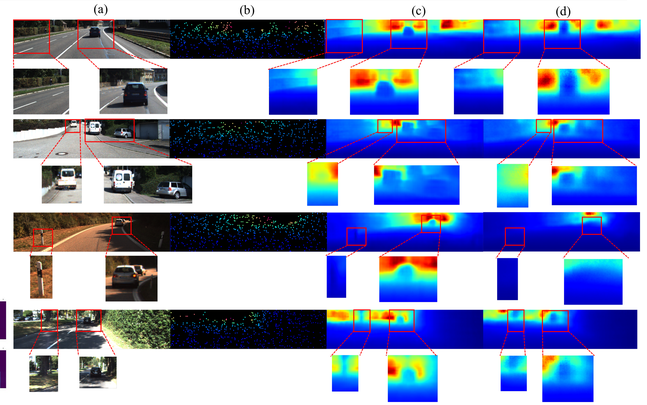
SLAM-RGBD

Sequence#	CFCNet	CNN-SLAM[1]	Laina[47]	Remode[48]
ICL/office0	**41.97**	19.41	17.19	4.47
ICL/office1	**43.86**	29.15	20.83	3.13
ICL/office2	**63.64**	37.22	30.63	16.70
ICL/living0	**51.76**	12.84	15.00	4.47
ICL/living1	**64.34**	13.03	11.44	2.42
ICL/living2	**59.07**	26.56	33.01	8.68
TUM/seq1	**54.70**	12.47	12.98	9.54
TUM/seq2	**66.30**	24.07	15.41	12.65
TUM/seq3	**74.61**	27.39	9.45	6.73
Average	**57.81**	22.46	18.44	7.64

KITTI 500 points w/ KITTI.

Methods	Input	Dataset	RMSE	δ1	δ2	δ3
Zhou *et al.* [36]	RGB	CS→K	7.580	57.7	84.0	93.7
Godard *et al.* [45]	RGB	CS→K	14.445	5.3	32.6	86.2
Aleotti *et al.* [46]	RGB	CS→K	14.051	6.3	39.4	87.6
CFCNet(50 pts)	RGB+sD	CS →K	7.841	78.3	92.7	97.0
CFCNet(100 pts)	RGB+sD	CS→K	**5.827**	**82.6**	**94.7**	**97.9**
Zhou *et al.* [36](cap 50m)	RGB	CS→K	6.148	59.0	85.2	94.5
CFCNet(50 pts, cap 50m)	RGB+sD	CS →K	6.334	79.2	93.2	97.3
CFCNet(100 pts, cap 50m)	RGB+sD	CS →K	**4.524**	**83.7**	**95.2**	**98.1**
CFCNet(50 pts, cap 50m)	RGB+sD	CS→CS	9.019	82.8	94.1	97.2
CFCNet(100 pts, cap 50m)	RGB+sD	CS→CS	6.887	88.9	96.1	98.1
CFCNet(100 pts, cap 50m)	RGB+sD	K→K	3.157	91.0	97.1	98.9





(a) Image (b) Sparse depth (c) CFCNet results (d) Sparse-to-dense method