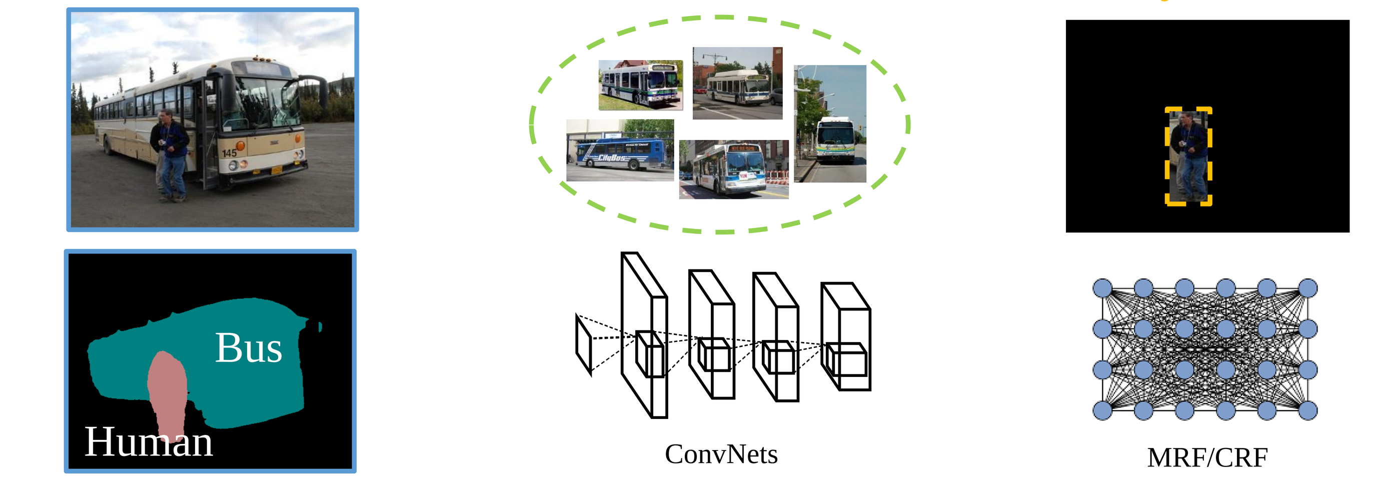
Semantic Image Segmentation via Deep Parsing Network

Ziwei Liu†, Xiaoxiao Li†, Ping Luo, Chen Change Loy, Xiaoou Tang
Department of Information Engineering, The Chinese University of Hong Kong
{lz013,lx015,pluo,ccloy,xtang}@ie.cuhk.edu.hk

1. Introduction

- **Task & General Approaches:**

Semantic Segmentation | Intra-object Variations | Inter-object Contexts



- **Motivation:**
  Combine ConvNets and MRF into a unified framework :
  ✓ End-to-end Training
  ✓ Rich Pairwise Relationship

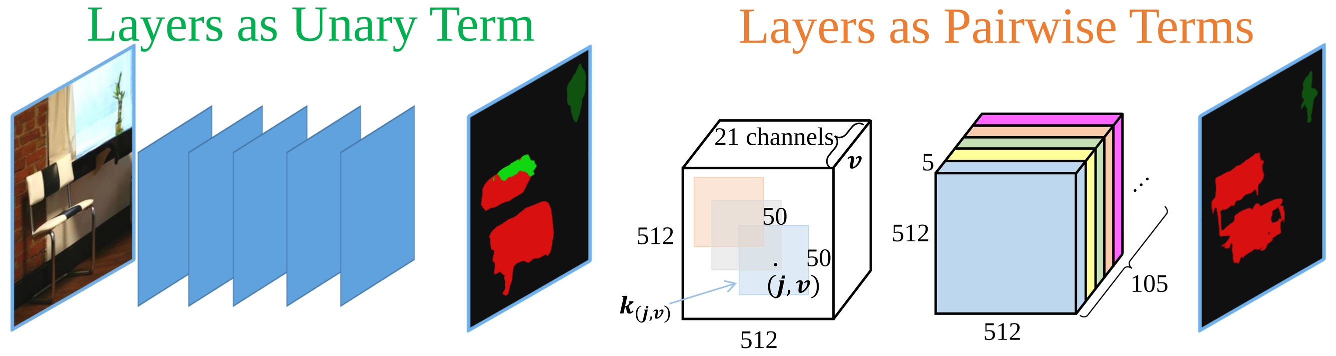
- **Existing Works:**

	Learned Features	Joint Training	# iterations
DenseCRF [NIPS 2011]	✗	-	10
FCN [CVPR 2015]	✓	-	-
DeepLab [ICLR 2015]	✓	✗	10
CRFasRNN [ICCV 2015]	✓	✓	10
DPN	✓	✓	1

- **Our Idea:**
  High-order MRF as One-pass CNN:

$$E(\mathbf{y}) = \sum_{\forall i \in \mathcal{V}} \Phi(\mathrm{y}_i^u) + \sum_{\forall i,j \in \mathcal{E}} \Psi(\mathrm{y}_i^u, \mathrm{y}_j^v)$$

Multiple Convolutional Layers as Unary Term | Additional Designed Layers as Pairwise Terms

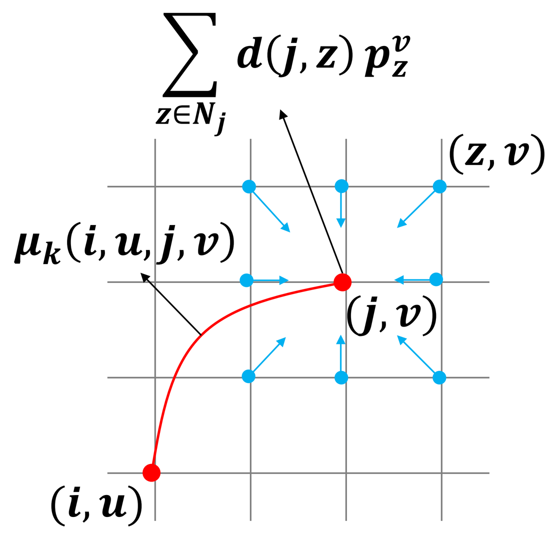


2. Approach

- **Unary Term**

$$\Phi(\mathrm{y}_i^u) = -\ln p_i^u$$

(p_i^u indicates the probability of the presence of label u at pixel i)



- **Pairwise Term**

$$\Psi(\mathrm{y}_i^u, \mathrm{y}_j^v) = \sum_{k=1}^{K} \lambda_k \mu_k(i,u,j,v) \sum_{\forall z \in \mathcal{N}_j} \mathrm{d}(j,z) p_z^v$$

- **Mean Field Solver**

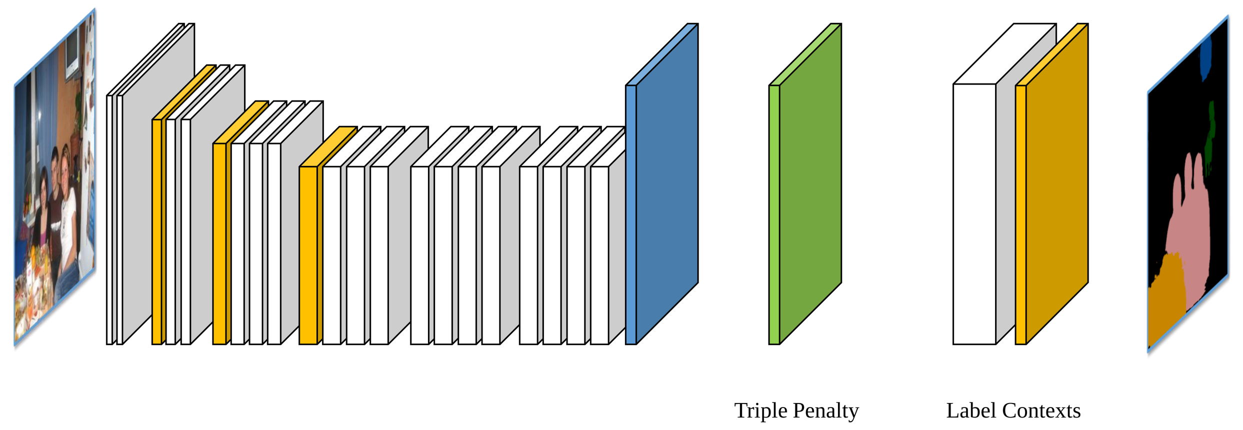
$$q_i^u \propto exp\left\{ \underbrace{-\Phi_i^u}_{\text{Unary Term}} - \underbrace{\sum_{k=1}^{K} \lambda_k \sum_{\forall v \in L, \forall j \in \mathcal{N}_i} \mu_k(i,u,j,v)}_{\text{Mixture of Label Contexts}} \underbrace{\sum_{\forall z \in \mathcal{N}_j} \mathrm{d}(j,z) q_z^v q_j^v}_{\text{Triple Penalty}} \right\}$$

(each q_i^u is initialized by the corresponding p_i^u)

3. Network Architecture

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
layer	2×conv	max	2×conv	max	3×conv	max	3×conv	3×conv	conv	conv	conv	lconv	conv	bmin	sum
filter–stride	3-1	2-2	3-1	2-2	3-1	2-2	3-1	5-1	25-1	1-1	1-1	50-1	9-1	1-1	1-1
#channel	64	64	128	128	256	256	512	512	4096	4096	21	21	105	21	21
activation	relu	idn	relu	idn	relu	idn	relu	relu	relu	relu	sigm	lin	lin	idn	soft
size	512	256	256	128	128	64	64	64	64	64	512	512	512	512	512

Deep Parsing Network (DPN) : 512×512×3 input image; 512×512×21 output label maps

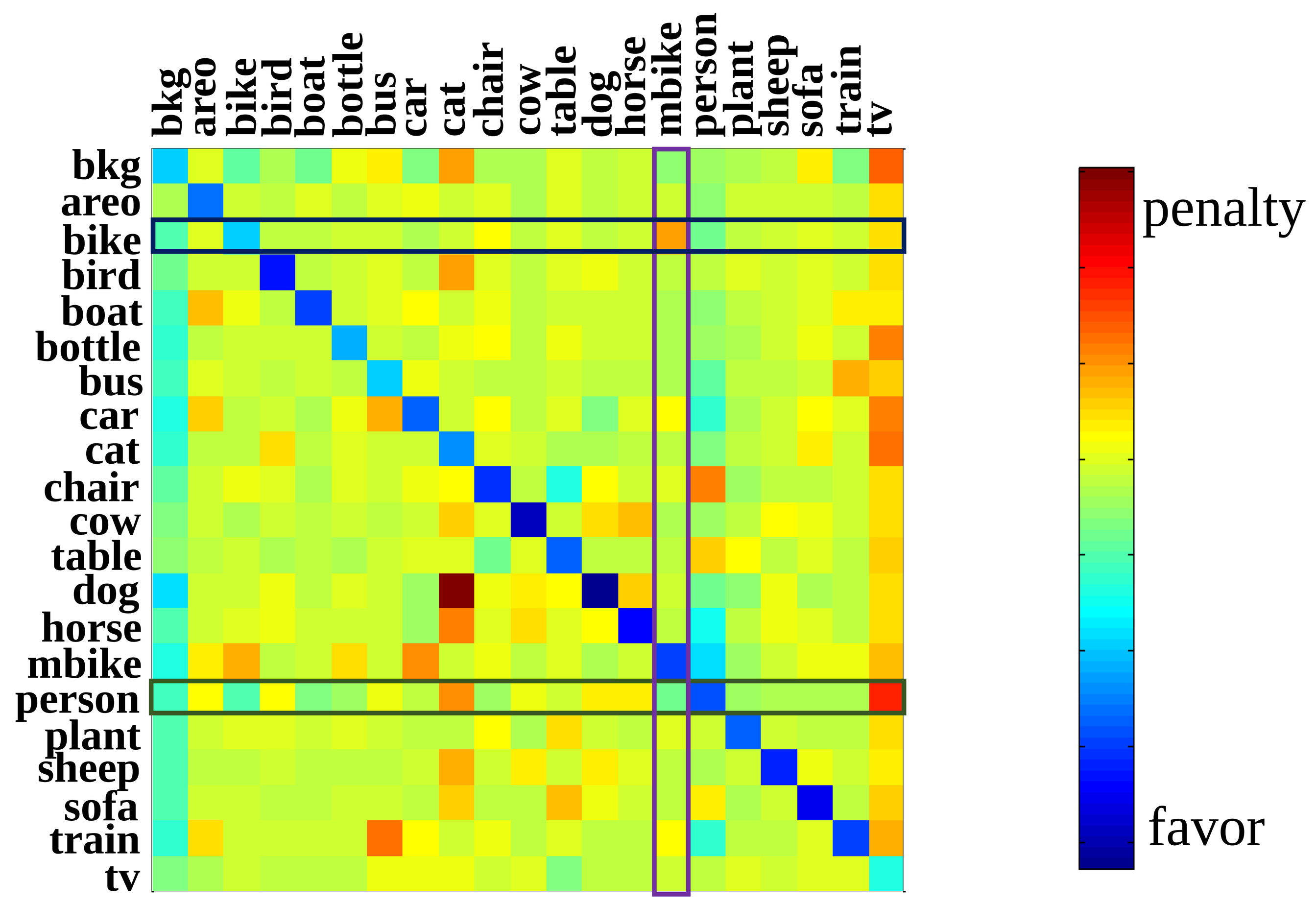


Convolution | Max Pooling | Deconvolution | Local Convolution

- **Project Page:** http://personal.ie.cuhk.edu.hk/~lz013/projects/DPN.html

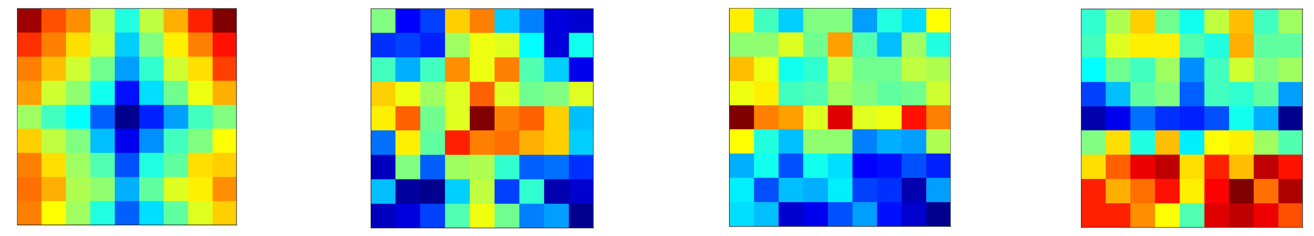
4. Effectiveness of DPN

- **Label-Label Space**



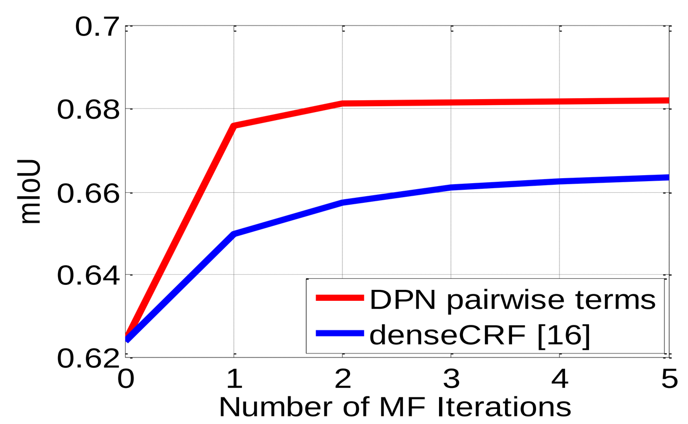
when 'motor bike' is presented, 'person' is more likely to present than 'bike'.

- **Spatial-Label Space**

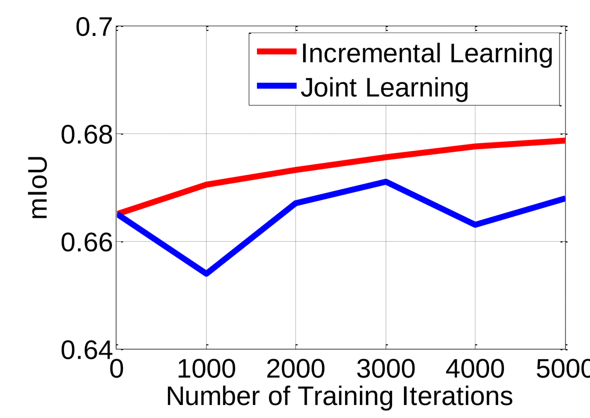


bottle : bottle | train : bkg | person : mbike | chair : person

- **Pairwise Terms Comparisons**



- **End-to-end Learning**

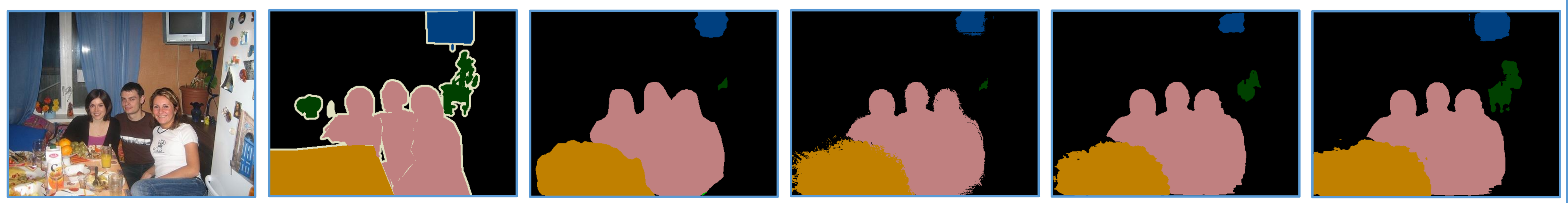


5. Overall Performance

	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
FCN	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
DeepLab†	89.2	46.7	88.5	63.5	68.4	87	81.2	86.3	32.6	80.7	62.4	81	81.3	84.3	82.1	56.2	84.6	58.3	76.2	67.2	73.9
RNN†	90.4	55.3	88.7	68.4	69.8	88.3	82.4	85.1	32.6	78.5	64.4	79.6	81.9	**86.4**	81.8	58.6	82.4	53.5	77.4	70.1	74.7
BoxSup†	**89.8**	38	**89.2**	**68.9**	68	89.6	83	**87.7**	34.4	83.6	**67.1**	81.5	83.7	85.2	83.5	58.6	84.9	55.8	**81.2**	**70.7**	75.2
DPN	87.7	59.4	78.4	64.9	70.3	89.3	83.5	86.1	31.7	79.9	62.6	81.9	80	83.5	82.3	60.5	83.2	53.4	77.9	65	74.1
DPN†	89	**61.6**	87.7	66.8	**74.7**	**91.2**	**84.3**	87.6	**36.5**	**86.3**	66.1	**84.4**	**87.8**	85.6	**85.4**	**63.6**	**87.3**	**61.3**	79.4	66.4	**77.5**

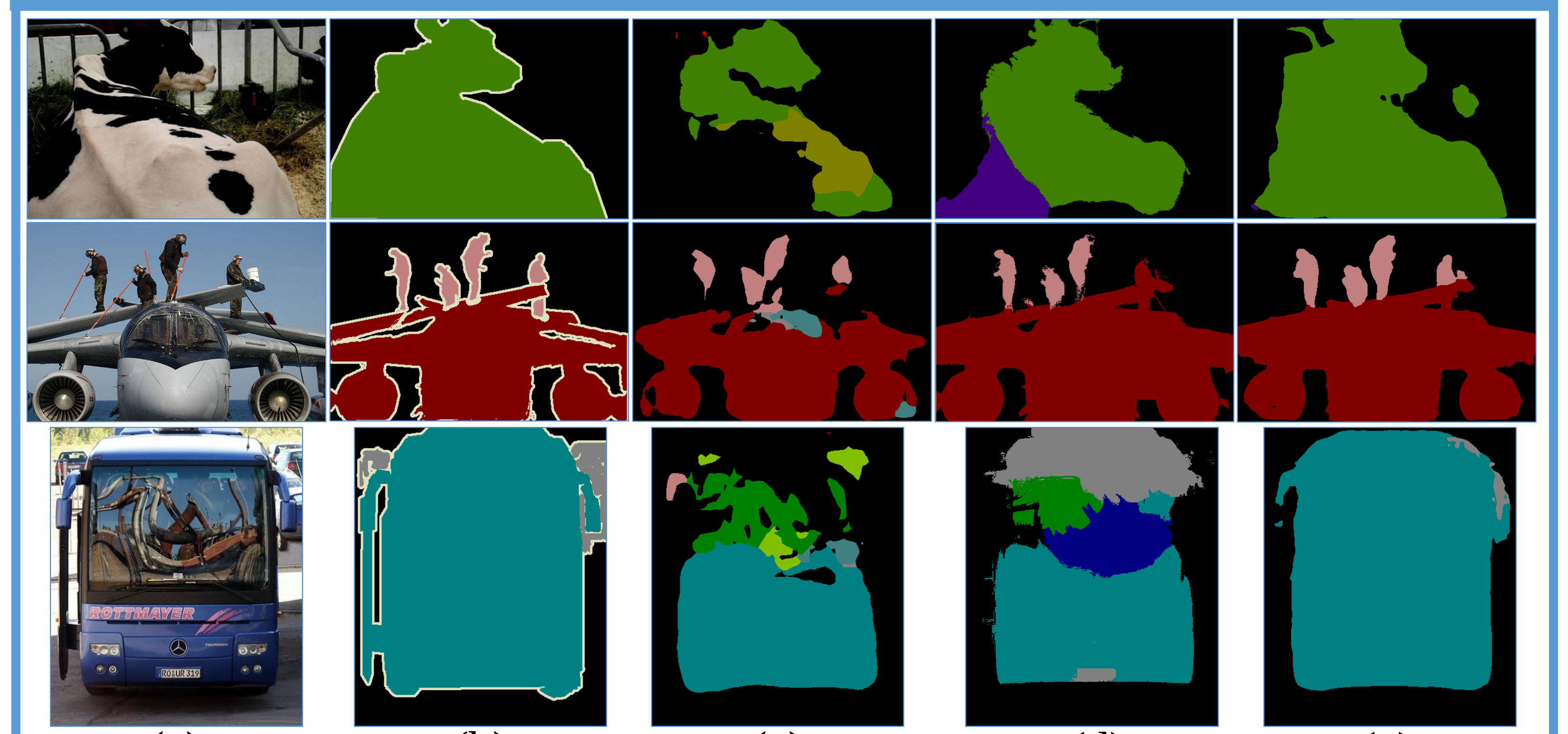
Per-class results on VOC12 test. The approaches pre-trained on COCO are marked with †.

6.1 Per-stage Visualization



Original Image | Ground Truth | Unary Term | Triple Penalty | Label Contexts | Joint Tuning

6.2 Visual Quality Comparisons



(a) (b) (c) (d) (e)

Visual quality comparison of different semantic image segmentation methods: (a) input image (b) ground truth (c) FCN (d) DeepLab and (e) DPN

7. Conclusion

DPN employs one-pass CNN to model high-order MRF
High performance by approximating one iteration of MF

DPN incorporates various types of pairwise terms
Rich contextual information

DPN contains only conventional operations of CNN
Easier to be parallelized and speeded up in GPU