UNIVERSITÀ DEGLI STUDI DI PADOVA

Enhancing Deep Semantic Segmentation of RGB-D data with Entangled Forest

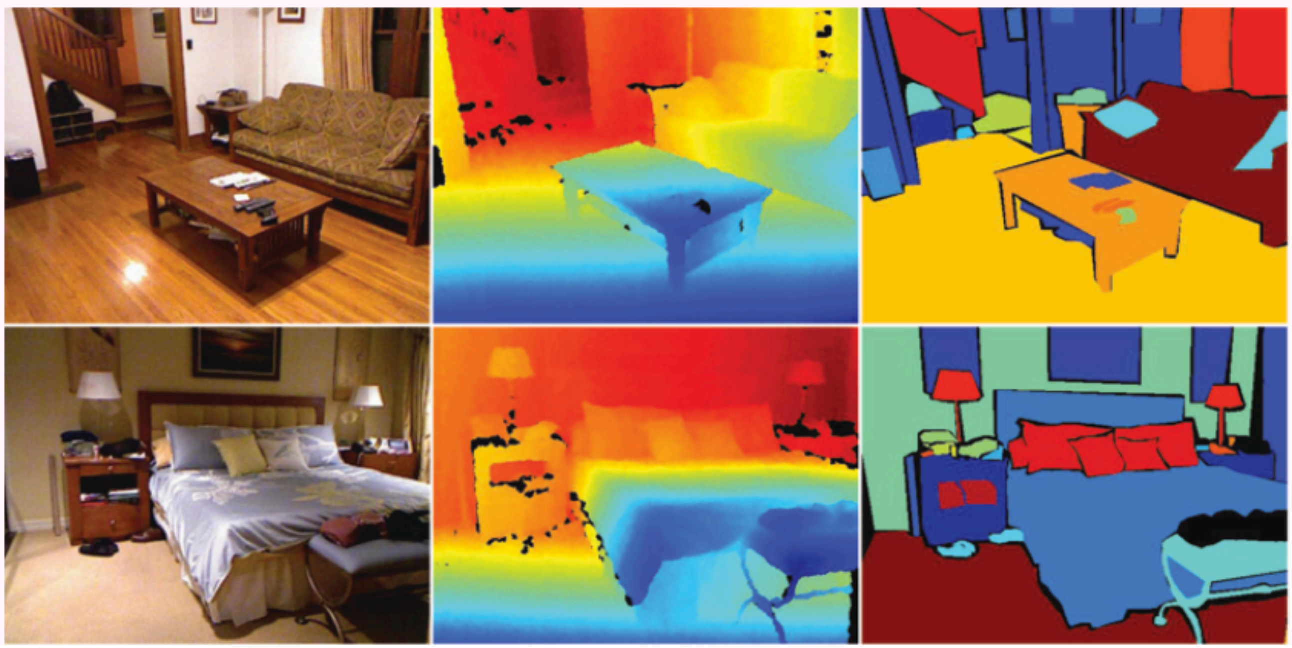
Matteo Terreran, Elia Bonetto, Stefano Ghidoni

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

INTELLIGENT AUTONOMOUS SYSTEMS LAB

Semantic segmentation

Semantic segmentation consists of a **pixel-wise classification** of the input data according to a given set of classes of interest (e.g. floor, table, bed)



(a) RGB (b) Depth (c) Labels

Contributions

Deep learning (DL) architectures

- high-end GPUs
- end-to-end training
- large training data

3D Entangled Forest classifier (3DEF)

- fast and runs on CPU
- hand-crafted features + classification
- less training data than DL

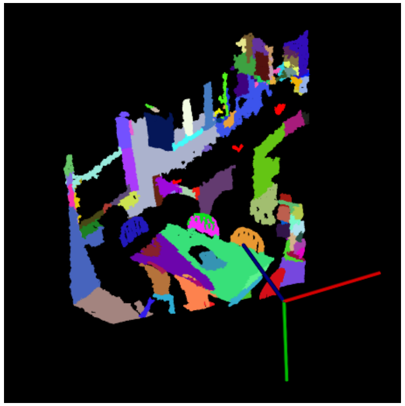
In this work we investigate how to combine the two approaches:

- Analysis of **different configurations to embed hand-crafted features** in a deep learning networks such as FuseNet;
- Study of the **potential of 3DEF hand-crafted features** in DL networks;
- **Possibility to shrink a DL network** without reducing performance by exploiting additional information.

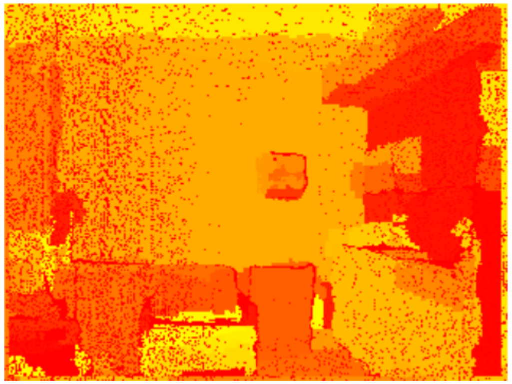
3D Entangled Forest

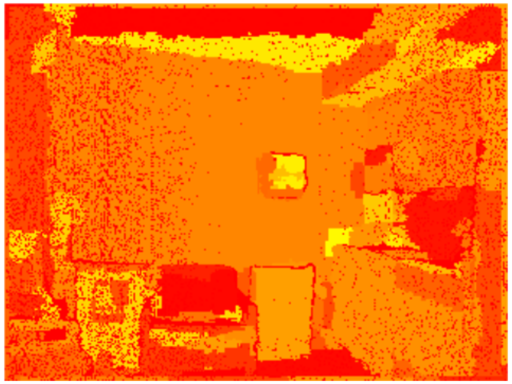
3DEF classifier relies on a preliminary over-segmentation in clusters and the computation of two types of hand-crafted features:

- **Unary features**, describe cluster information (e.g. color);
- **Entangled features**, describe relations between clusters to consider also information in the space domain.

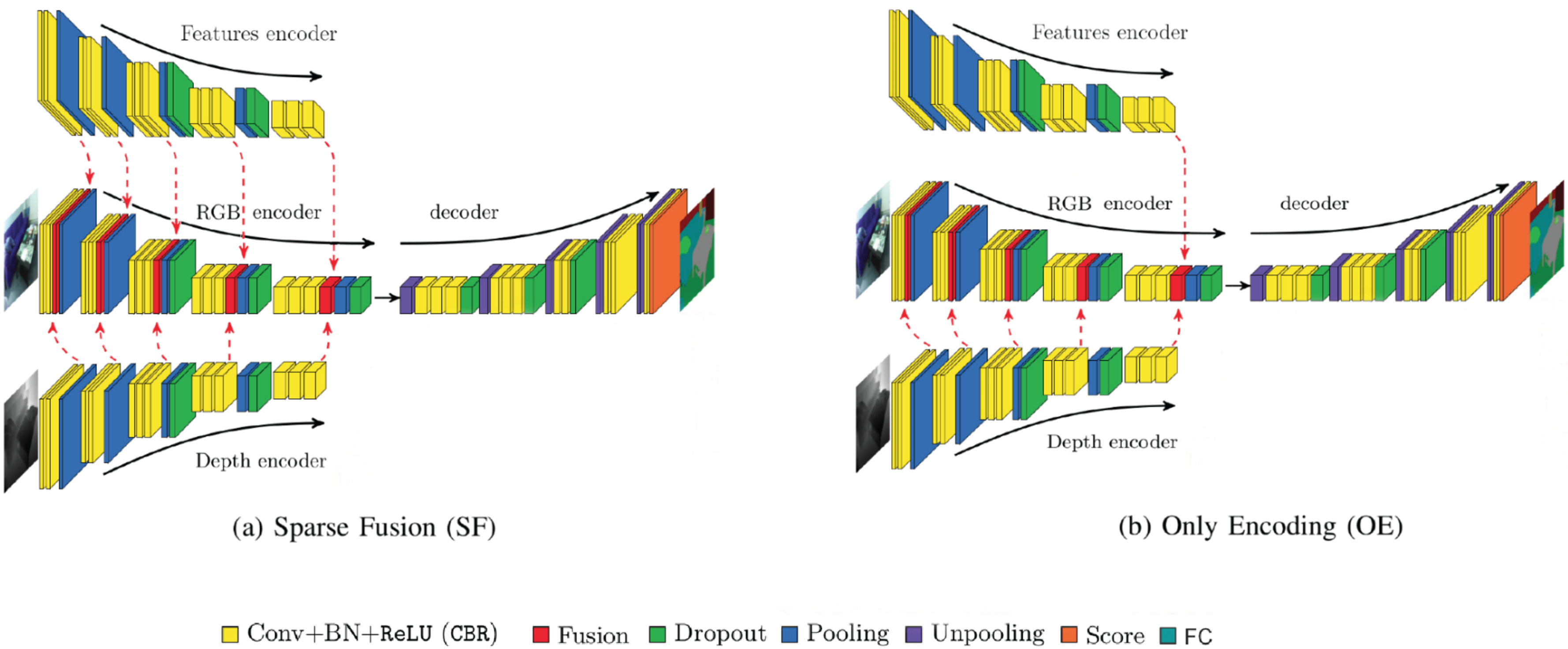


3DEF features represent cluster information, not pixels as in FuseNet architecture; to obtain a similar representation we redefined 3DEF features in a 2D space





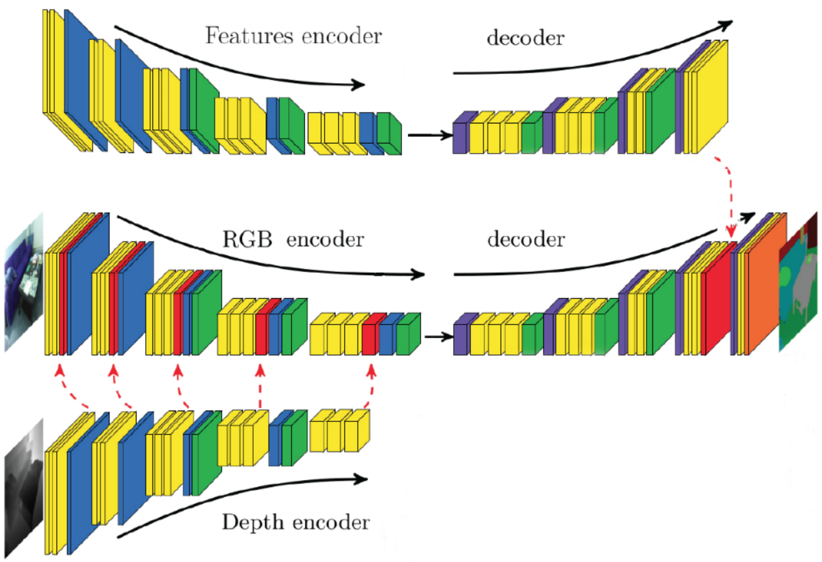
Configurations



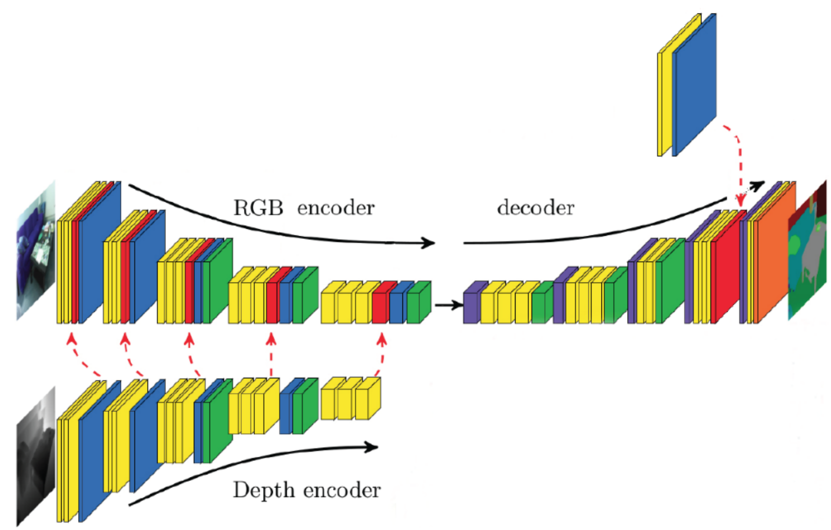
(a) Sparse Fusion (SF)

(b) Only Encoding (OE)

Configurations



(c) Parallel Autoencoder (PA)



(d) Direct Fusion (DF)

Conv+BN+ReLU (CBR) Fusion Dropout Pooling Unpooling Score FC

Experimental results

Experiments on NYU Depth v2 dataset with 13 and 40 classes mapping:

- Slightly better performance than FuseNet using 3DEF features
- Improved loss function convergence

Configuration		13-classes Global acc.	13-classes Mean acc.	13-classes IoU	40-classes Global acc.	40-classes Mean acc.	40-classes IoU
SF	UN	76.44	67.52	54.93	68.63	47.08	35.72
	ENT	75.19	65.28	52.83	66.69	42.82	32.95
	ENT+UN	76.86	67.63	55.29	68.60	47.95	36.28
OE	UN	76.29	66.81	54.36	68.76	47.24	35.82
	ENT	76.91	67.61	55.21	68.16	44.65	34.05
	ENT+UN	76.56	67.73	55.03	68.76	46.67	35.54
PA	UN	76.50	67.02	54.74	**69.02**	46.27	36.10
	ENT	76.16	67.31	54.06	67.35	42.88	33.24
	ENT+UN	76.77	67.75	55.27	68.77	47.06	35.89
DF	UN	76.61	67.59	54.94	68.72	46.86	35.86
	ENT	76.17	67.26	54.14	68.13	44.71	34.71
	ENT+UN	**76.92**	**68.21**	**55.49**	69.00	**47.64**	**36.62**
FuseNet		76.40	66.93	54.74	68.76	46.42	35.48

Experimental results

- Larger improvements considering reduced configurations

Configuration		13-classes Global acc.	13-classes Mean acc.	13-classes IoU	40-classes Global acc.	40-classes Mean acc.	40-classes IoU
SF	UN	67.92	56.95	43.73	58.73	32.11	23.02
	ENT	**71.20**	**60.35**	**47.79**	**62.28**	**36.84**	**26.51**
	ENT+UN	67.83	56.66	43.59	59.04	33.83	24.00
OE	UN	67.77	56.52	43.46	59.03	32.77	23.34
	ENT	70.11	58.42	46.12	60.78	34.44	24.79
	ENT+UN	68.03	57.06	43.80	59.05	33.13	23.71
PA	UN	67.82	56.31	43.38	58.95	32.68	23.30
	ENT	68.38	55.89	43.90	59.68	33.59	23.82
	ENT+UN	68.10	57.11	43.87	59.12	33.14	23.53
DF	UN	67.82	56.36	43.47	59.02	32.71	23.40
	ENT	68.02	54.98	43.10	59.52	32.91	23.44
	ENT+UN	67.76	56.62	43.43	59.31	32.69	23.50
FuseNet reduced		68.01	56.50	43.70	59.02	33.00	23.42

Conclusions

- We investigate how to integrate hand-crafted features in a deep learning model proposing different fusion strategies;
- Larger improvements when considering reduced configurations;
- Entangled features can be used to obtain lighter state-of-the-art models, with lower computational cost.

Network	Original		Reduced	
	Param	Epoch time	Param	Epoch time
Sparse Fusion	58.908	125	6.968	83
Only Encoding	58.908	125	6.968	83
Parallel AE	73.591	143	8.666	94
Direct Fusion	44.224	104	5.269	70
FuseNet	44.173	90	5.218	60

Thanks for your attention