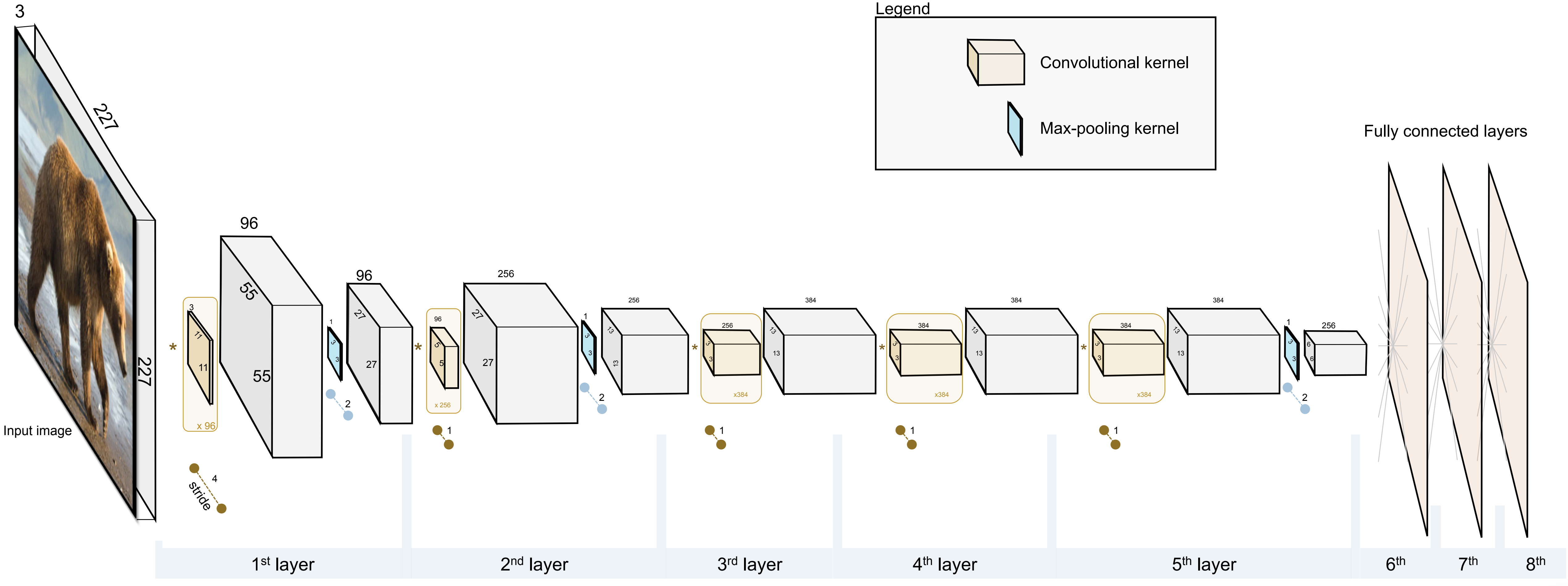
Semantic Image Segmentation via Deep Learning

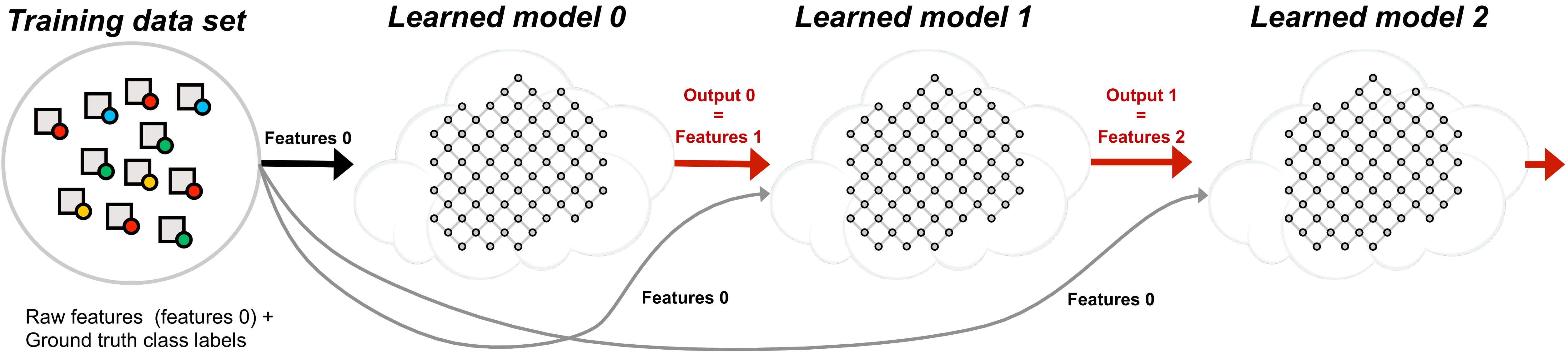
What is deep learning ?

Deep Convolutional Neural Networks



Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). **ImageNet Classification with Deep Convolutional Neural Networks.** NIPS 2012

“Stacked” classifiers



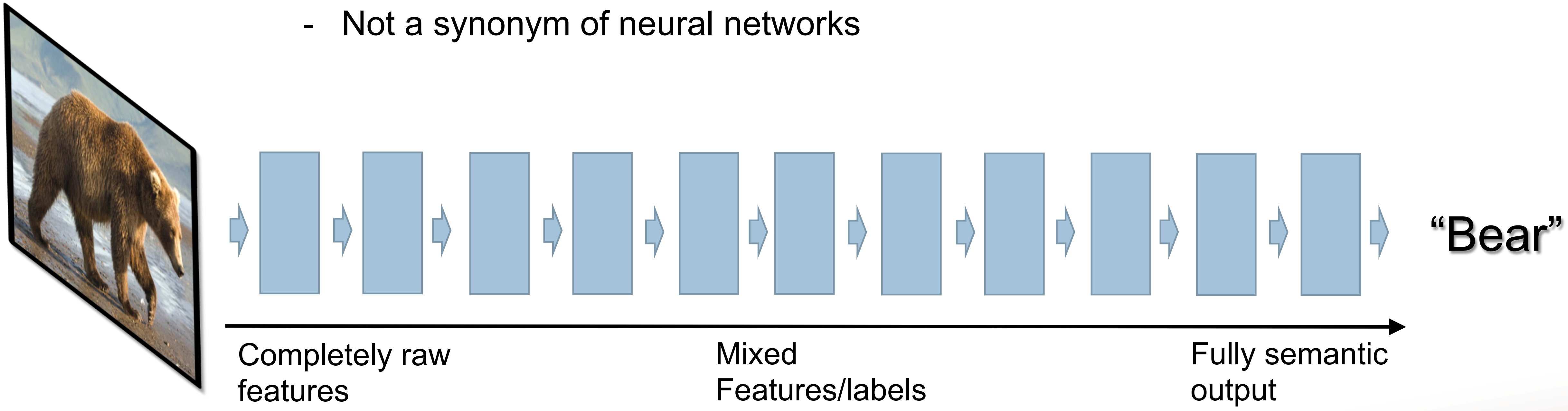
Wolpert, D., ***Stacked Generalization.***, Neural Networks, 5(2), pp. 241-259., 1992

Deep learning (my definition)

Properties

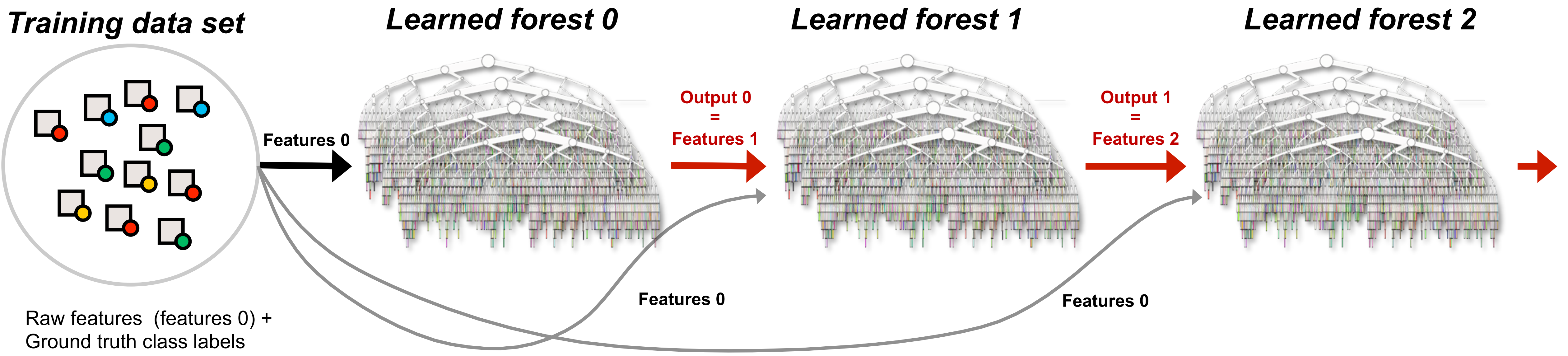
- Functions of functions of the input data (e.g. conv of conv)
- Representation learning. Data transformation in learned stages
- Non linearities (merging, pooling etc.) in between layers

- Not a synonym of neural networks



“Stacked” classifiers a.k.a. “AutoContext”

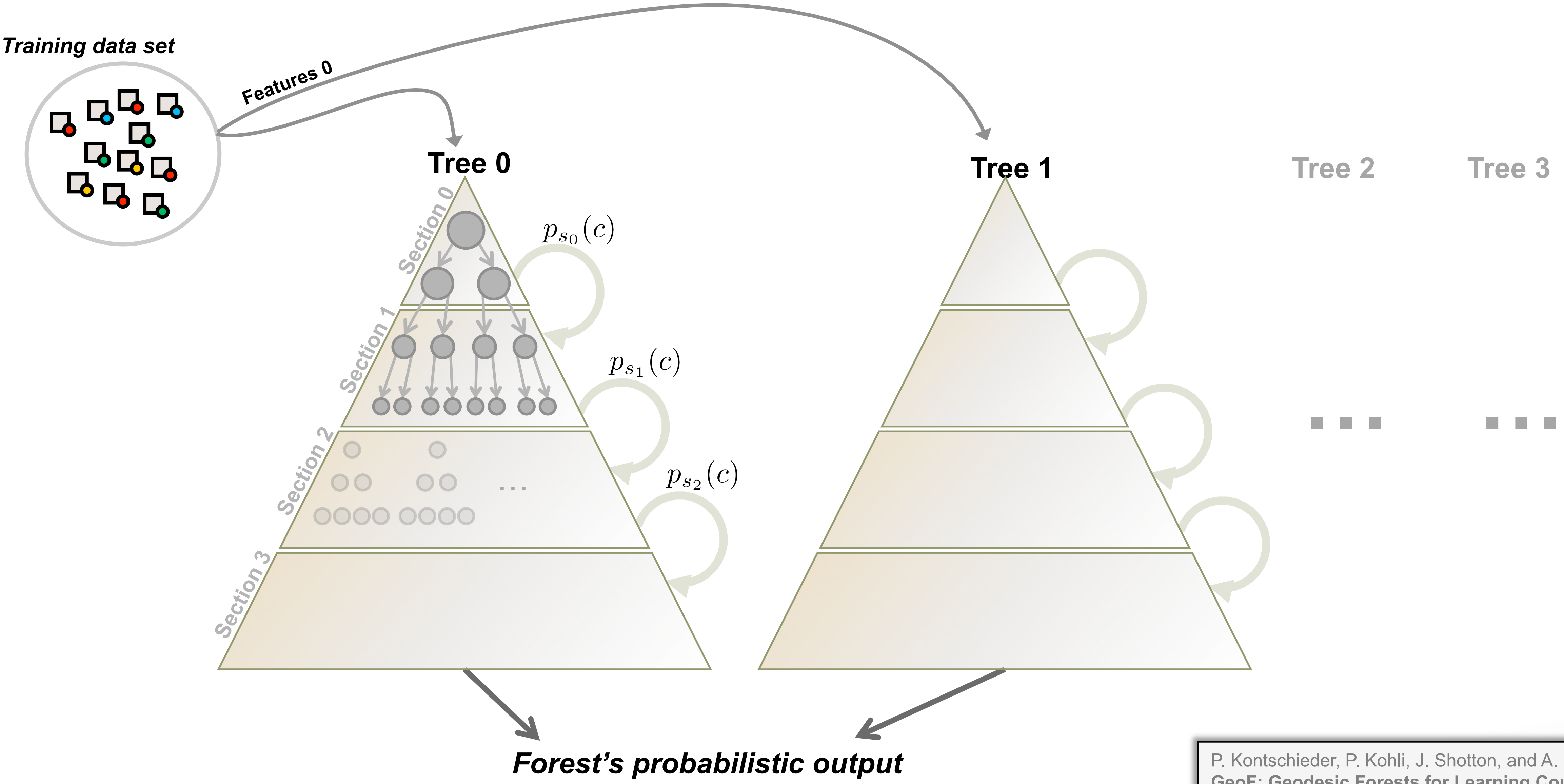
In the medical image analysis literature



- Demonstrated to exploit a learned model of **spatial context**
- Applied successfully to semantic segmentation
- Applied successfully to medical images

Zhuowen Tu and Xiang Bai, **Auto-context and Its Application to High-level Vision Tasks and 3D Brain Image Segmentation**, IEEE Trans. on PAMI

Another form of deep learning: **Entangled decision forests**



P. Kontschieder, P. Kohli, J. Shotton, and A. Criminisi, **GeoF: Geodesic Forests for Learning Coupled Predictors**, in *Proc. CVPR*, IEEE, June 2013

Deep Forests for Semantic Segmentation

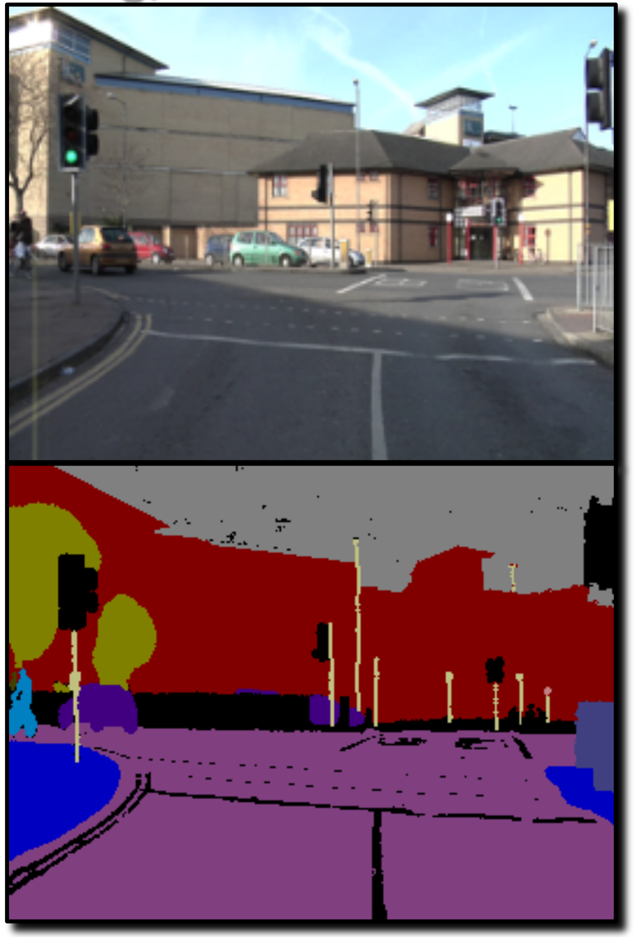
P. Kontschieder, P. Kohli, J. Shotton, and A. Criminisi, **GeoF: Geodesic Forests for Learning Coupled Predictors,** in *Proc. CVPR*, IEEE, June 2013

Semantic Segmentation

Spatial smoothness

e.g. spatially compact segments

Long, thin structures



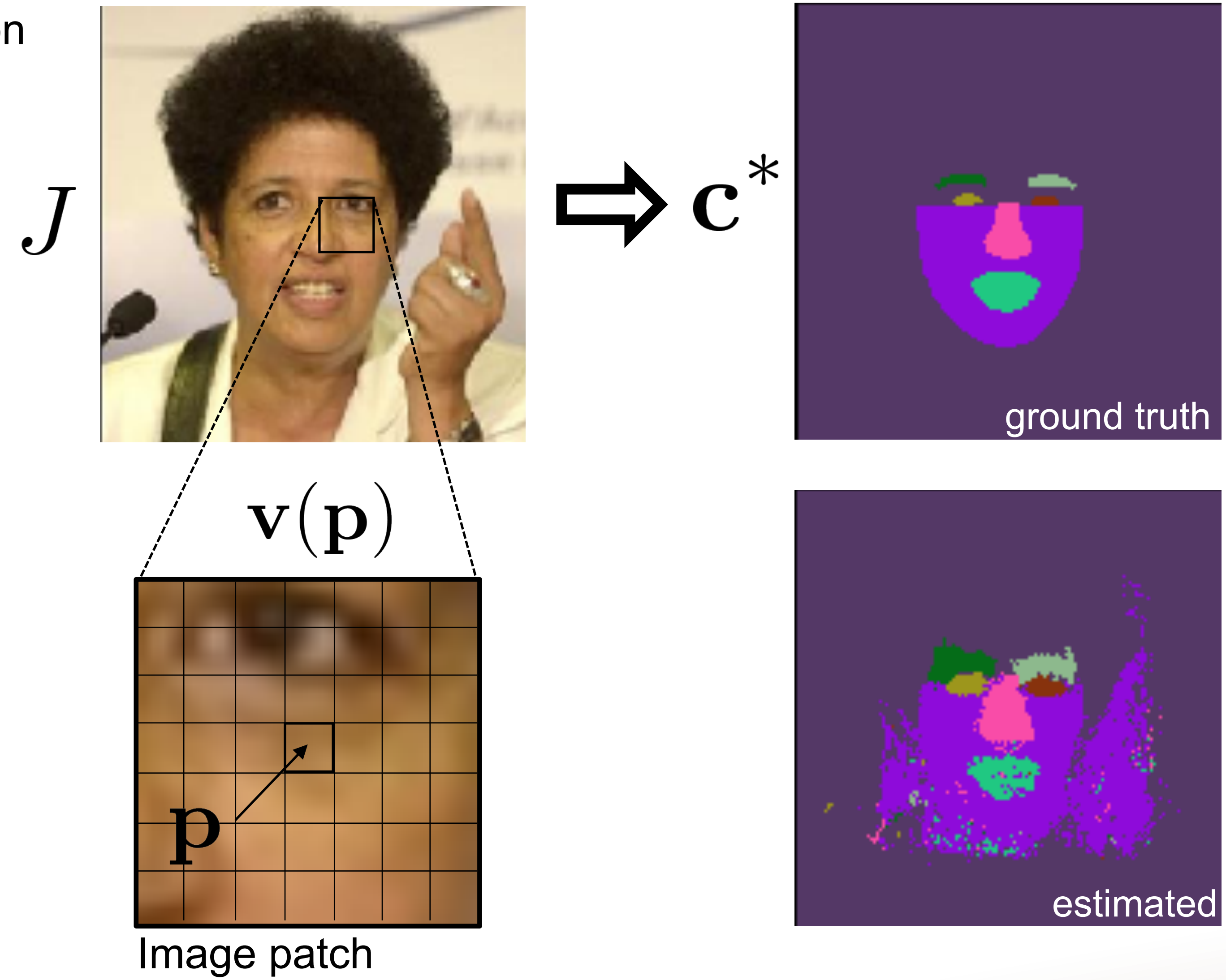
e.g. lamp posts, blood vessels

Semantic context

e.g. heart in between lungs, liver below heart.

Background: Pixel-wise labeling

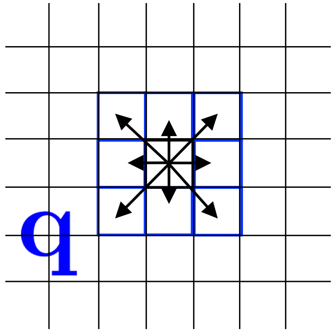
Semantic image segmentation
as pixel-wise classification



[Amit & Geman, NC'1997]
[Breiman, ML'2001]

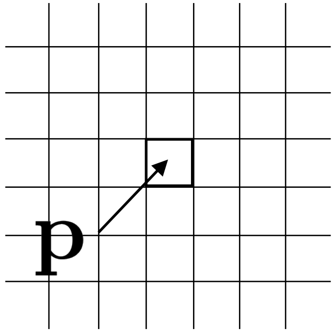
Background: Graphical Models for spatial prior

+

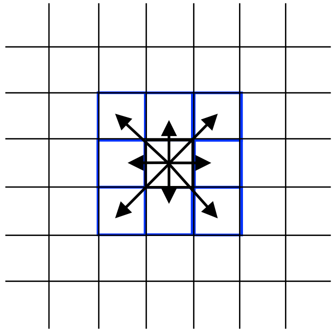


$\mathbf{q} \in \mathcal{N}_8(\mathbf{p})$

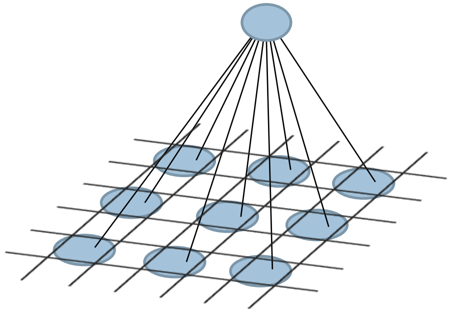
[Geman & Geman, PAMI'1984]
[Blake et al., ECCV'2004]



+

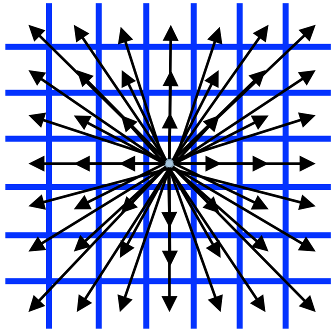


+



[Kohli et al., CVPR'2008]
[Ladicky et al., ICCV'2009]

+



[Krähenbühl & Koltun, NIPS 2012]

Unary term + Pairwise term + Higher order term

Background: Classification Forest Labelling

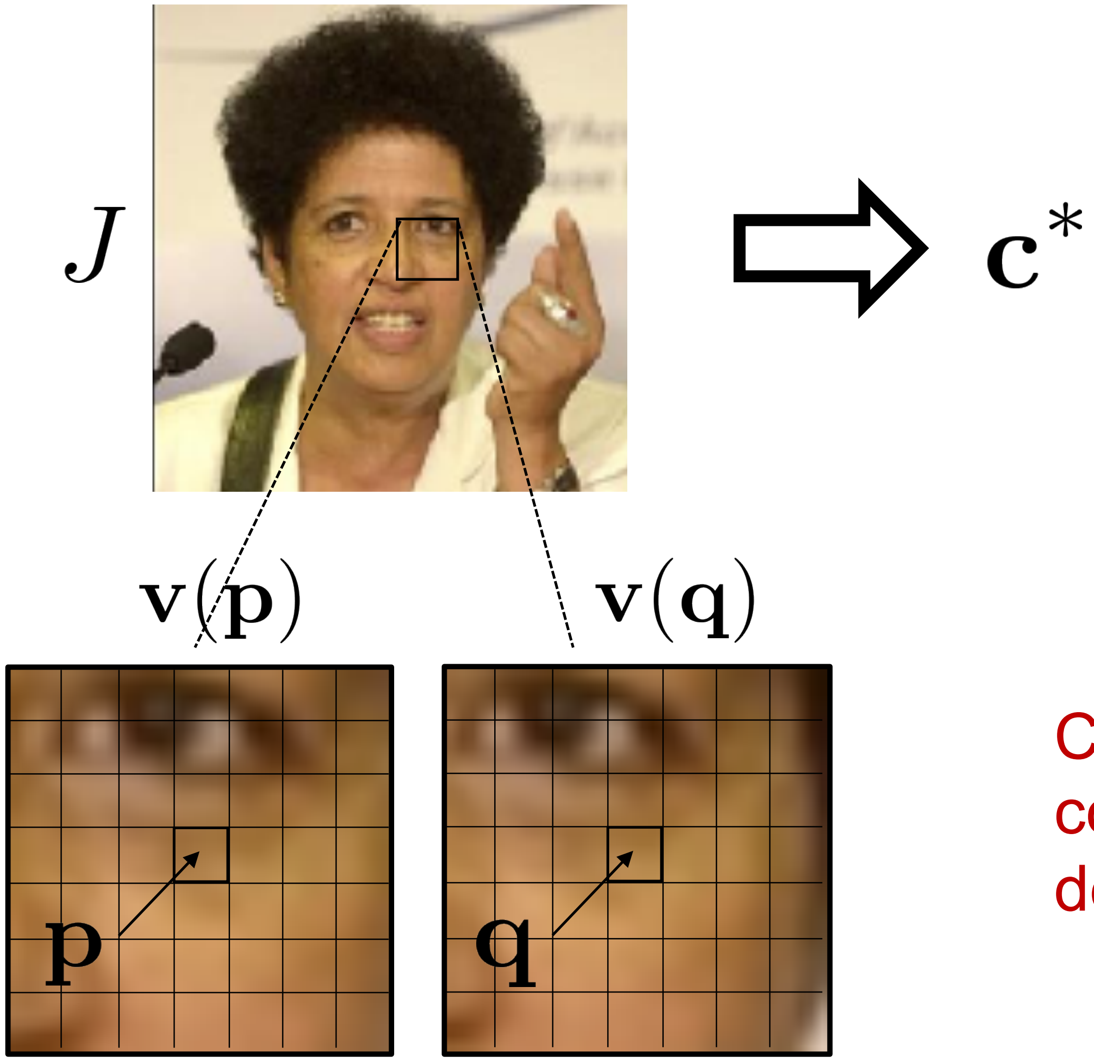


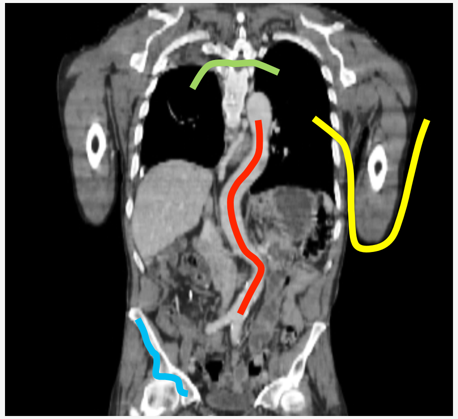
Image patches for 2 adjacent pixels

Can ly
cons rom a
decis ier alone?

Entangled Geodesic Forests

Efficient, soft connectivity features

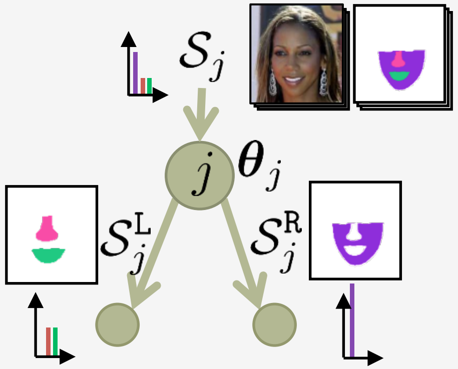
(modification at feature level)



Better features capturing spatial smoothness

Field-Inspired Training Objective

(modification of training energy)



Better surrogate training function

Soft Connectivity Features for Capturing Spatial Smoothness

Semantic segmentation – in Kinect

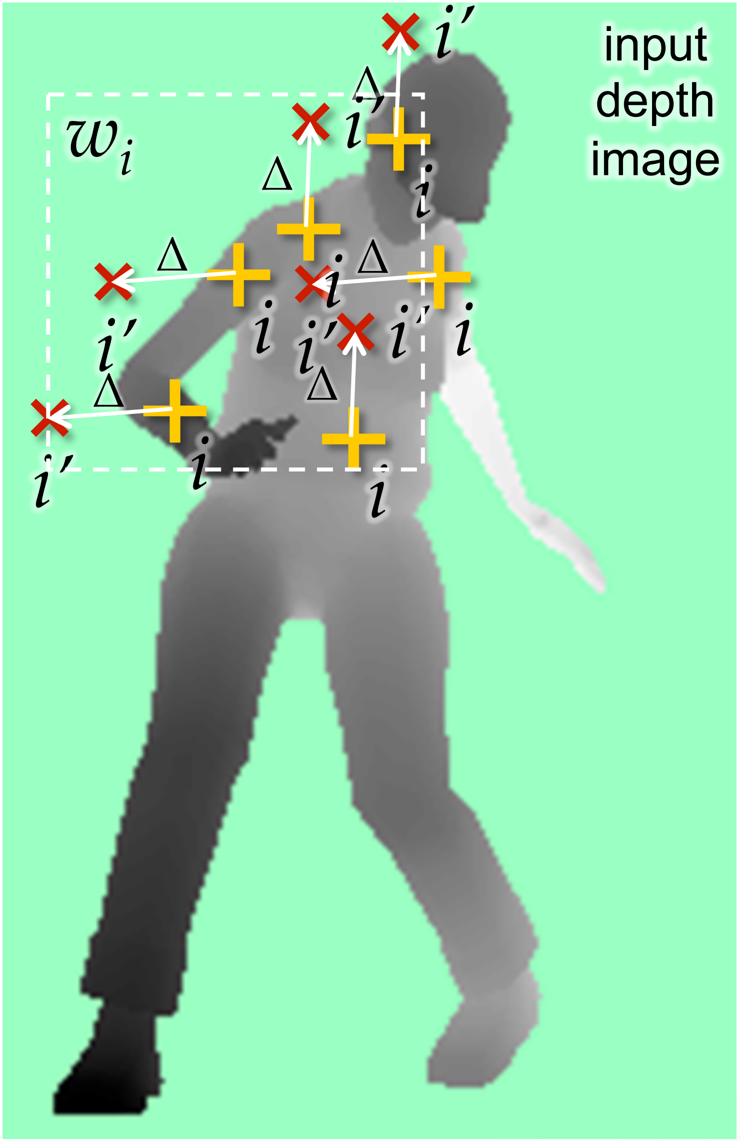
input depth image
from Kinect depth camera

inferred body parts
from our algorithm running on the XBox

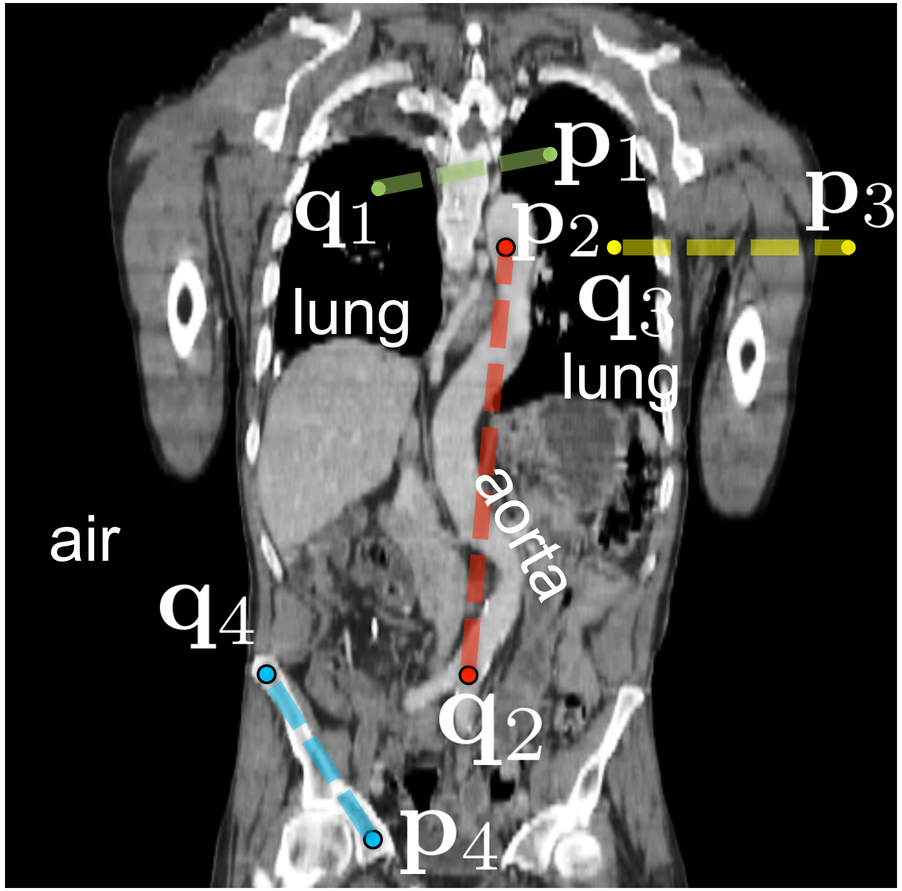
Pixel-wise comparison features – in Kinect

- Depth comparisons:
  - $f(i\,;\Delta) = d(i) - d(i')$
    where $i' = i + \Delta$
- Background pixels
  - d = large constant

desired
body parts

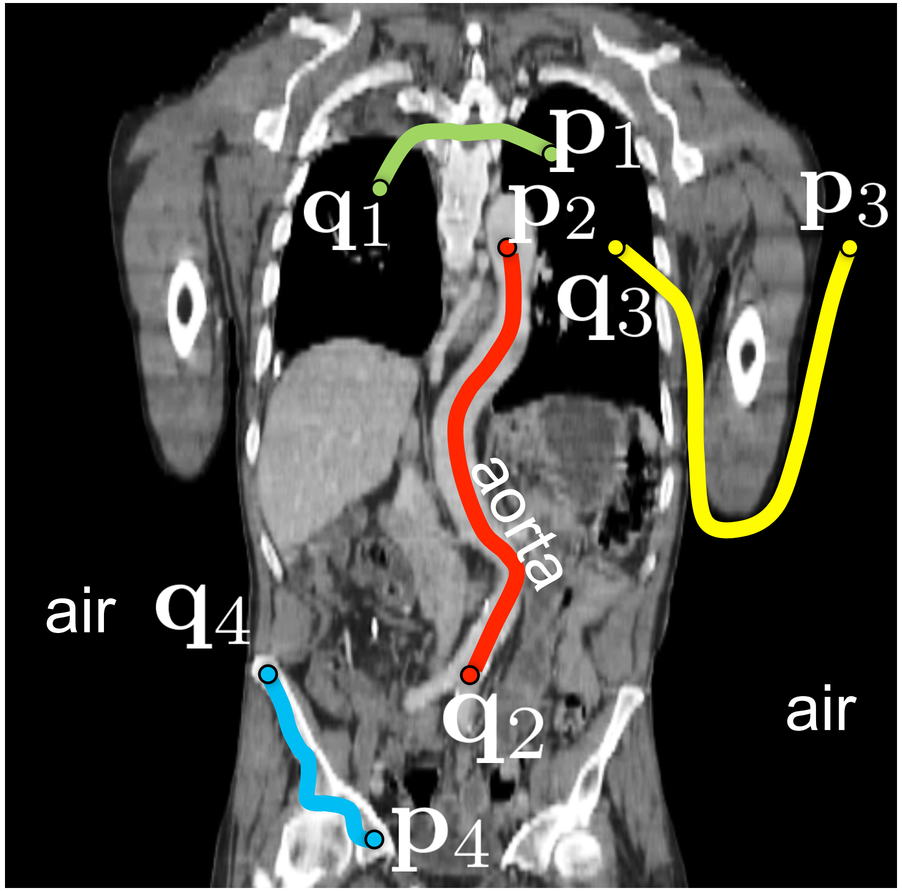


Soft connectivity features



Features:

comparing pairs of pixels
(as used in Kinect)



Features:

exploiting the intensity profile along
shortest path/s connecting the two pixels

Soft connectivity features

Input image

Soft input mask (e.g. likelihood ratio)

Generalized geodesic distance

Image
$$J(\mathbf{p}) : \Psi \subset \mathbb{N}^2 \to \mathbb{R}$$

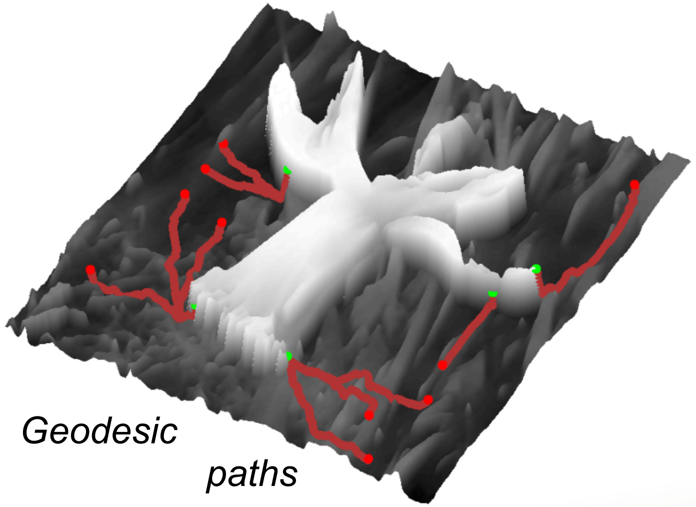
Real valued mask
$$M(\mathbf{p}) : \Psi \subset \mathbb{N}^2 \to [0,1]$$

Generalized geodesic distance

$$Q(\mathbf{p}; M, \nabla J) = \min_{\mathbf{p}' \in \Psi} \left(\delta(\mathbf{p}, \mathbf{p}') + \nu M(\mathbf{p}')\right)$$

$$\delta(\mathbf{p}, \mathbf{q}) = \inf_{\mathbf{\Gamma} \in \mathcal{P}_{\mathbf{p},\mathbf{q}}} \int_0^{l(\mathbf{\Gamma})} \sqrt{1 + \gamma^2 (\nabla J(s) \cdot \mathbf{\Gamma}'(s))^2} ds.$$

$$\mathbf{\Gamma}'(s) = \frac{\partial \mathbf{\Gamma}}{\partial s}$$



Geodesic paths

A. Criminisi, T. Sharp, and P. Perez.
Geodesic image and video editing. SIGGRAPH 2011

Soft connectivity features

ground truth segmentation

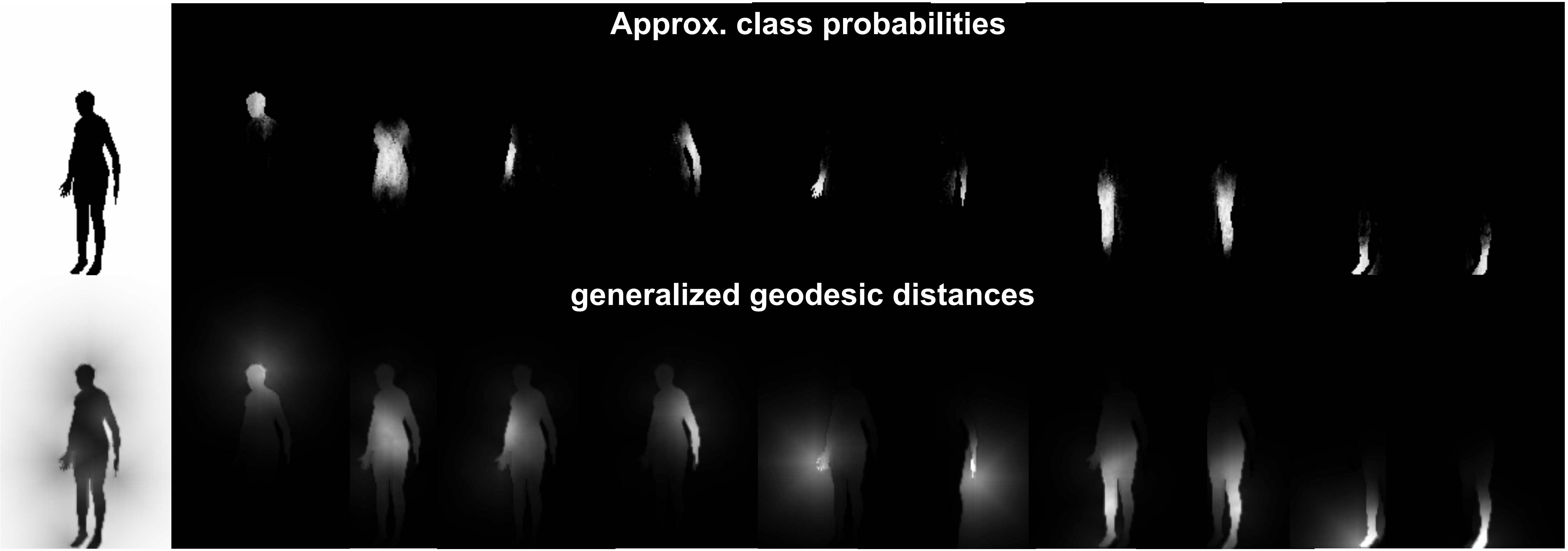
approximate class probabilities

class: torso

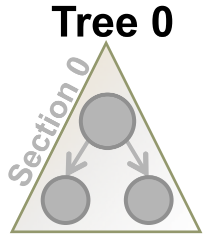
class: left leg

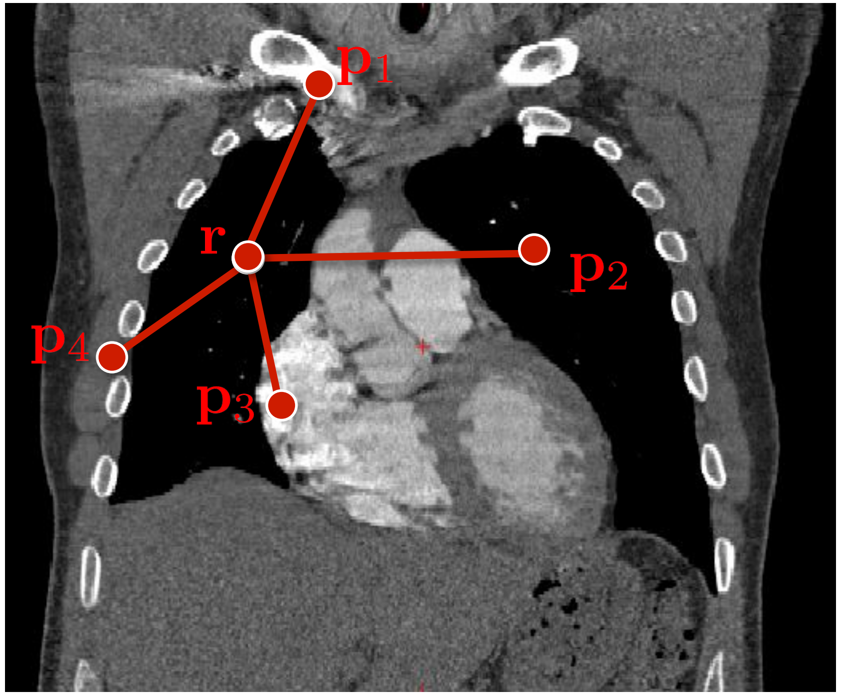
generalized geodesic distances

Soft connectivity features



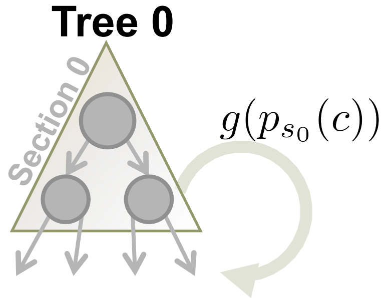
Entangled Geodesic Forests

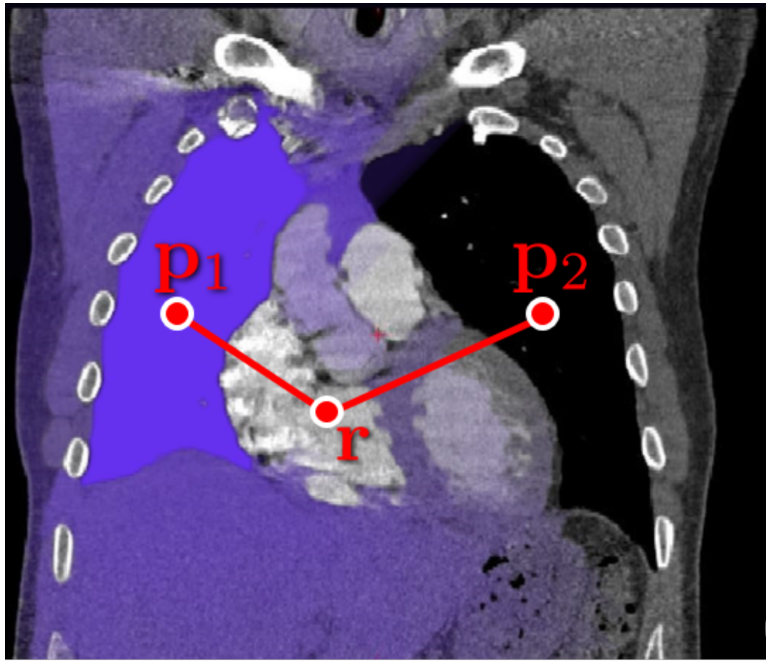




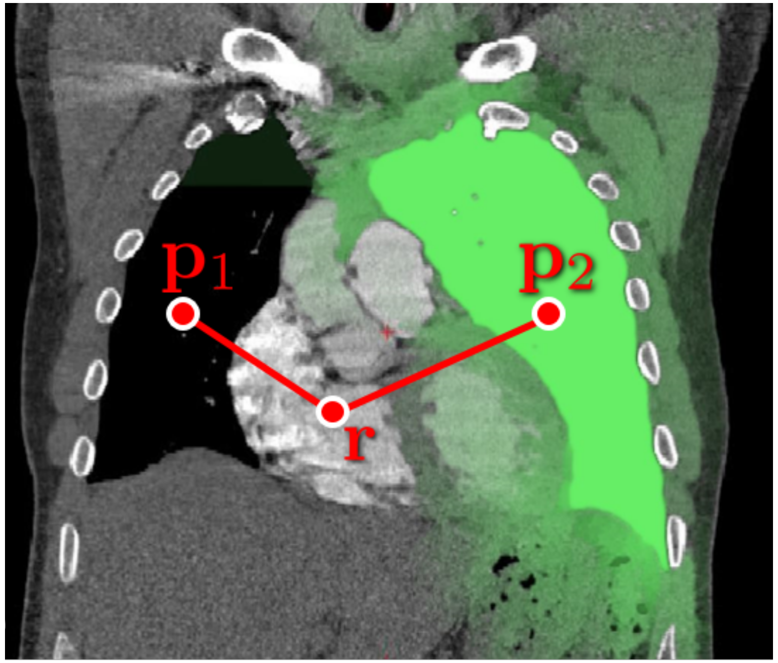
Conventional pixel-comparison features

Entangled geodesic trees





$g(p_{s_0}(c = \texttt{rightlung}))$

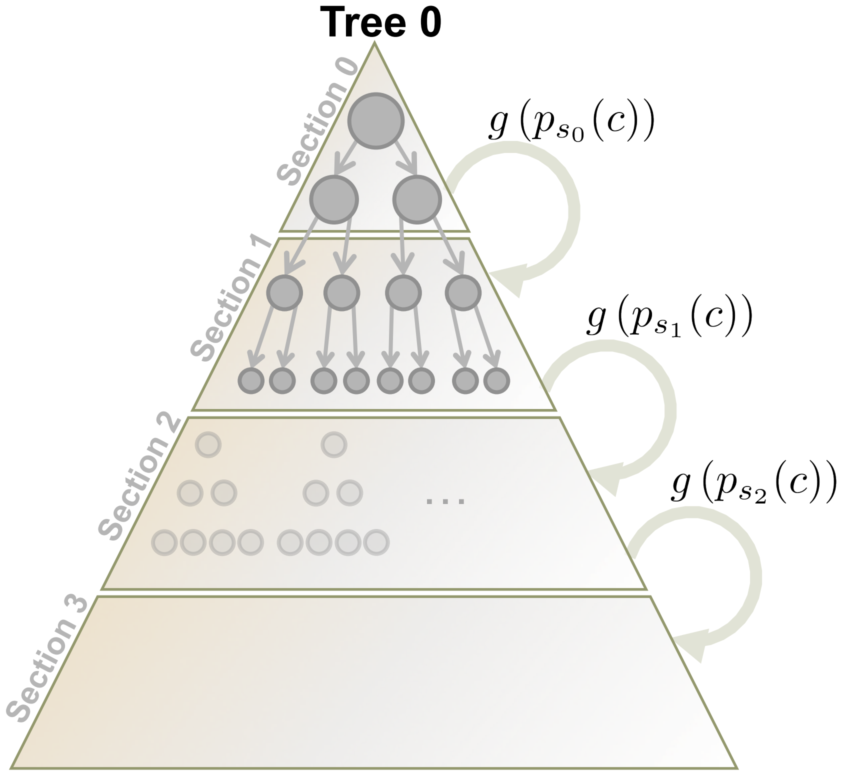


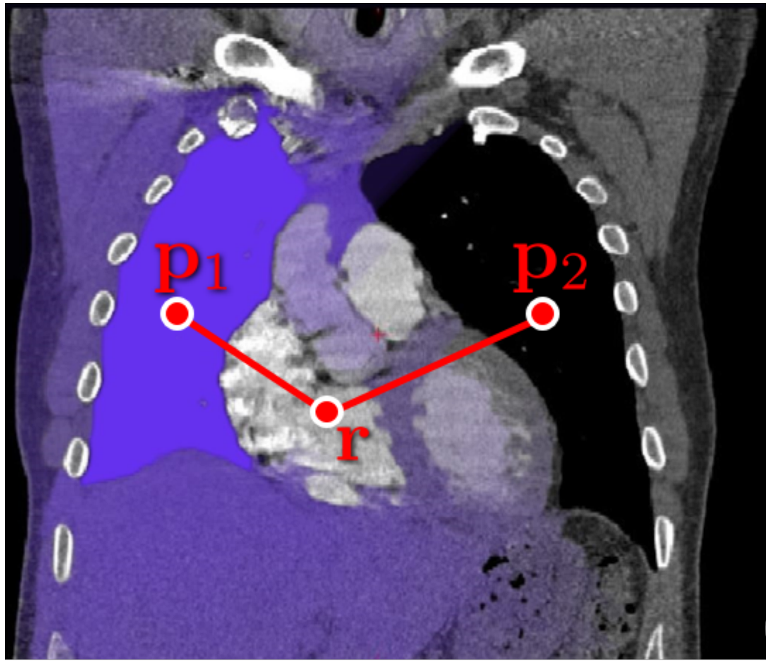
$g(p_{s_0}(c = \texttt{leftlung}))$

Pixel-comparison features on <u>geodesic-transformed probabilities</u>

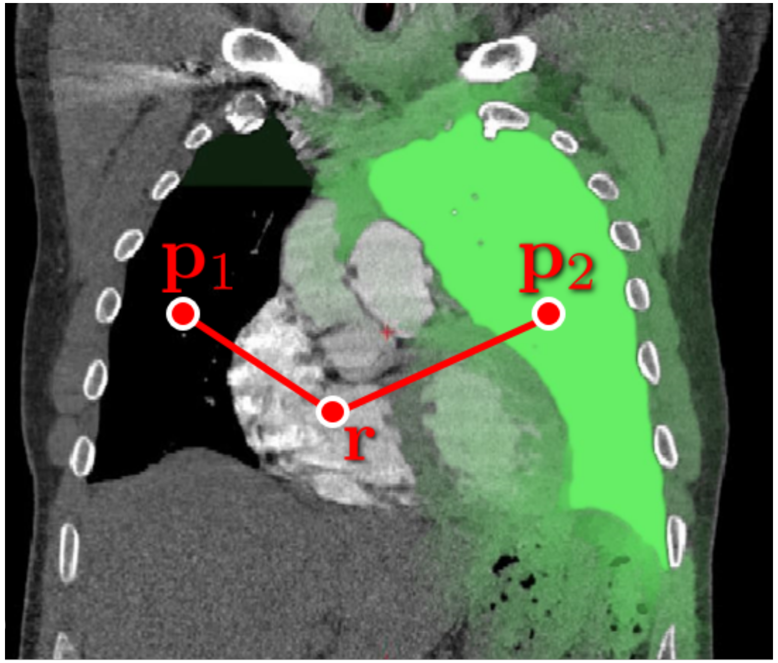
Capturing semantic context

Entangled geodesic trees





$g(p_{s_0}(c = \texttt{rightlung}))$

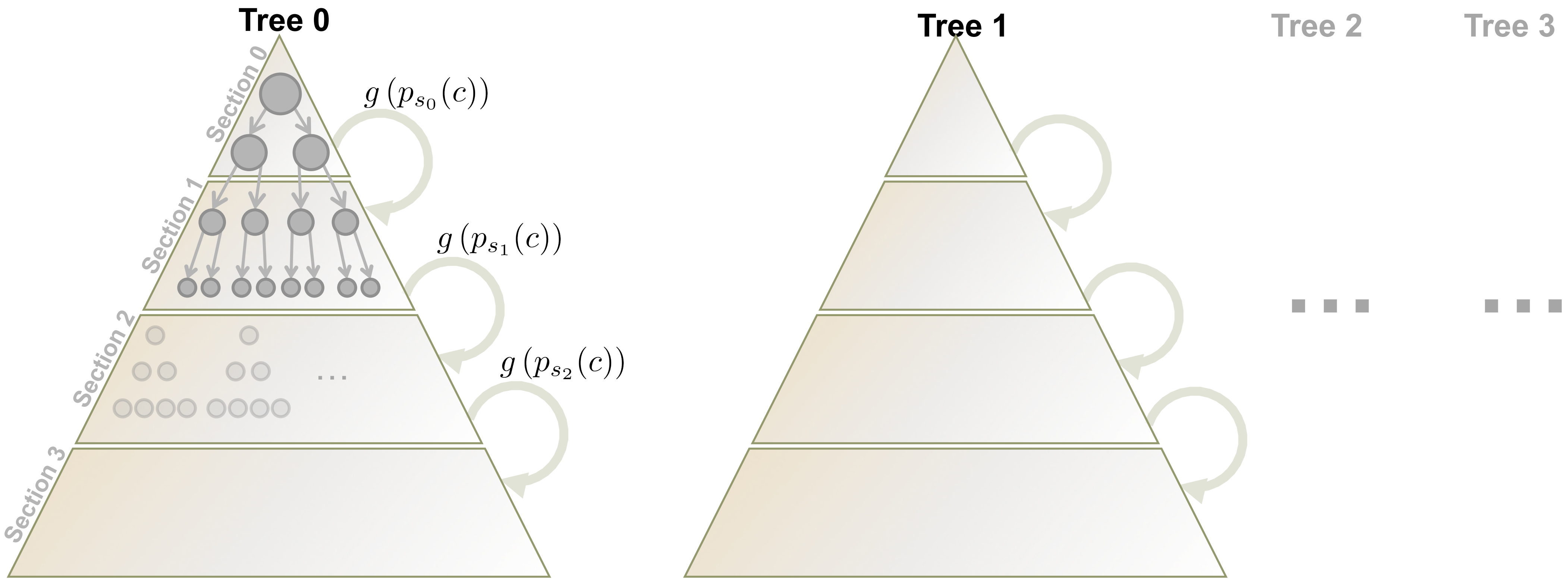


$g(p_{s_0}(c = \texttt{leftlung}))$

Pixel-comparison features on <u>geodesic-transformed probabilities</u>

Capturing semantic context

Entangled geodesic forests



Field-inspired Training Objective

Field-Inspired Training Objective

We wish the forest to learn to apply the “right” level of spatial smoothness.


Input


Ground truth


Std. Class. Forest

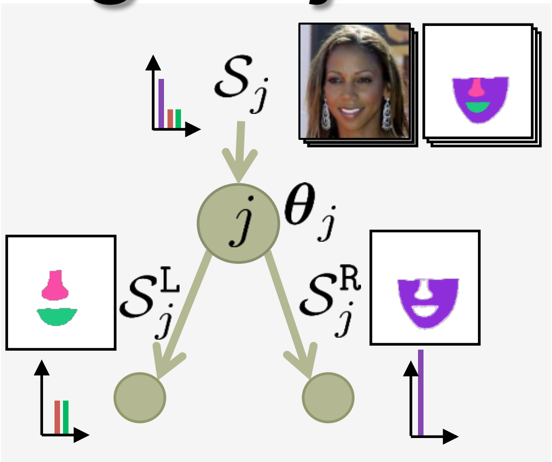

Std. Entanglement

Proposed

entanglement + generalized geodesic distances

+ field-inspired training objective

The Training Objective Function



Node training

$$\boldsymbol{\theta}_j = \arg\min_{\boldsymbol{\theta}\in\mathcal{T}_j} E_{\mathtt{IT}}(\mathcal{S}_j, \boldsymbol{\theta})$$

Node training

$$\boldsymbol{\theta}_j = \arg\min_{\boldsymbol{\theta}\in\mathcal{T}_j} E_{\mathtt{RF}}(\mathcal{S}_j, \boldsymbol{\theta})$$

IG-based energy

$$E_{\mathtt{IT}}(\mathcal{S}_j, \boldsymbol{\theta}) = -\sum_{i\in\{\mathtt{L},\mathtt{R}\}} |\mathcal{S}_j^i| \sum_{c\in\mathcal{C}} p(c|\mathcal{S}_j^i) \log p(c|\mathcal{S}_j^i)$$

Random field-based energy

$$E_{\mathtt{RF}}(\mathcal{S}_j, \boldsymbol{\theta}) \quad = \quad \sum_{i\in\{\mathtt{L},\mathtt{R}\}} \Bigg(- \sum_{k \mid \mathbf{z}_k\in\mathcal{S}_j^i} \log p(c = c(\mathbf{z}_k)|\mathcal{S}_j^i) + \lambda \sum_{k \mid \mathbf{z}_k\in\mathcal{S}_j^i, \mathbf{r}\in\mathcal{N}(\mathbf{z}_k)} [c(\mathbf{z}_k) \neq c(\mathbf{r})] \Bigg)$$

- When are resulting segmentations smoother? When are they more accurate?
- Are the geodesic features used? When are they selected more often?
- Have we been able to remove the need for an MRF/CRF post-processing step?

A closer look at the unary term

IG-based energy (unaries only)

$$-|\mathcal{S}| \sum_{c\in\mathcal{C}} p(c|\mathcal{S}) \log p(c|\mathcal{S}) =$$

$$-|\mathcal{S}| \sum_{c\in\mathcal{C}} \frac{n_c}{|\mathcal{S}|} \log \frac{n_c}{|\mathcal{S}|} =$$

$$E_{\mathtt{IT}} = -\sum_{c\in\mathcal{C}} n_c \log \frac{n_c}{|\mathcal{S}|}$$

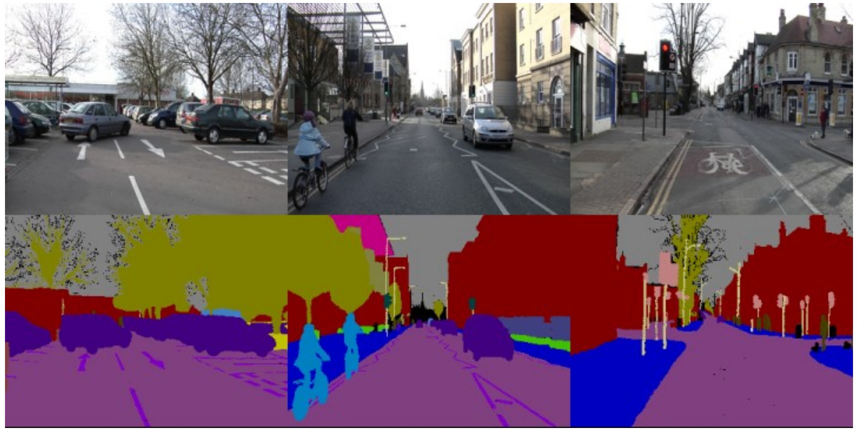
Random field-based energy (unaries only)

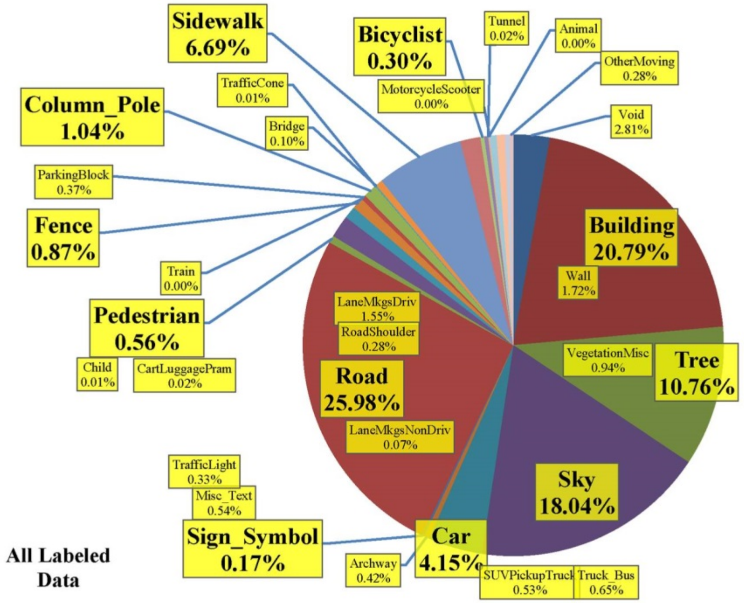
$$-\sum_{k \,|\, \mathbf{z}_k \in \mathcal{S}} \log p(c = c(\mathbf{z}_k)|\mathcal{S}) =$$

$$-\left(n_0 \log \frac{n_0}{|\mathcal{S}|} + n_1 \log \frac{n_1}{|\mathcal{S}|} + \ldots\right) =$$

$$E_{\mathtt{RF}} = -\sum_{c\in\mathcal{C}} n_c \log \frac{n_c}{|\mathcal{S}|}$$

Dealing with Unbalanced Datasets





CamVid Dataset [Brostow et al.,

Global sample reweighing according to inverse frequency!

$\mathcal{S}_0$

$$\omega_c = \frac{|\mathcal{S}_0|}{n(c, \mathcal{S}_0)}$$

Root node training set statistics

$$Z(\mathcal{S}_j) = \sum_{k \in \mathcal{C}} w_k \, n(k, \mathcal{S}_j)$$

Node-based normalization factor

Dealing with Unbalanced Datasets

IG-based energy (unaries only)

$$-Z(\mathcal{S}) \sum_{c\in\mathcal{C}} p(c|\mathcal{S}, w_c) \log p(c|\mathcal{S}, w_c) =$$

$$-Z(\mathcal{S}) \sum_{c\in\mathcal{C}} \frac{w_c n_c}{Z(\mathcal{S})} \log \frac{w_c n_c}{Z(\mathcal{S})} =$$

$$E_{\mathtt{IT}} = -\sum_{c\in\mathcal{C}} w_c n_c \log \frac{w_c n_c}{Z(\mathcal{S})}$$

Random field-based energy (unaries only)

$$-Z(\mathcal{S}) \sum_{k\,|\,\mathbf{z}_k\in\mathcal{S}} \log p(c = c(\mathbf{z}_k)|\mathcal{S}, w_c) =$$

$$-Z(\mathcal{S}) \left(n_0 \log \frac{w_0 n_0}{Z(\mathcal{S})} + n_1 \log \frac{w_1 n_1}{Z(\mathcal{S})} + \dots \right) =$$

$$E_{\mathtt{RF}} = -Z(\mathcal{S}) \sum_{c\in\mathcal{C}} n_c \log \frac{w_c n_c}{Z(\mathcal{S})}$$

Proposed forest training energy

Experiments and Results

Experimental Evaluation

Twelve challenging and very diverse image datasets

Lab. Faces in the Wild

Daimler stereo

...

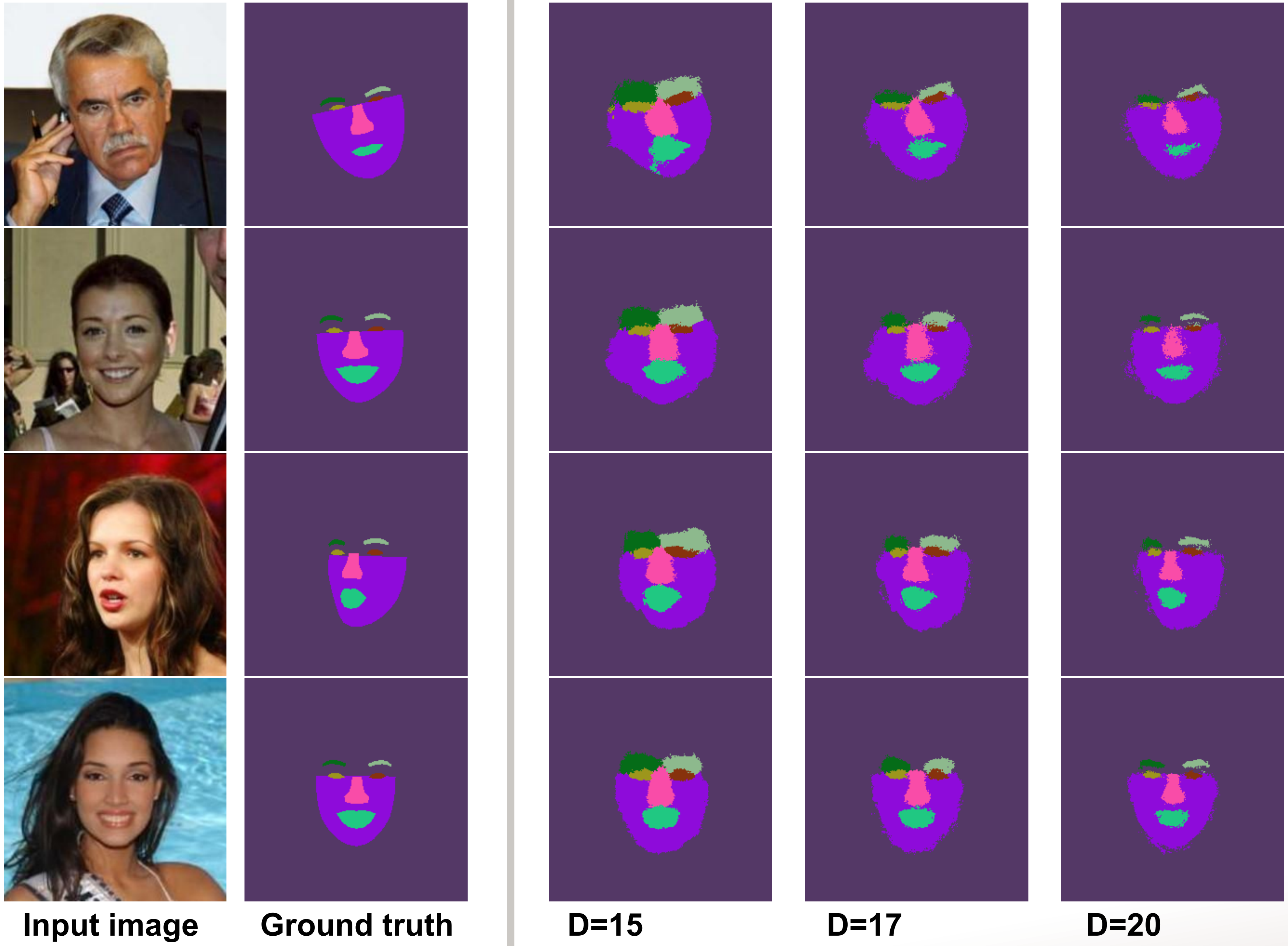
Computed Tomography

KinectBG

CamVid

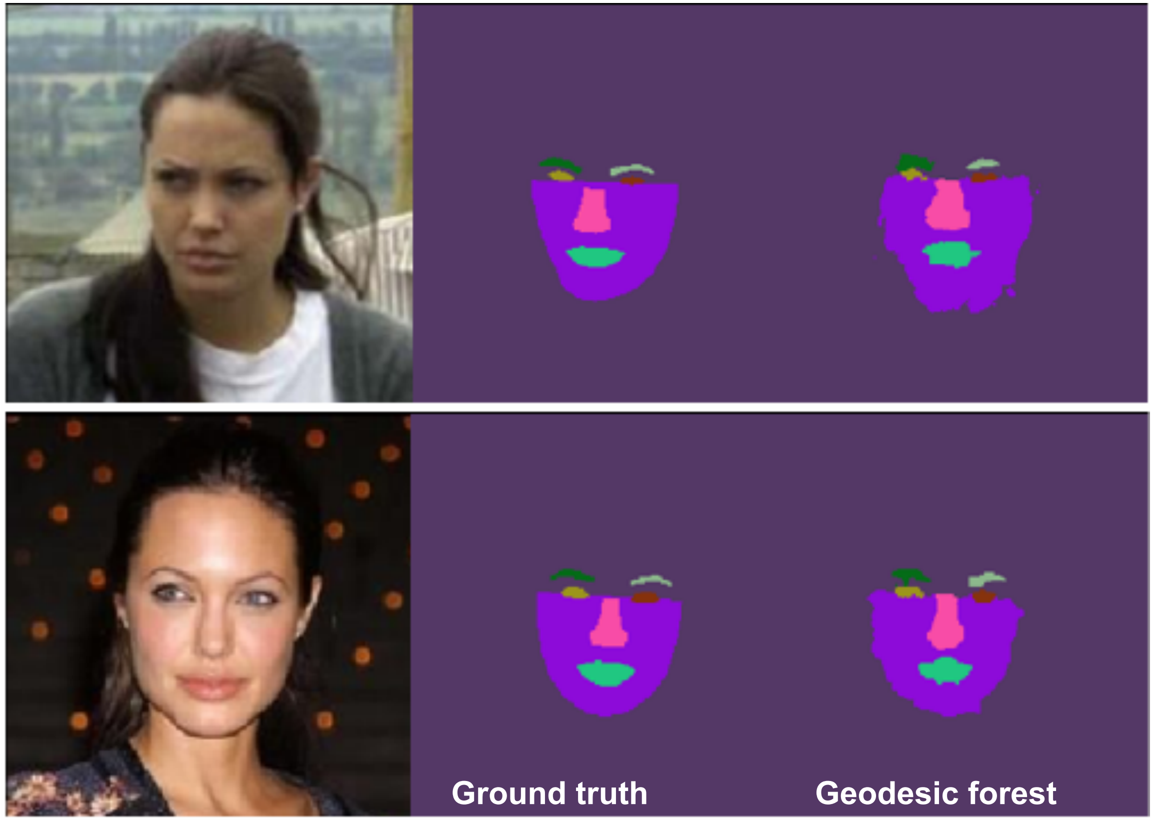
...

Qualitative results on the LFW dataset

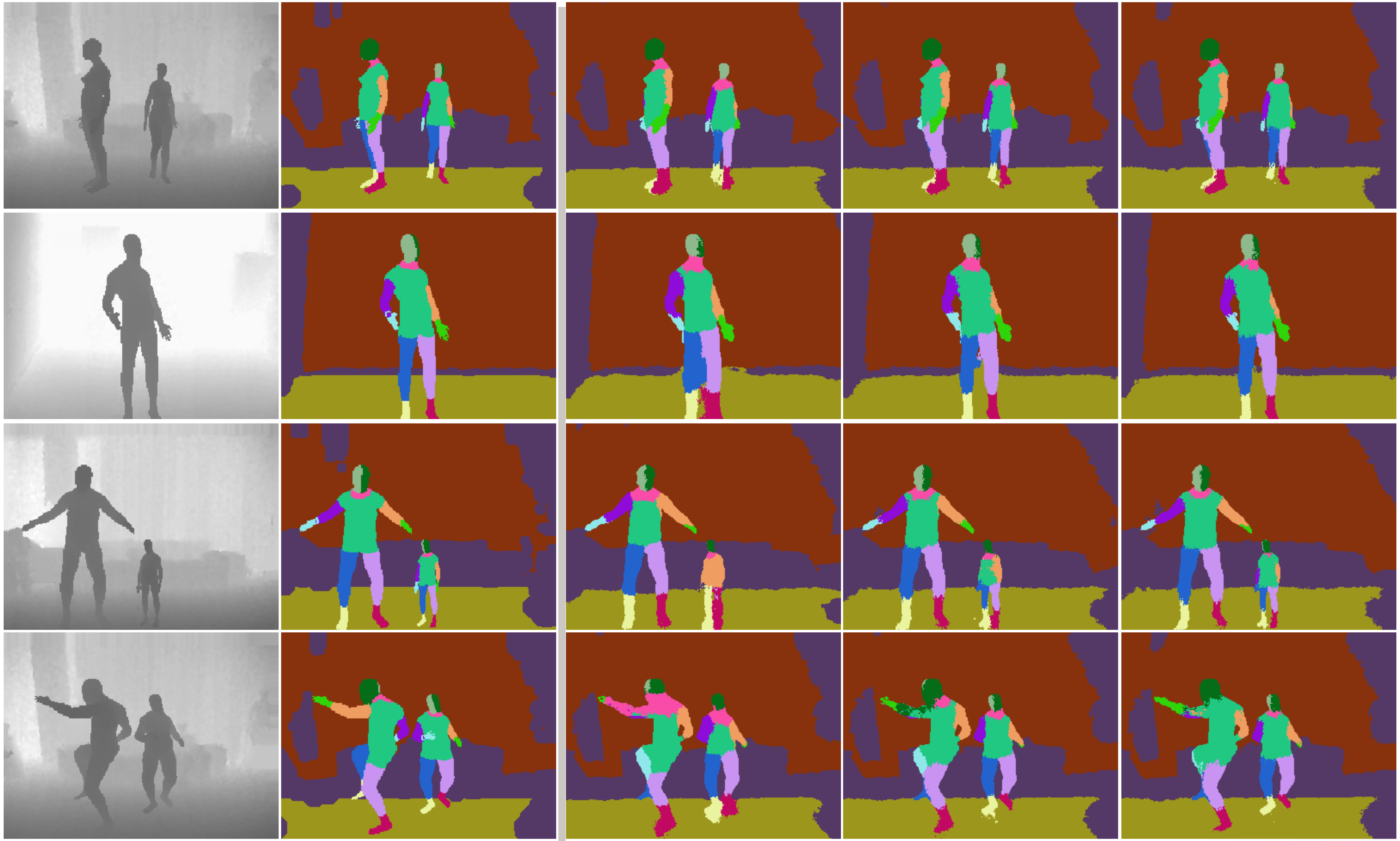


Qualitative results on LFW

Dealing with occlusions, pose
and illumination changes

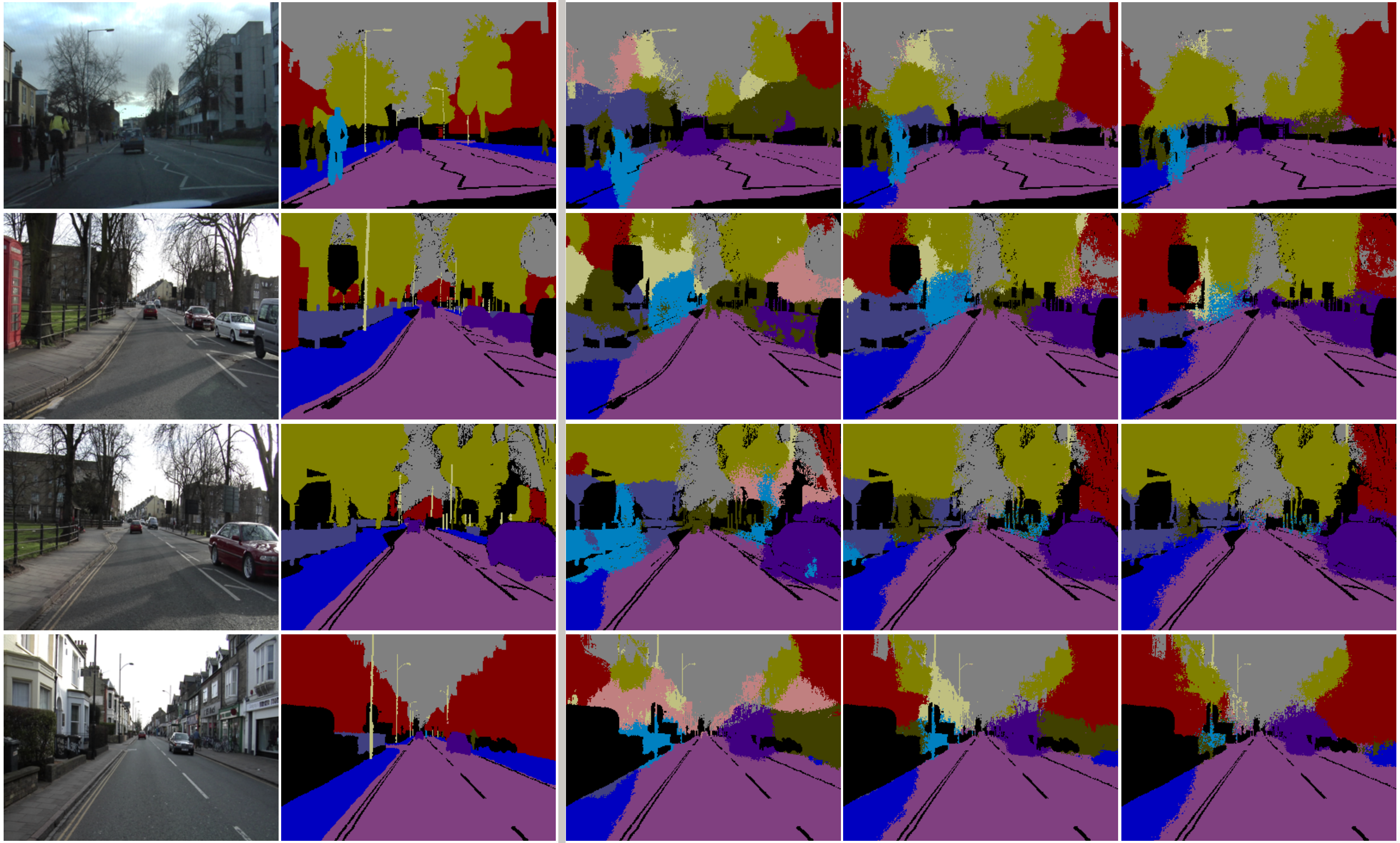


Qualitative results on the Kinect-BG dataset



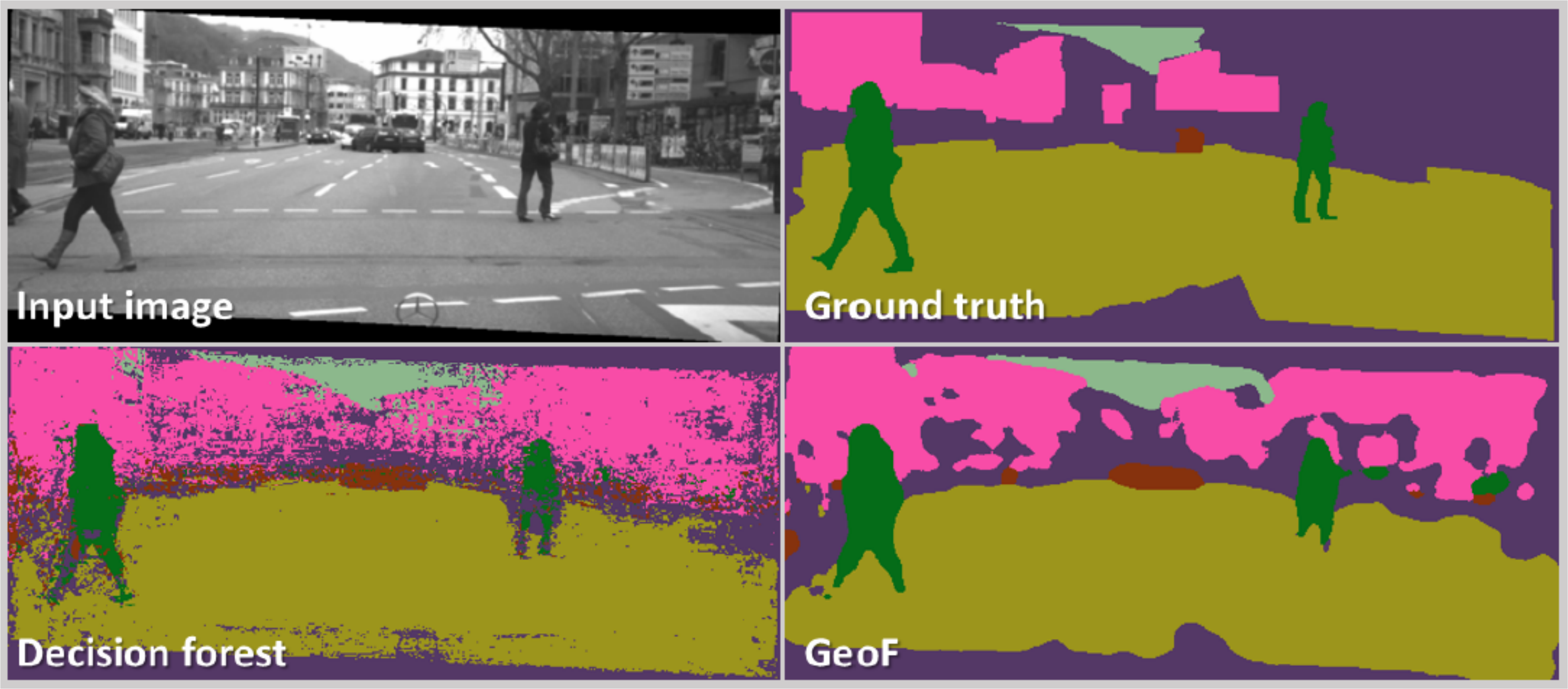
Input image Ground truth D=15 D=17 D=20

Qualitative results on the CamVid dataset

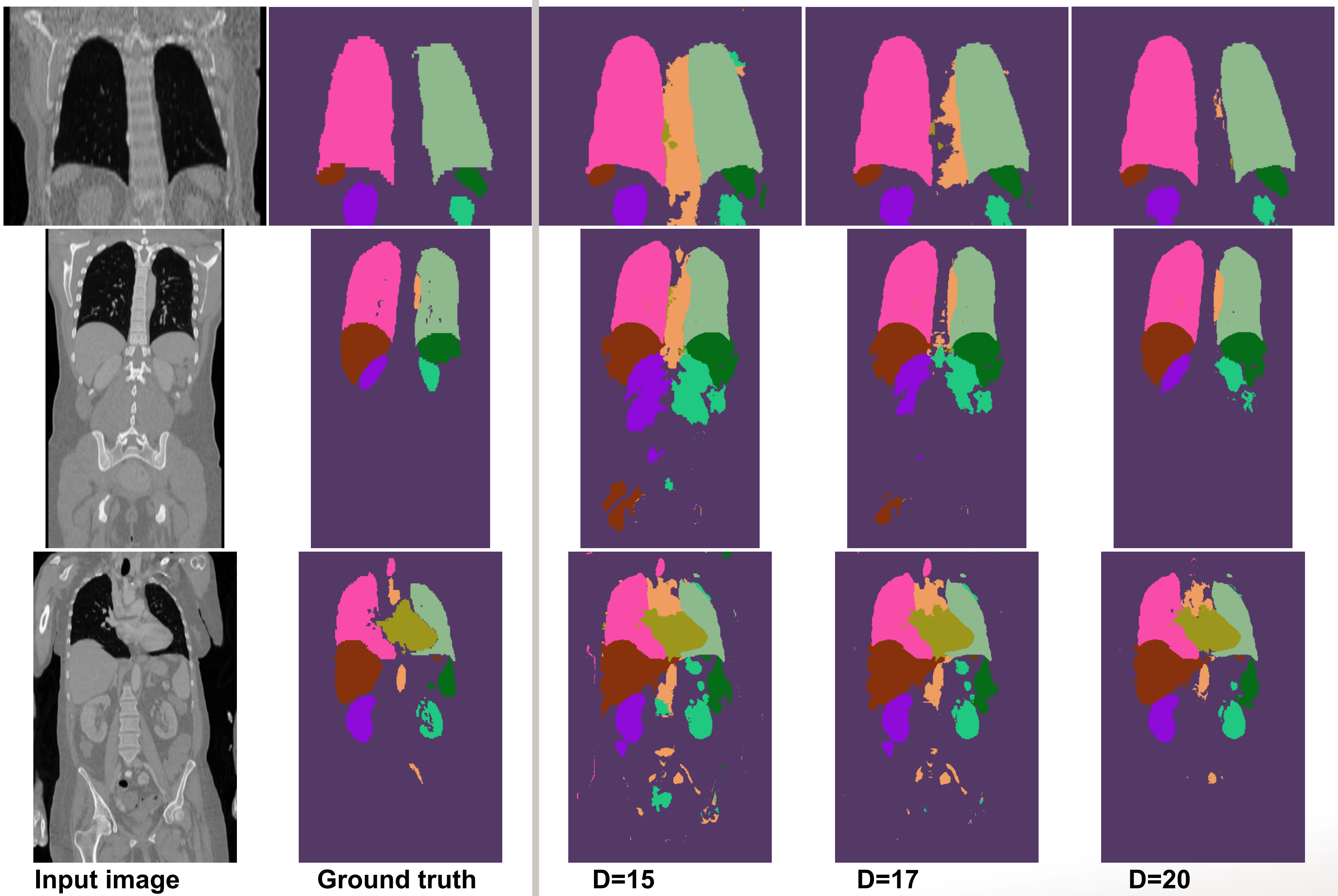


Input image Ground truth D=12 D=15 D=17

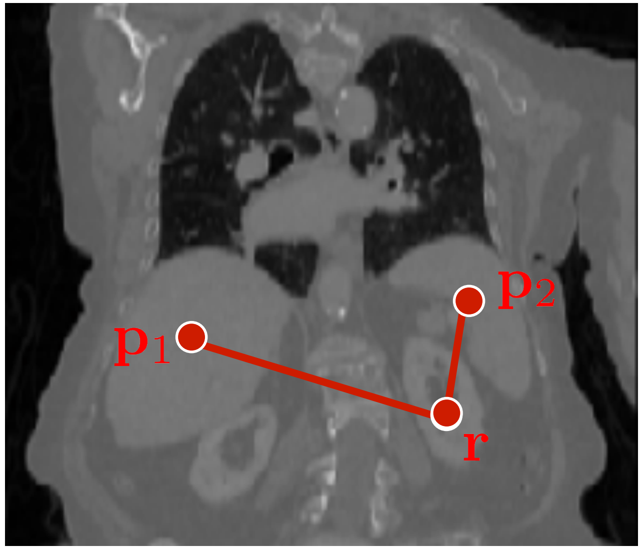
Qualitative results on Daimler dataset (stereo images)

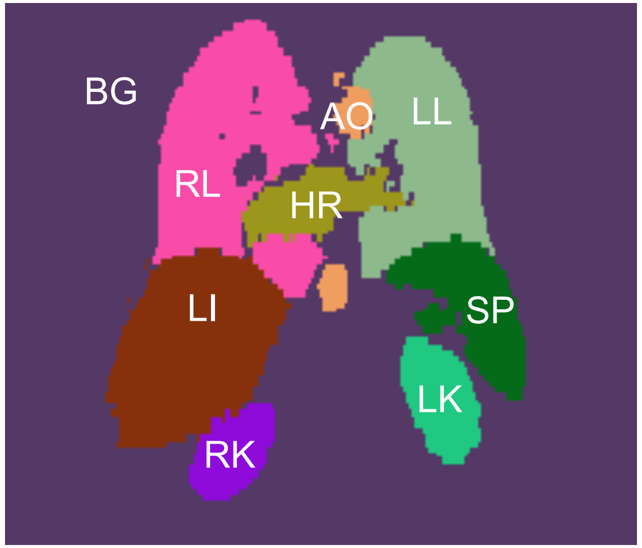


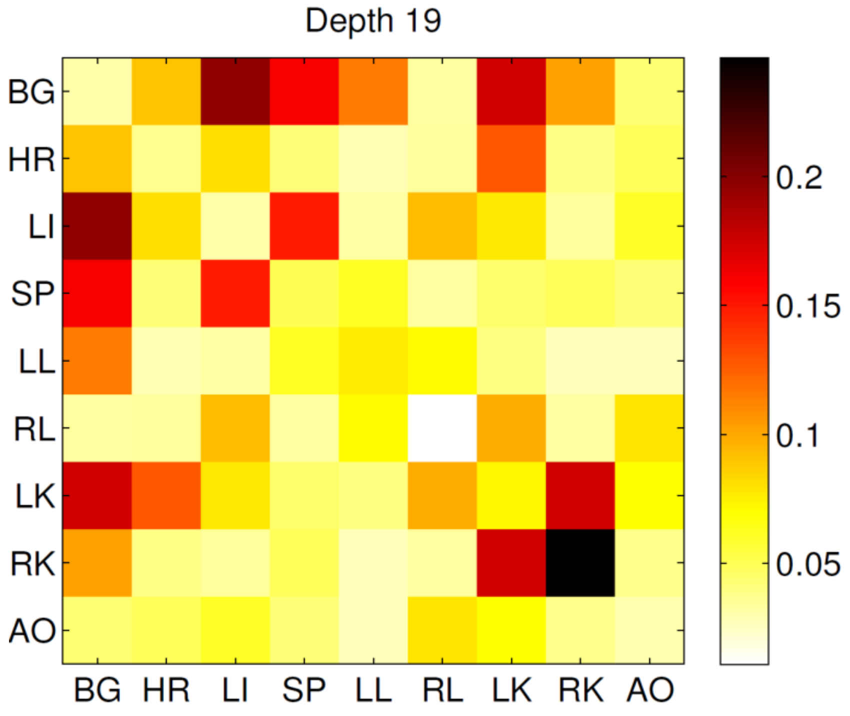
Qualitative results on the CT dataset: **the role of context**



Qualitative results on the CT dataset: the role of context







Selected, discriminative probe pairs, when reference on left kidney (LK)

Quantitative results: on 12 image datasets

	FC-8	FC-3	VC-2	VC-D
Decision forest	55.7, 96.7	73.5, 90.7	80.0, 92.4	23.1, 80.2
Dec. frst + CRF	59.3, 97.1	76.5, 92.1	87.5, 95.7	26.1, 83.5
Geodesic forest	**62.0, 97.4**	**77.6, 92.5**	**88.7, 96.3**	**27.5, 85.2**

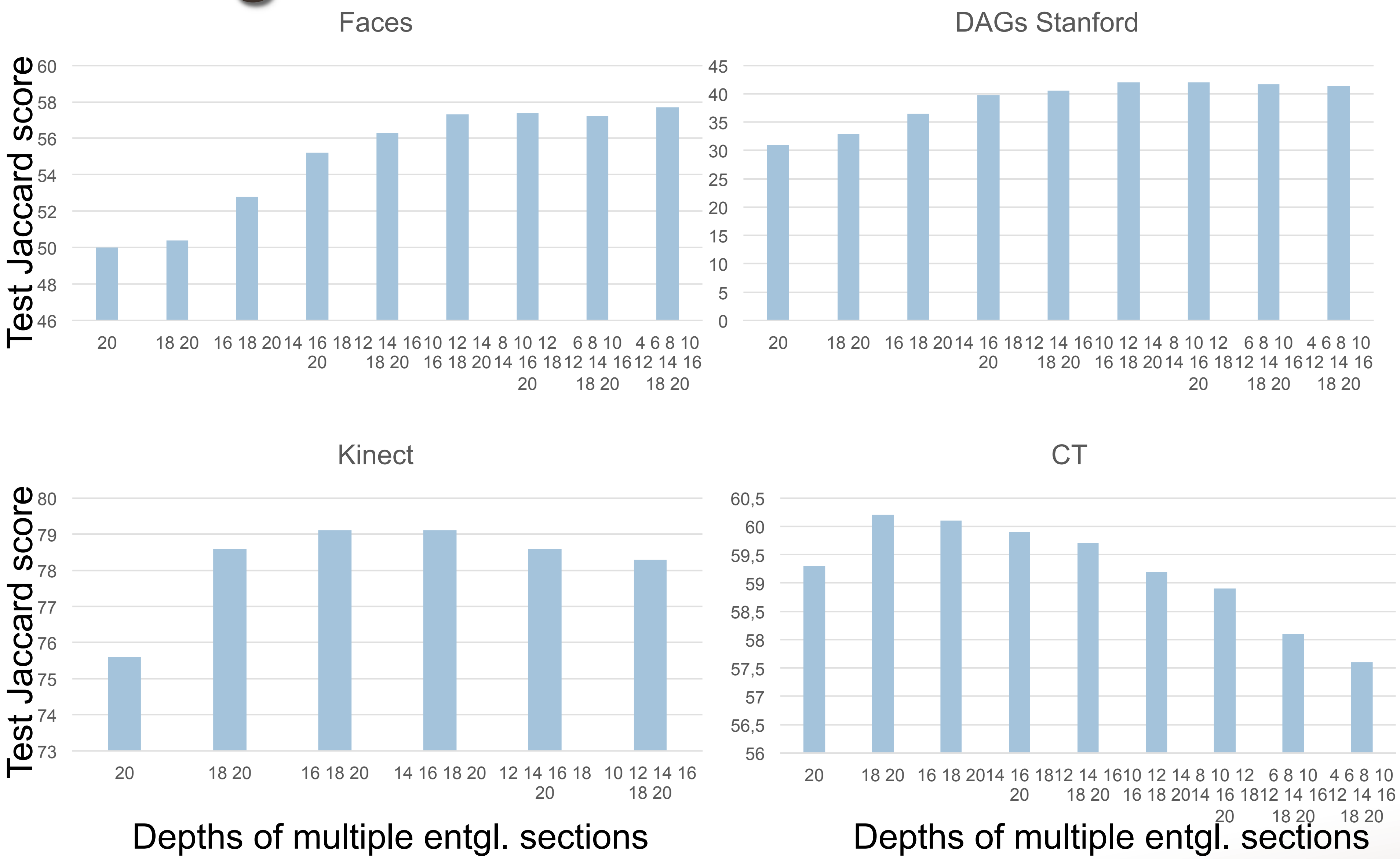
	SF-8	KI-3	CT-9	KT-15
Decision forest	39.0, 61.3	45.0, 90.1	64.0, 94.8	55.8, 90.6
Dec. frst + CRF	45.4, 69.5	50.7, 94.9	71.2, **96.1**	62.4, 92.2
Geodesic forest	**47.1, 70.5**	**56.8, 95.3**	**71.9**, 96.0	**62.6, 92.8**

	NY-15	CV-11	DA-6	DA-5
Decision forest	24.2, 50.3	33.5, 70.2	54.0, 73.0	71.4, 90.9
Dec. frst + CRF	30.3, 60.4	42.4, 80.4	57.3, 75.1	73.9, 91.9
Geodesic forest	**32.2, 61.8**	**43.8, 81.0**	**59.2, 76.4**	**78.0, 93.1**

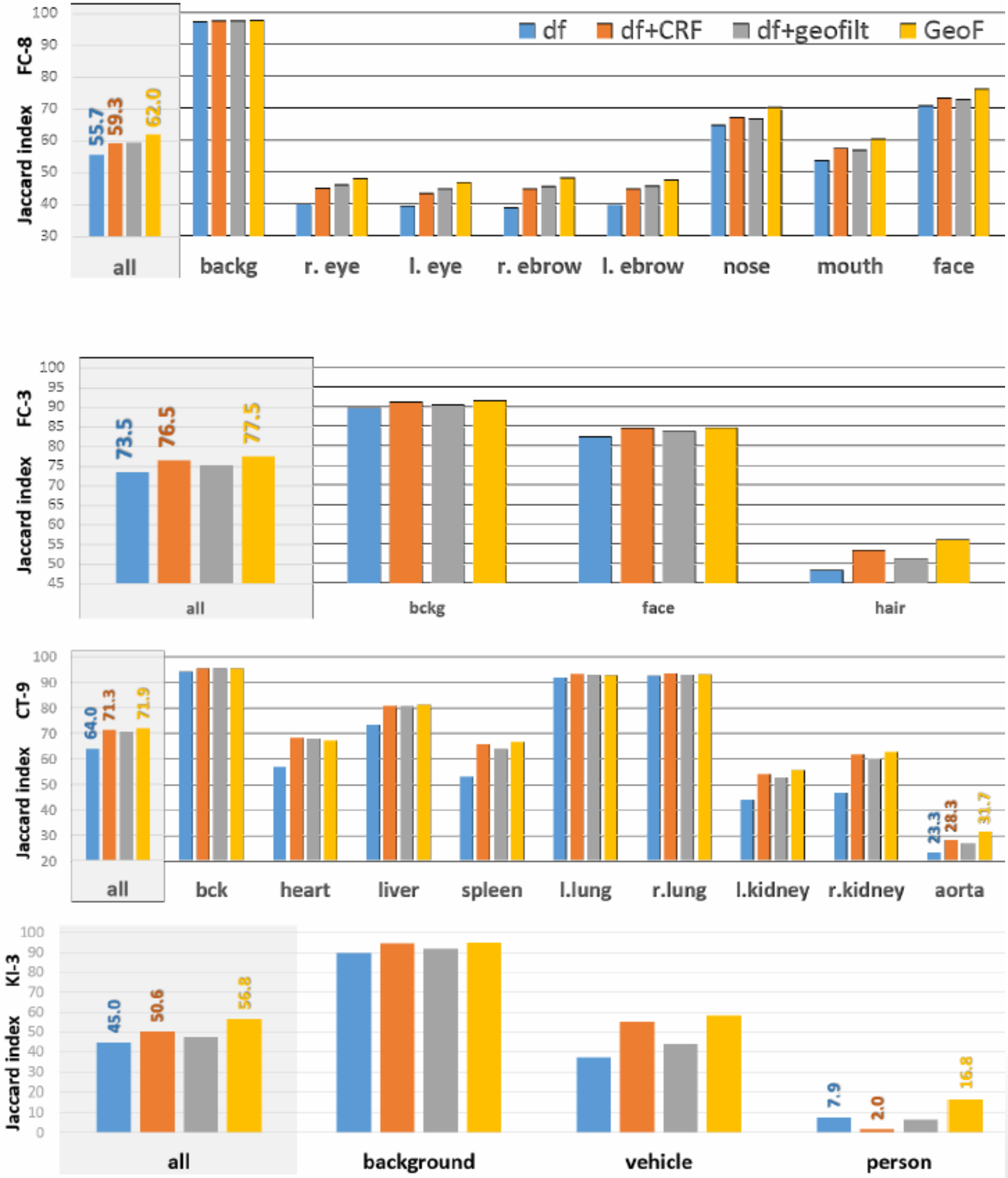
Jaccard / **Accuracy**

Free parameters optimized
Individually for each algorithm.

The effect of geodesic entanglament



Class-based analysis

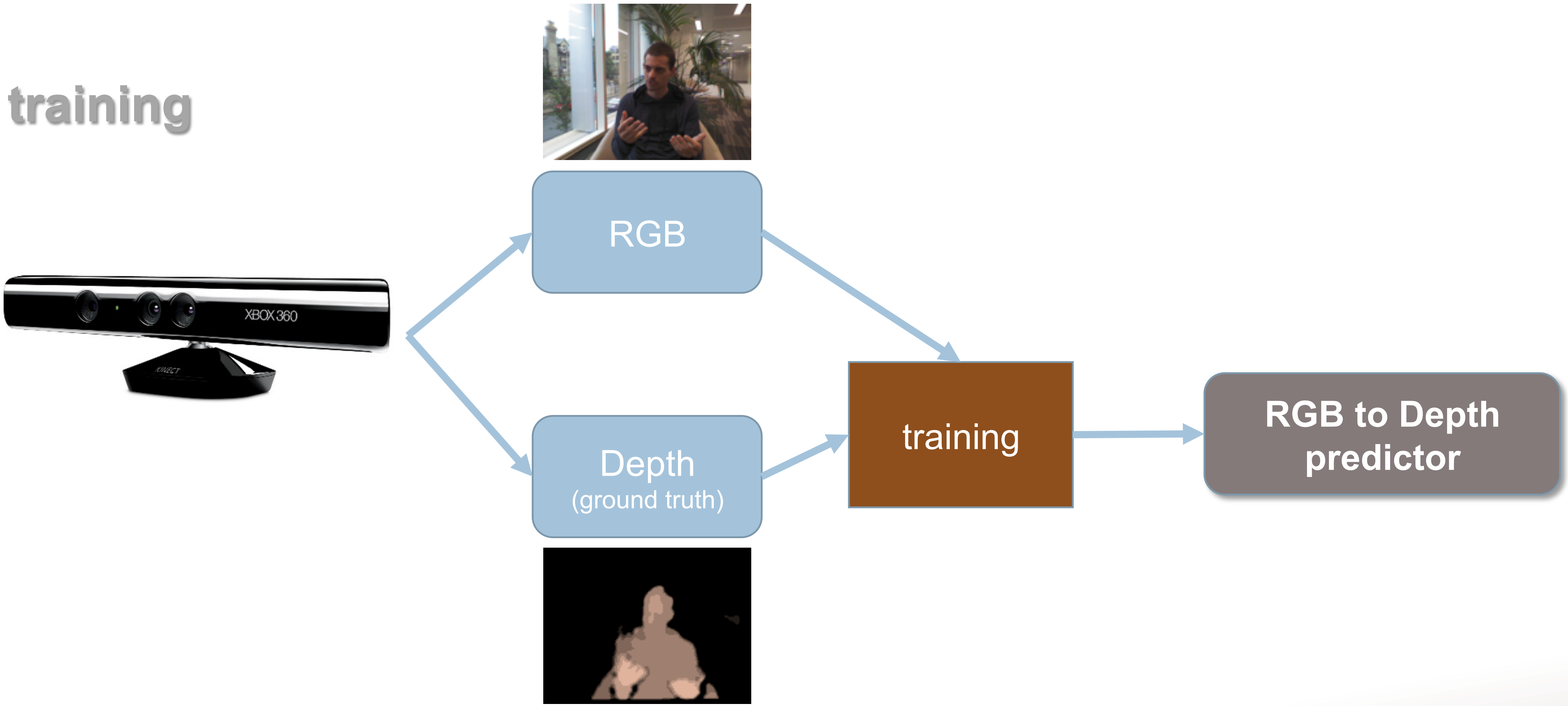


Geodesic forests help more on the more difficult classes.
e.g. the classes with fewer training pixels or thin and long.

One last application

Depth for Free

Turning conventional web cams into depth sensing devices



Turning conventional web cams into depth sensing devices

runtime

Any conventional webcam

RGB → RGB to Depth predictor → **Depth (estimated)**

Preliminary results

- No extra hardware required. Just a web-cam
    - Low-cost
    - Application to mobile devices
- Real-time depth prediction

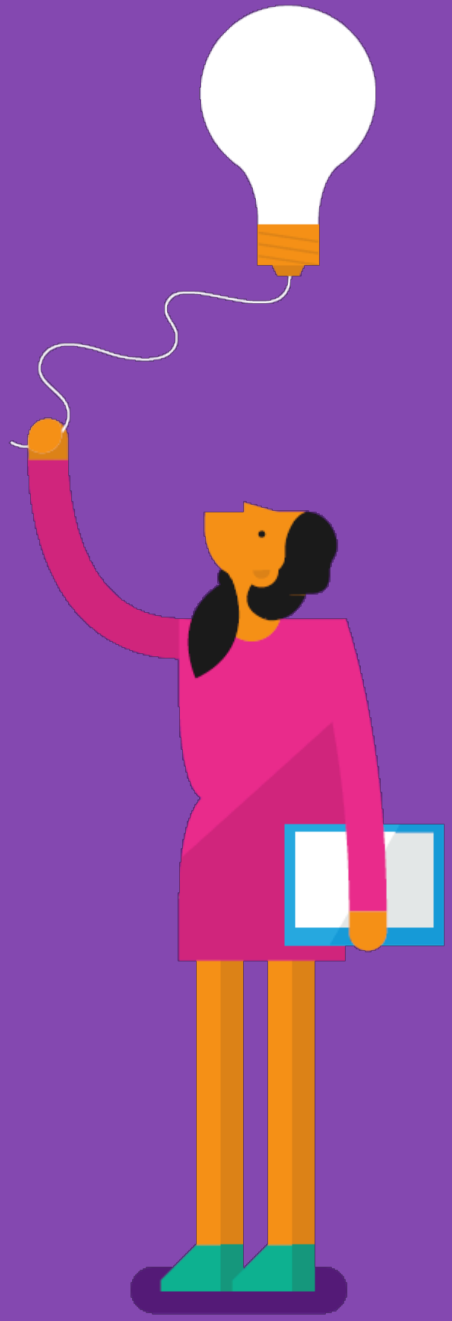
S.R. Fanello, C. Keskin, S. Izadi, P. Kohli, D. Kim, D. Sweeney, A. Criminisi, J. Shotton, S.B. Kang, T. Paek. **Learning to be a depth camera for close-range human capture and interaction.** In *ACM SIGGRAPH and Transaction On Graphics* 2014.

Summary

- **Deep learning** can be achieved with neural networks, decision forests and other classifiers too
- Here we have explored **entangled decision forests** with
  - Efficient, **soft connectivity** features
  - A new surrogate **training energy**
- State of the art results in **semantic segmentation** without the use of graph-based inference.
- Validated on a wide range of medical and non medical image and video **datasets**.

Microsoft Research Bright Minds Competition

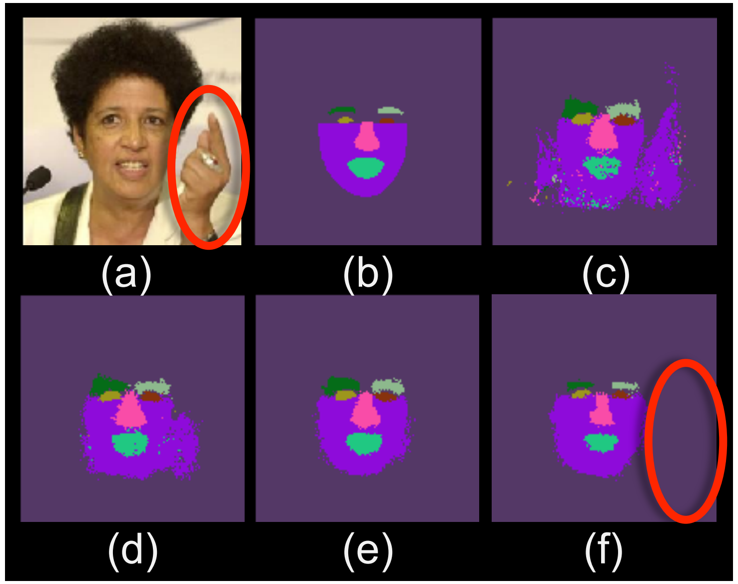
research.microsoft.com/undergrad

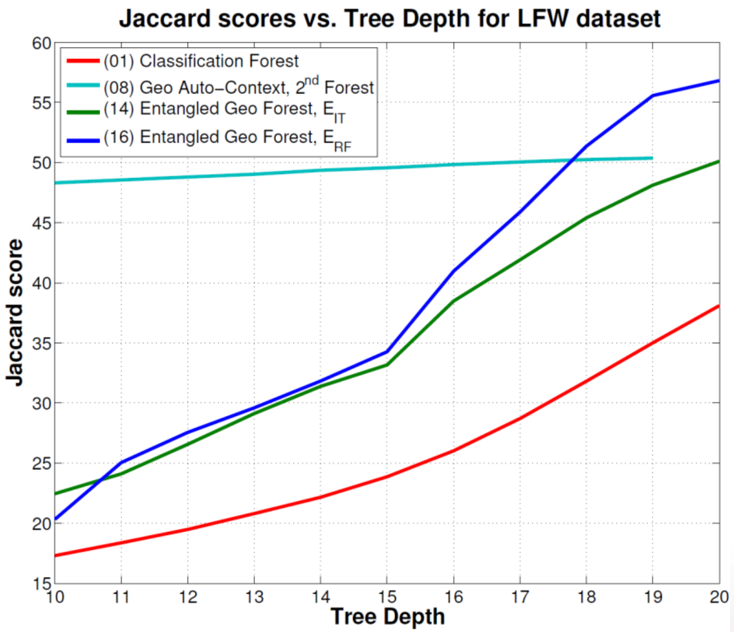


Labelled Faces in the Wild (LFW)

Algorithm	LFW
Number of training / testing imges	1000 / 250
Conventional Classification Forest (c)	38.1
Classification forest + (CRF)	45.2
Auto-context classification forest	48.1
Entangled classification forest (d)	43.2
Auto-context geodesic forests E_{IT}	50.4
Entangled geodesic forests (1 section) E_{IT}	46.2
Entangled geodesic forests (1 section) E_{RF}	54.6
Entangled geodesic forests (2 sections) (e) E_{IT}	50.1
Entangled geodesic forests (2 sections) (f) E_{RF}	**56.8**

Algorithm	Runtime (s/ frame)
Classification forest + (CRF)	0.71
Entangled geodesic forests (1 section)	0.42





Kinect + Background (KinBG)

Algorithm		KinBG
Number of training / testing imges		2500/ 250
Conventional Classification Forest		57.1
Classification forest + (CRF)		60.0
Auto-context classification forest		61.9
Entangled classification forest		55.7
Auto-context geodesic forests	E_{IT}	**63.9**
Entangled geodesic forests (1 section)	E_{IT}	55.4
Entangled geodesic forests (1 section)	E_{RF}	60.0
Entangled geodesic forests (2 sections)	E_{IT}	56.8
Entangled geodesic forests (2 sections)	E_{RF}	60.3

Algorithm	Runtime (s/frame)
Classification forest + (CRF)	1.35
Auto-context geodesic forests	1.39
Entangled geodesic forests (1 section)	0.64

CamVid Dataset

Algorithm	CamVid
Number of training / testing imges	367/ 233
Conventional Classification Forest	33.3
Classification forest + (CRF)	**41.7**
Auto-context classification forest	35.2
Entangled classification forest	35.5
Structured class-labels in RF's [ICCV'11]	36.2
Local label descriptors [ECCV'12]	29.6
Auto-context geodesic forests E_{IT}	36.6
Entangled geodesic forests (1 section) E_{IT}	35.1
Entangled geodesic forests (1 section) E_{RF}	37.7
Entangled geodesic forests (2 sections) E_{IT}	38.0
Entangled geodesic forests (2 sections) E_{RF}	38.3

Algorithm	Runtime
Classification forest + (CRF)	1.07
Entangled geodesic forests (1 section)	0.56

2D Computed Tomography (CT)

Algorithm		CT
Number of training / testing imges		512/ 250
Conventional Classification Forest		53.2
Classification forest + (CRF)		68.3
Auto-context classification forest		65.9
Entangled classification forest		58.3
Auto-context geodesic forests	$E_{\mathtt{IT}}$	69.2
Entangled geodesic forests (1 section)	$E_{\mathtt{IT}}$	60.2
Entangled geodesic forests (1 section)	$E_{\mathtt{RF}}$	**72.3**
Entangled geodesic forests (2 sections)	$E_{\mathtt{IT}}$	61.1
Entangled geodesic forests (2 sections)	$E_{\mathtt{RF}}$	72.2

Algorithm	Runtime (s/ frame)
Classification forest + (CRF)	1.20
Entangled geodesic forests (1 section)	0.72

Input

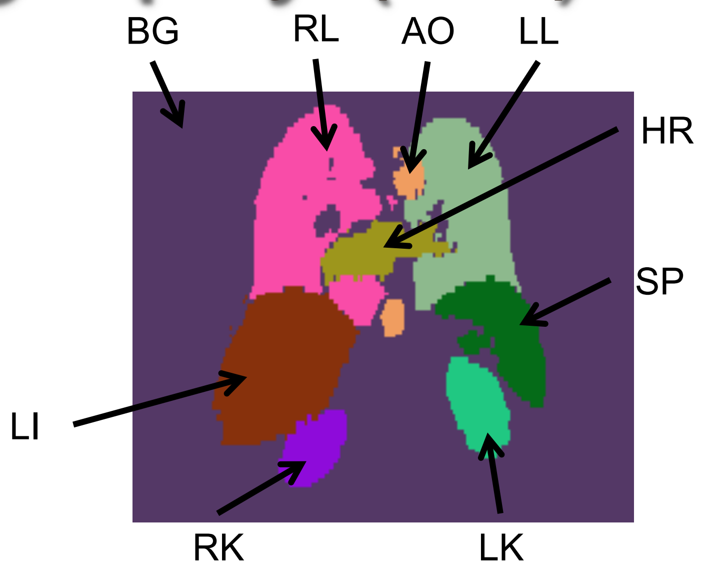
Ground truth

Our result

2D Computed tomography (CT)

Algorithm		CT
Number of training / testing imges		512/ 250
Conventional Classification Forest		53.2
Classification forest + (CRF)		68.3
Auto-context classification forest		65.9
Entangled classification forest		58.3
Auto-context geodesic forests	E_{IT}	69.2
Entangled geodesic forests (1 section)	E_{IT}	60.2
Entangled geodesic forests (1 section)	E_{RF}	**72.3**
Entangled geodesic forests (2 sections)	E_{IT}	61.1
Entangled geodesic forests (2 sections)	E_{RF}	72.2

Algorithm	Runtime (s/frame)
Classification forest + conventional CRF	1.20
Entangled geodesic forests (1 section)	0.72



Selected, discriminative probe pairs, when reference on left kidney (LK)