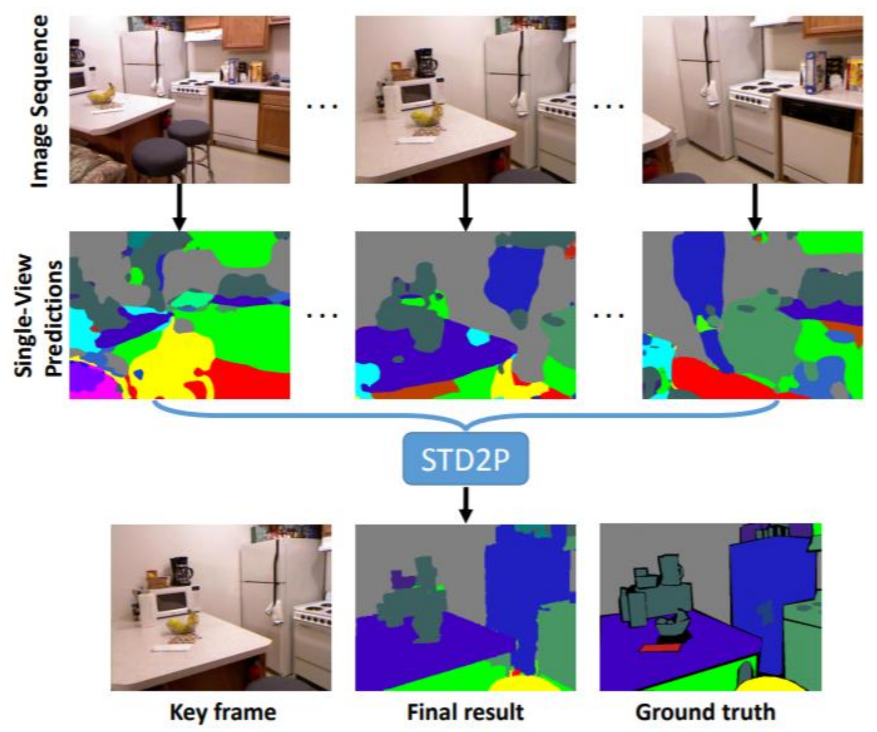
STD2P: RGBD Semantic Segmentation Using Spatio-Temporal Data-Driven Pooling

Yang He[1], Wei-Chen Chiu[1], Margret Keuper[2], Mario Fritz[1]

[1] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany [2] University of Mannheim, Germany

Motivation

- Rich information from videos
- Accurate boundary and data-driven receptive field from superpixels
- Robust region correspondence from superpixels
- Leverage large-scale unlabeled data



Contribution

- ***Multi-view region-based neural network*** for semantic segmentation
- ***Semi-supervised learning*** from partially labeled video
- ***State-of-the-art results*** on various datasets
- ***New layer*** compatible with existing architectures.

Please visit our project page for ***download***:

Previous work

- Comparison to bilateral inceptions [1]: ***faster computation and less memory cost*** with pooling operations.
- Comparison to region-based semantic segmentation with end-to-end training [2]: ***temporal pooling*** allows us to utilize unlabeled frames as well as better prediction.
- Comparison to SemanticFusion [3]: our network is ***trained with multiple-frame input*** and their correspondences instead of training a single frame prediction model and a fusion model separately.

Method

- Region Correspondence

  Superpixel: RGBD MCG [4]

  Optical flow: EpicFlow [5]

  a matching rejection scheme:

$$\min\left(\overrightarrow{IoU_{tu}}, \overleftarrow{IoU_{tu}}\right) > \tau$$

- Spatial pooling

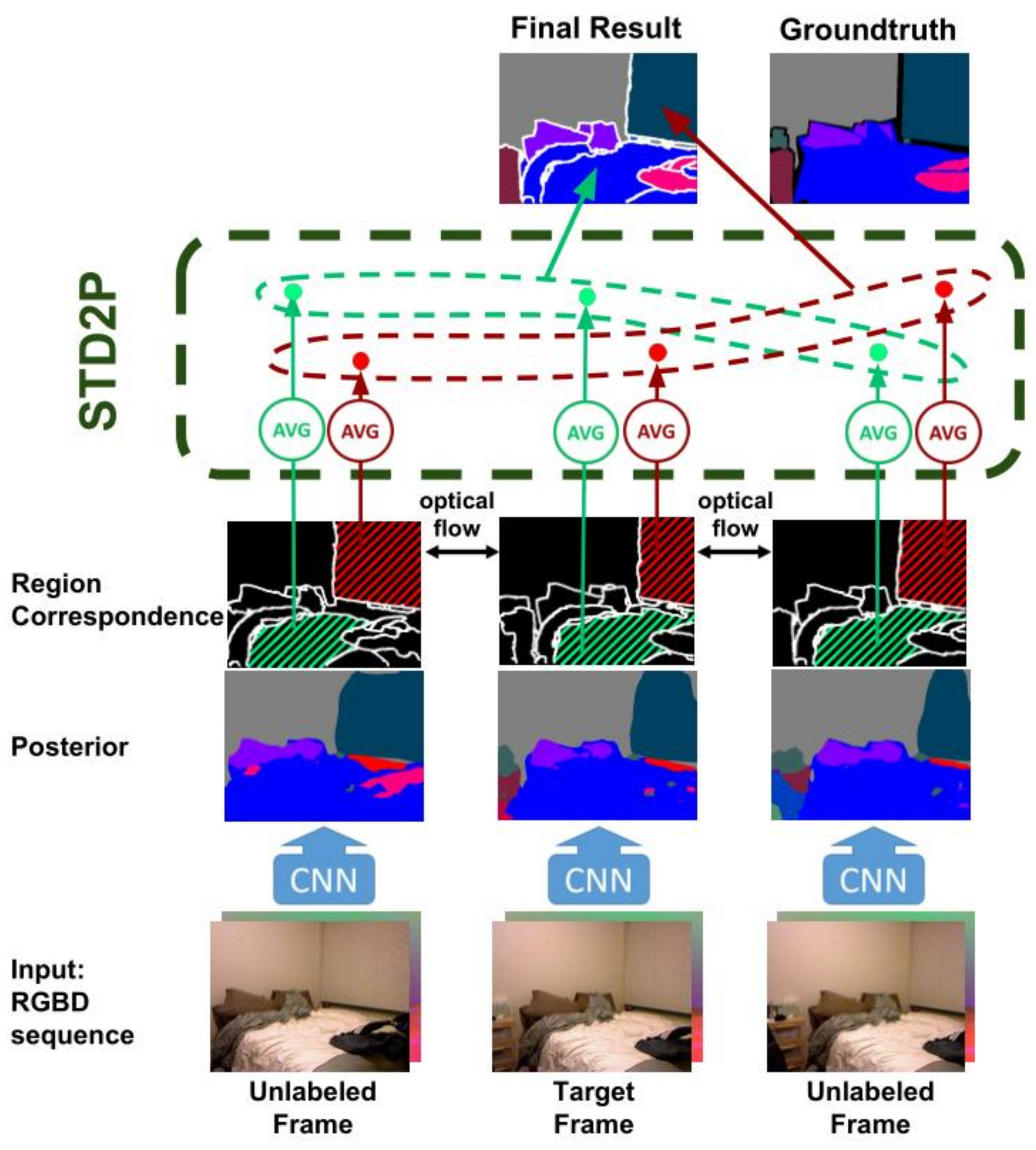
$$O_s(i,c,j) = \frac{1}{|\Omega_{ij}|} \sum_{(x,y)\in\Omega_{ij}} I_s(i,c,x,y)$$

- Temporal pooling

$$O_t(c,j) = \frac{1}{K} \sum_{\Omega_{ij}\neq\emptyset} I_t(i,c,j)$$

- Region-to-pixel mapping

$$O_r(c,x,y) = I_r(c,j), \quad S_{tar}(x,y) = j$$



Evaluation settings

- Datasets and tasks:

  NYUDv2 40-class, 13-class and 4-class tasks,

  SUN3D 33-class task.

- Four evaluation metrics:

  Pixel Acc., Mean Acc., Mean IoU, f.w. IoU.

- The different settings about leveraging temporal information:

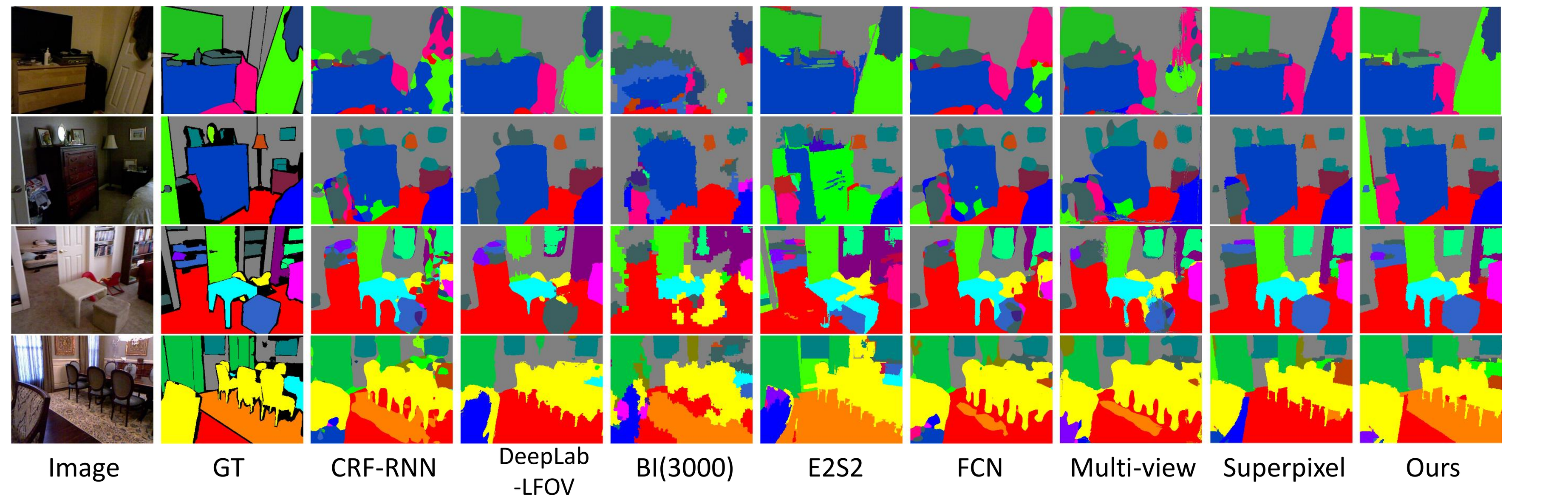
Model	Training	Test
Superpixel model	×	×
Superpixel+ model	✓	×
Full model	✓	✓

References

[1] R. Gadde *et al.*, Superpixel Convolutional Networks using Bilateral Inceptions, ECCV, 2016.
[2] H. Caesar *et al.*, Region-based semantic segmentation with end-to-end training, ECCV, 2016.
[3] J. McCormac et al., SemanticFusion: Dense 3D Semantic Mapping with Convolutional Neural Networks. ICRA, 2017.
[4] S. Gupta et al., Learning Rich Features from RGB-D Images for Object Detection and Segmentation, ECCV, 2014.
[5] J. Revaud et al., EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow, CVPR, 2015.

Results

Qualitative results on NYUDv2 40-class task



- Comparison to state-of-the-art methods on NYUDv2 40-class task

Methods	Pixel Acc.	Mean Acc.	Mean IoU	f.w. IoU
Mutex Constraints	63.8	-	31.5	48.5
RGBD R-CNN	60.3	-	28.6	47.0
Bayesian SegNet	68.0	45.8	32.4	-
Multi-Scale CNN	65.6	45.1	34.1	51.4
CRF-RNN	66.3	48.9	35.4	51.0
DeepLab	68.7	46.9	36.8	52.5
DeepLab-LFOV	**70.3**	49.6	39.4	54.7
BI (1000)	57.7	37.8	27.1	41.9
BI (3000)	58.9	39.3	27.7	43.0
E2S2	58.1	52.9	31.0	44.2
FCN	65.4	46.1	34.0	49.5
Ours (*superpixel*)	68.5	48.7	36.0	52.9
Ours (*superpixel+*)	68.4	52.1	38.1	54.0
Ours (*full model*)	70.1	**53.8**	**40.1**	**55.7**

- Results on SUN3D dataset

Methods	Pixel Acc.	Mean Acc.	Mean IoU	f.w. IoU
Mutex Constraints	**65.7**	-	28.2	51.0
CRF-RNN	59.8	-	25.5	43.3
DeepLab	60.9	30.7	24.0	44.1
DeepLab-LFOV	62.3	35.3	28.2	46.2
BI (1000)	53.8	31.1	20.8	37.1
BI (3000)	53.9	31.6	21.1	37.4
E2S2	56.7	**47.7**	27.2	43.3
FCN	58.8	38.5	26.1	43.9
Ours (*superpixel+*)	62.5	40.8	29.4	47.8
Ours (*full model*)	65.5	41.2	**32.9**	**51.5**

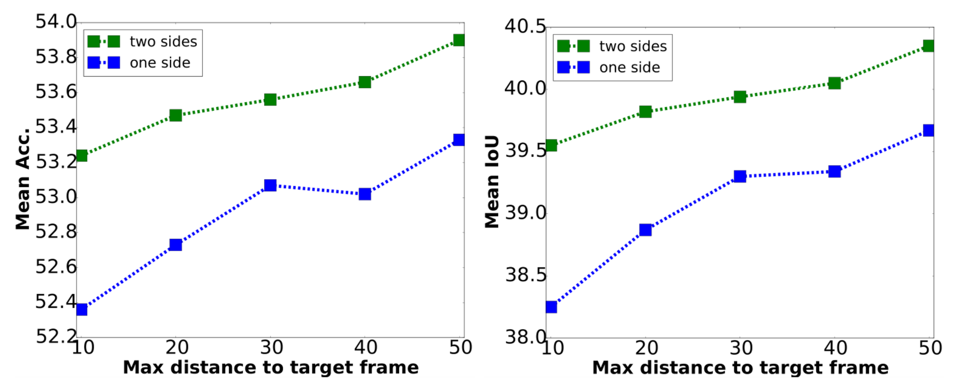
- Average vs. Max

Spatial/Temporal	Pixel Acc.	Mean Acc.	Mean IoU	f.w. IoU
AVG / AVG	**70.1**	**53.8**	**40.1**	**55.7**
AVG / MAX	69.4	51.0	38.0	54.4
MAX / AVG	66.4	45.4	33.8	49.6
MAX / MAX	64.9	44.5	32.1	47.9

- Region based vs. Pixel based

Methods	Pixel Acc.	Mean Acc.	Mean IoU	f.w. IoU
FCN	65.4	46.1	34.0	49.5
Pixel Correspondence	66.2	45.9	34.6	50.2
Superpixel Correspondence	**70.1**	**53.8**	**40.1**	**55.7**

- Wider is better



- Comparison to multi-view methods on NYUDv2 4-class and 13-class tasks

Methods	Pixel Acc.	Mean Acc.	Pixel Acc.	Mean Acc.
Couprie *et al.*	64.5	63.5	52.4	36.2
Hermans *et al.*	69.0	68.1	54.2	48.0
Stückler *et al.*	70.6	66.8	-	-
McCormac *et al.*	-	-	69.9	63.6
Wang *et al.*	-	65.3	-	42.2
Wang *et al.*	-	74.7	-	52.7
Eigen *et al.*	83.2	82.0	75.4	66.9
Ours (*superpixel+*)	82.7	81.3	74.8	67.0
Ours (*full model*)	**83.6**	**82.5**	**75.8**	**68.4**