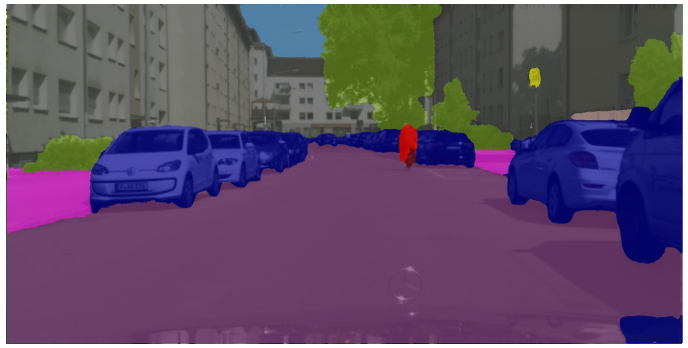
Scene Understanding with Deep Structured Models

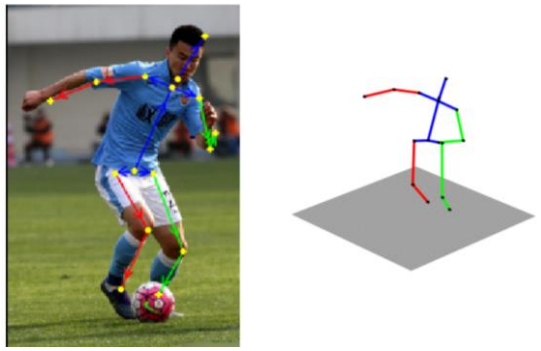
Anurag Arnab

Scene Understanding

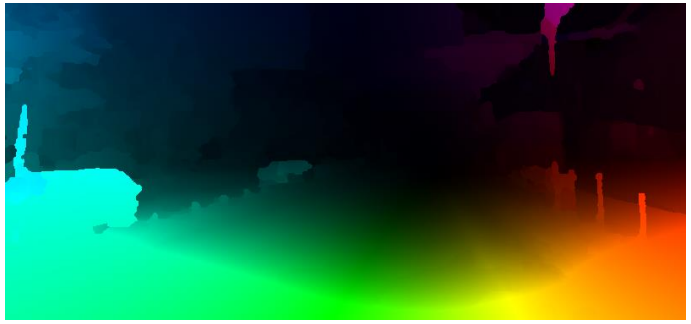


Semantic segmentation

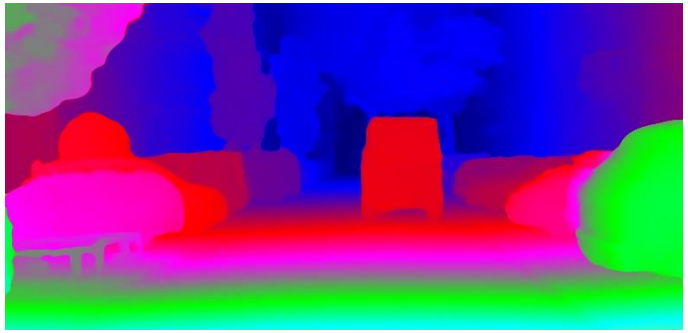
Instance/Panoptic segmentation



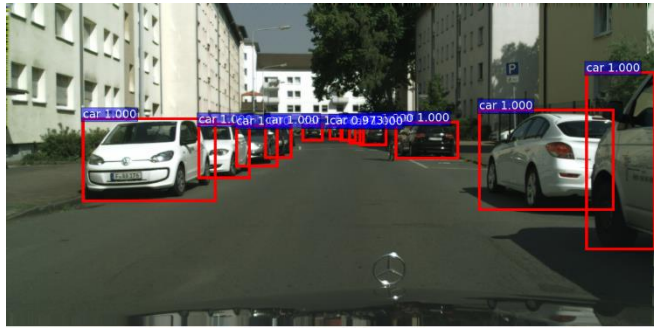
Pose estimation



Optical flow

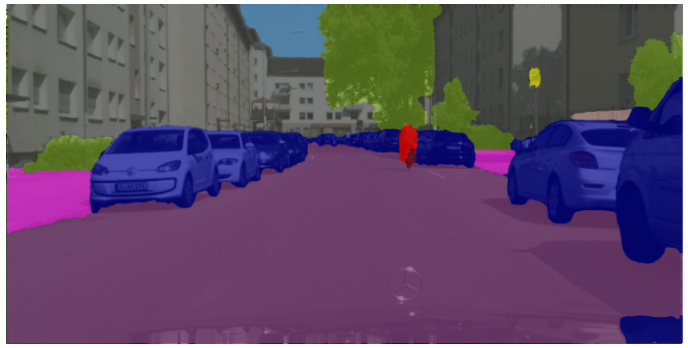


Stereo



Object detection

Scene Understanding

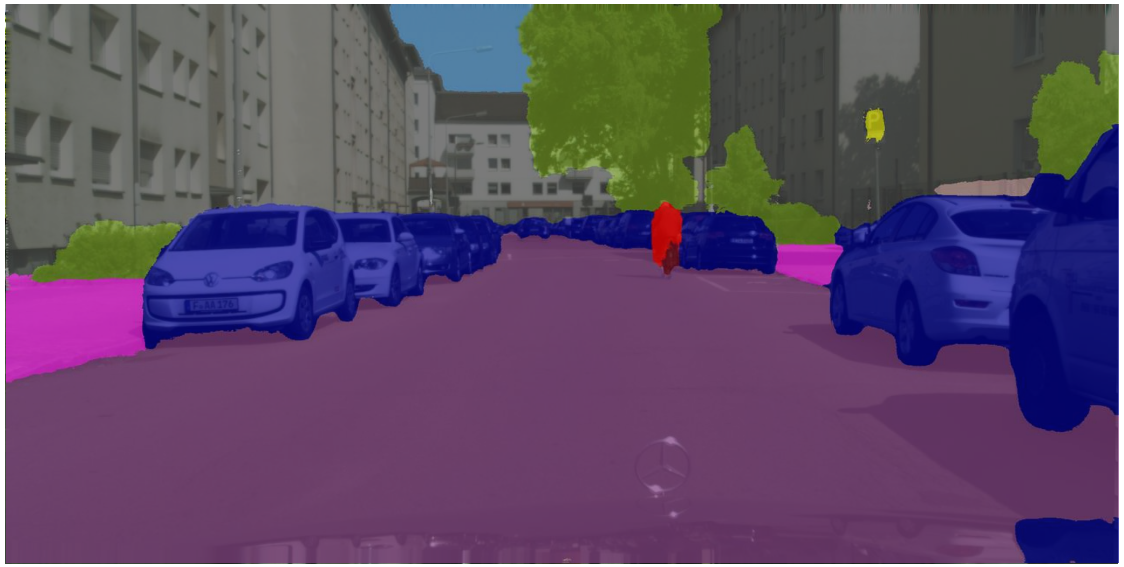


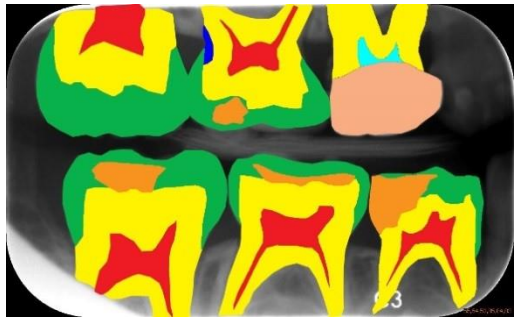
Semantic segmentation

Instance/Panoptic segmentation

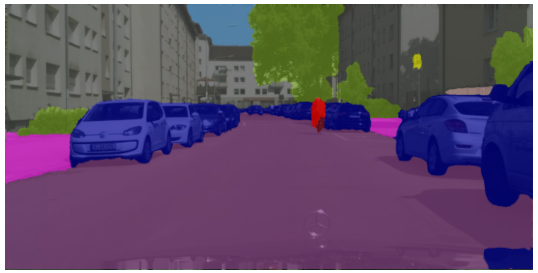
Why Pixel-level Scene Understanding

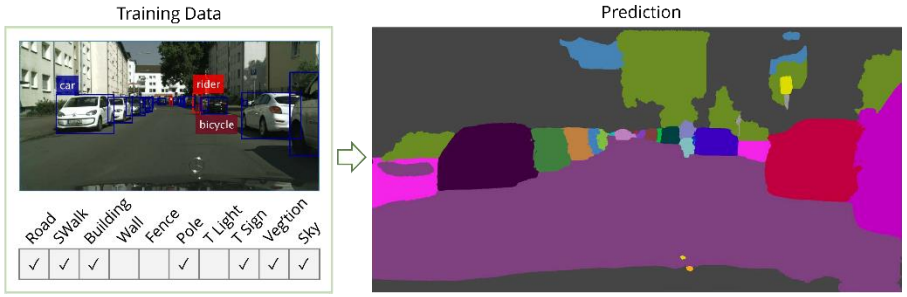
- Autonomous navigation
- Assisting the partially sighted
- AR / VR
- Medical diagnosis
- Image editing
- Many more …





Outline





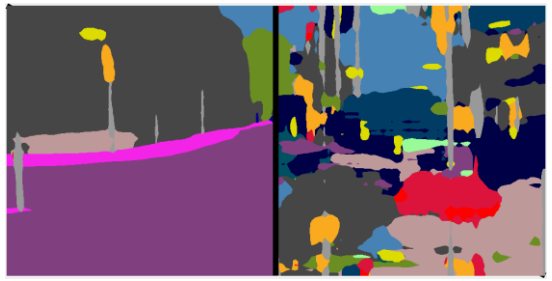
Higher Order CRFs in Deep Neural Networks (ECCV 2016)

Pixelwise Instance Segmentation with a Dynamically Instantiated Network (CVPR 2017)

Weakly and Semi-Supervised Panoptic Segmentation (ECCV 2018)

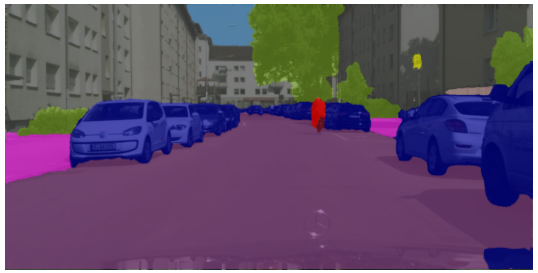
Outline

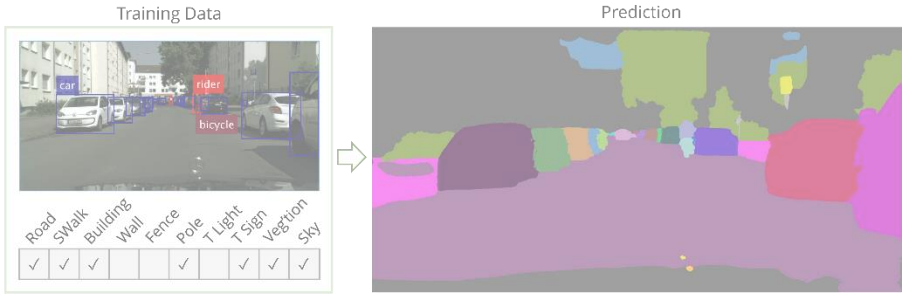
Exploiting Temporal Context for 3D Pose Estimation in the Wild (CVPR 2019) – DeepMind internship



On the Robustness of Semantic Segmentation Models to Adversarial Attacks (CVPR 2018)

Outline





Higher Order CRFs in Deep Neural Networks (ECCV 2016)

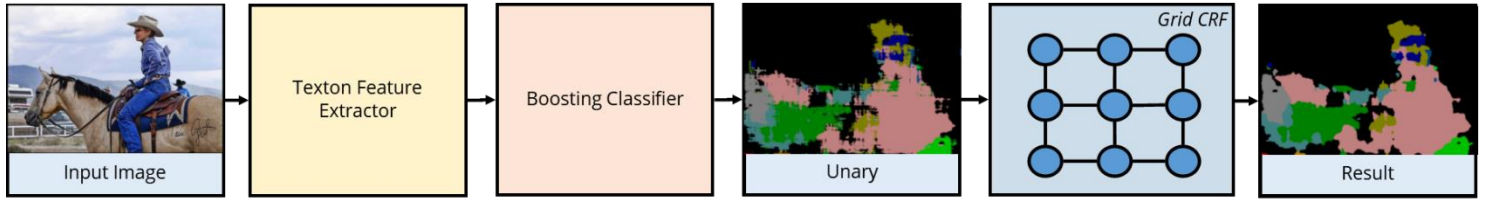
Pixelwise Instance Segmentation with a Dynamically Instantiated Network (CVPR 2017)

Weakly and Semi-Supervised Panoptic Segmentation (ECCV 2018)

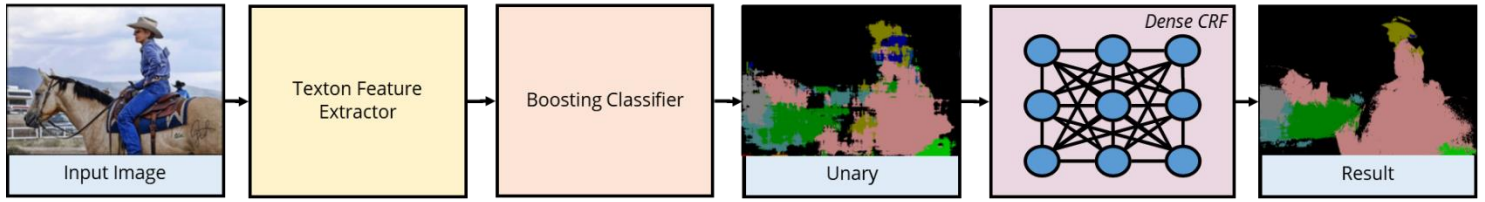
Structured Prediction

- Structured prediction (predicting many correlated variables – ie pixels)

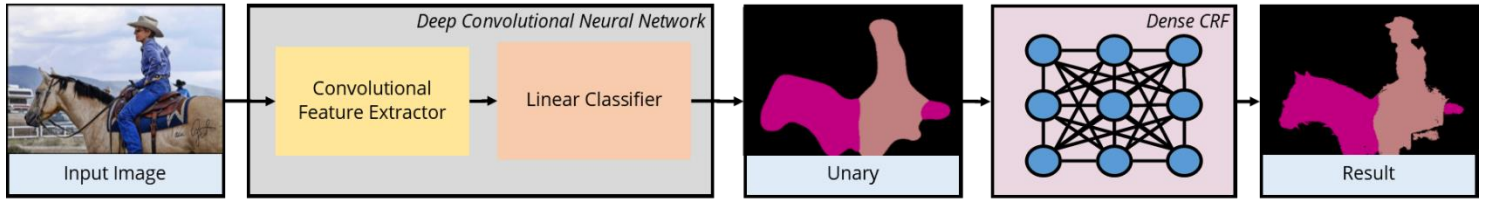
Ladicky et al. ICCV 2009



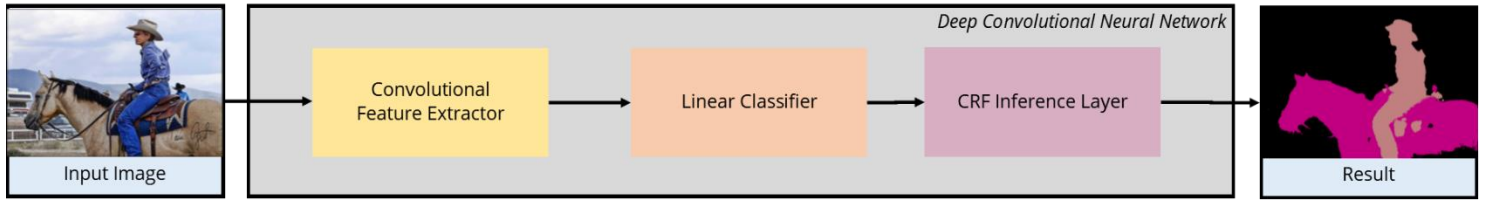
Krahenbuhl and Koltun NIPS 2011



Chen et al. ICLR 2015



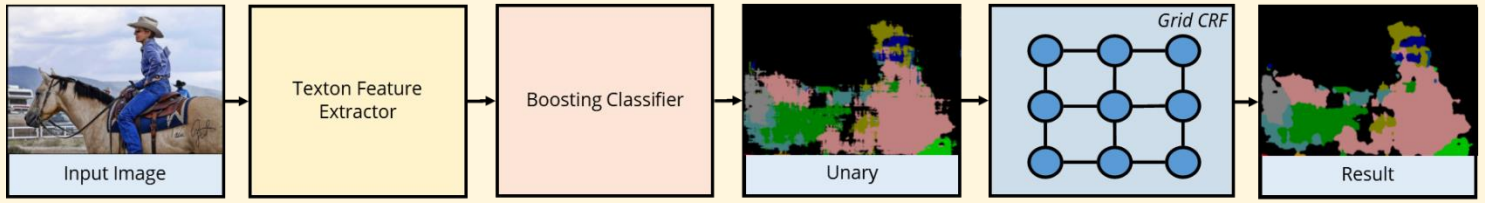
Arnab et al. ECCV 2016



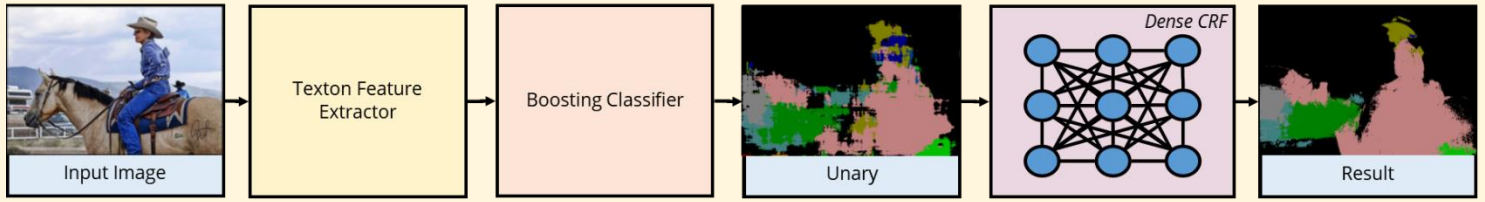
Structured Prediction

- Structured prediction (predicting many correlated variables – ie pixels)

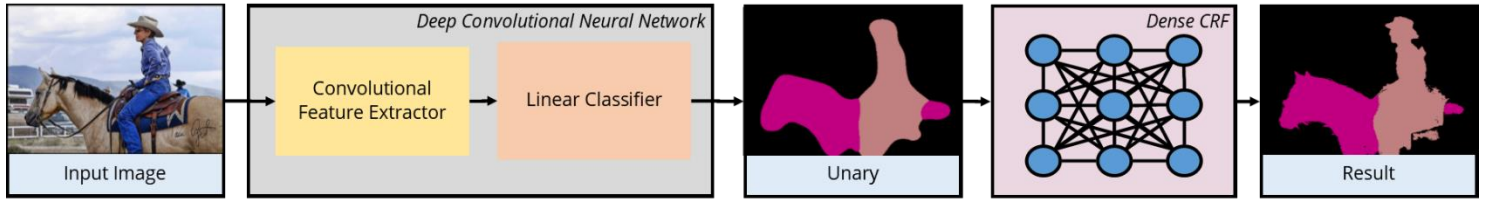
Ladicky et al. ICCV 2009



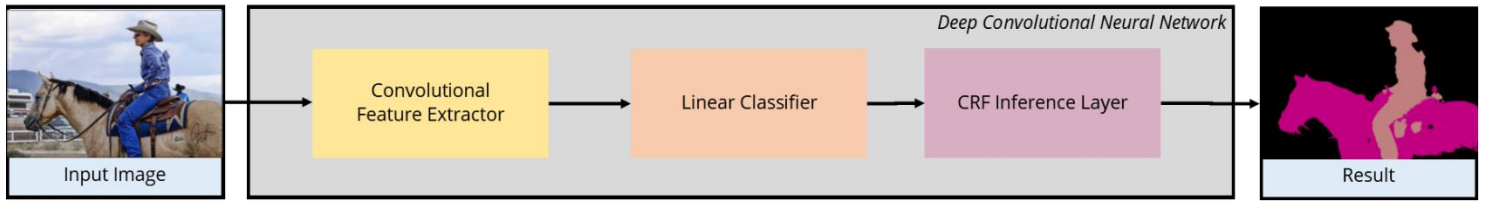
Krahenbuhl and Koltun NIPS 2011



Chen et al. ICLR 2015



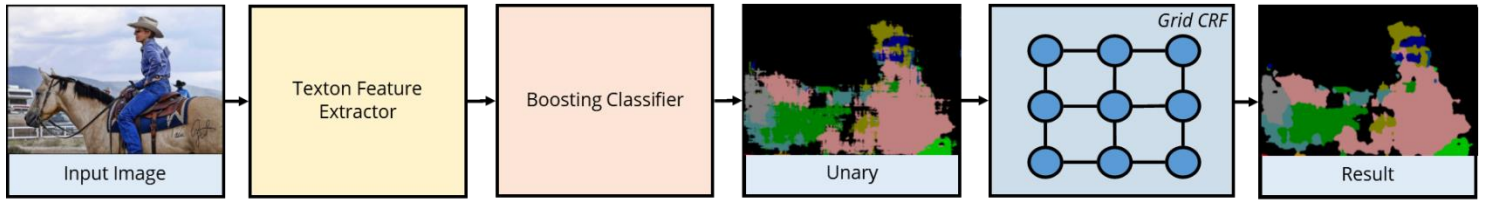
Arnab et al. ECCV 2016



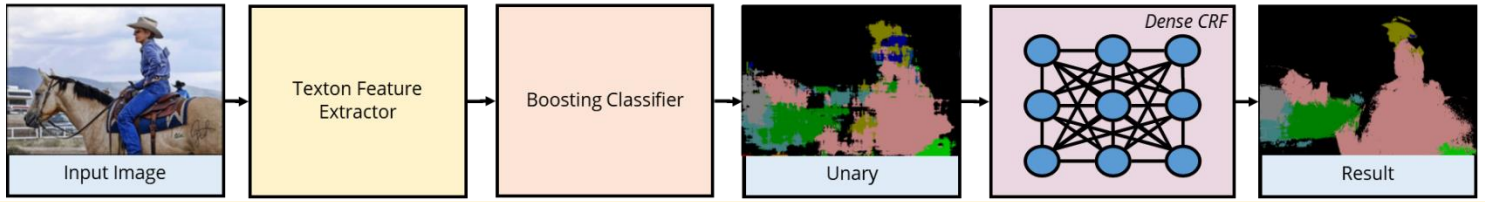
Structured Prediction

- Structured prediction (predicting many correlated variables – ie pixels)

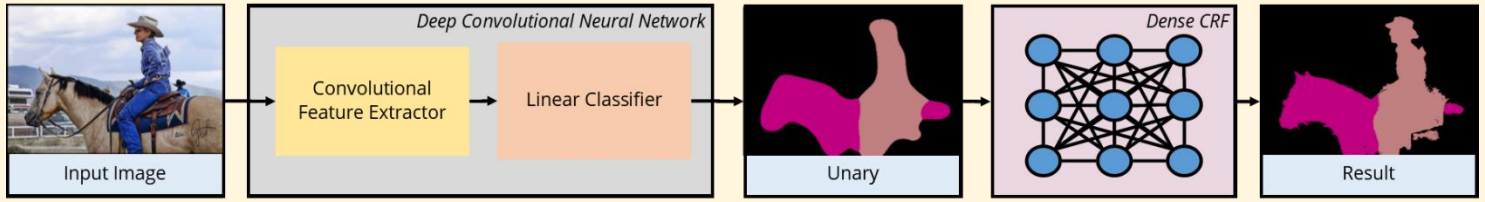
Ladicky et al. ICCV 2009



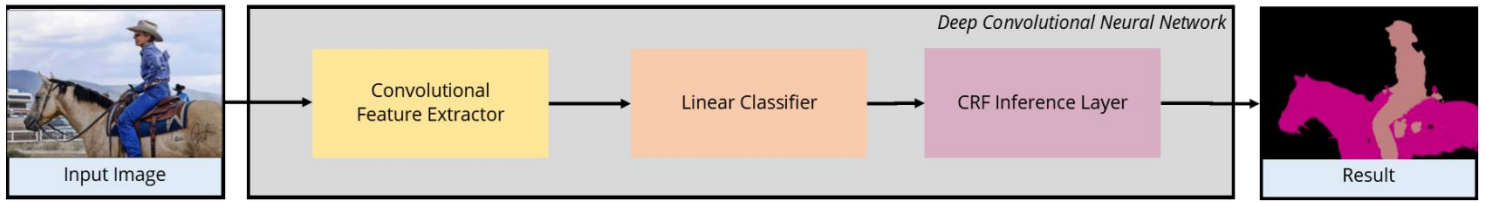
Krahenbuhl and Koltun NIPS 2011



Chen et al. ICLR 2015



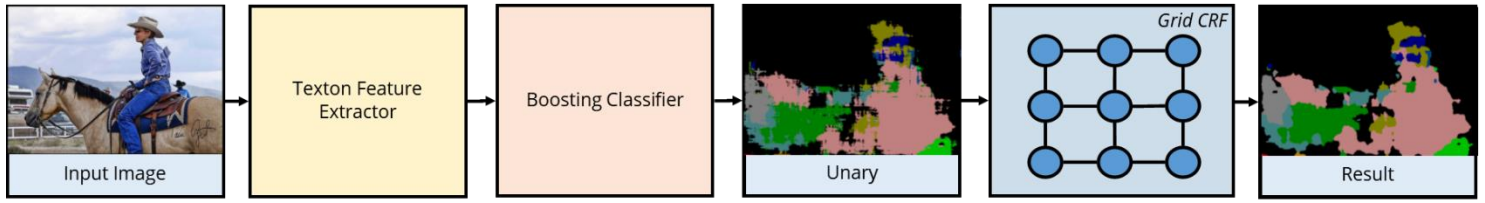
Arnab et al. ECCV 2016



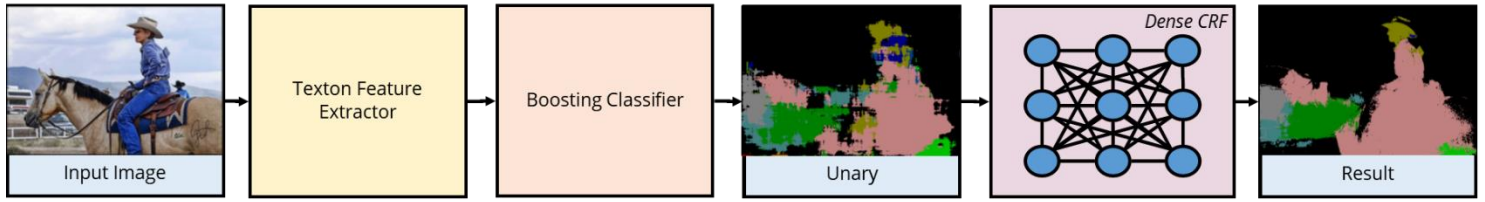
Structured Prediction

- Structured prediction (predicting many correlated variables – ie pixels)

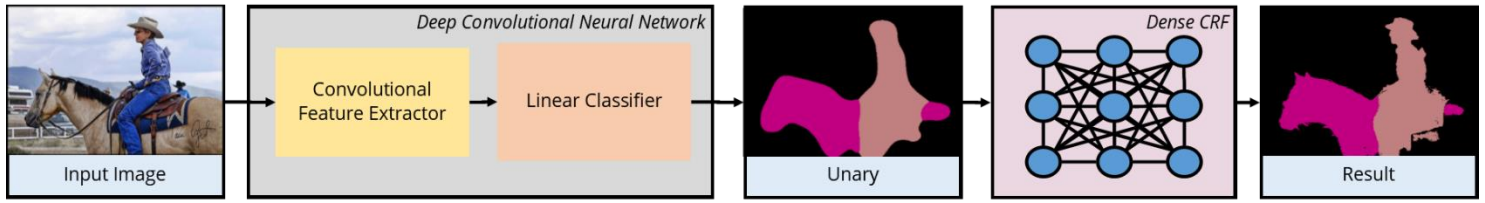
Ladicky et al. ICCV 2009



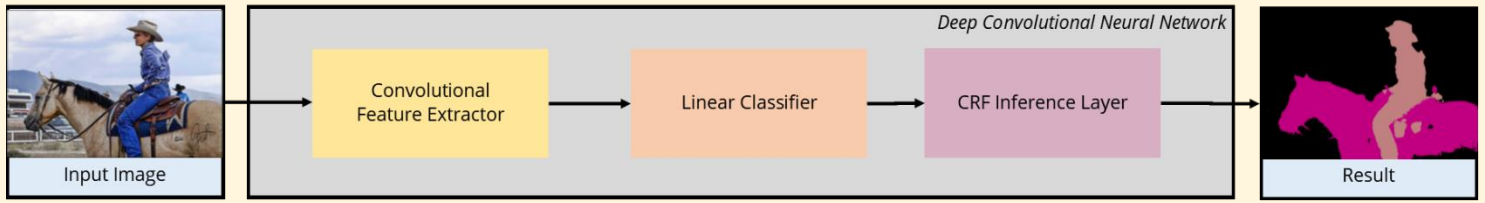
Krahenbuhl and Koltun NIPS 2011



Chen et al. ICLR 2015

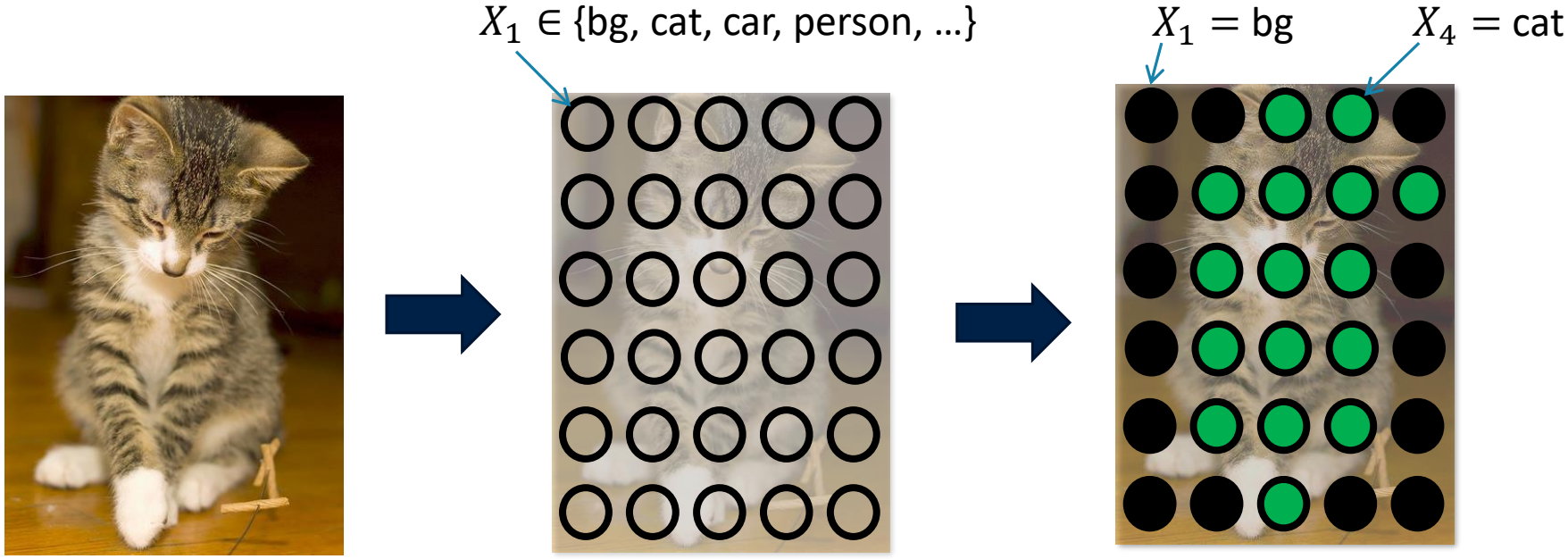


Arnab et al. ECCV 2016



Zheng *et al.* ICCV 2015. Schwing and Urtasun. arXiv 2015.

Conditional Random Fields



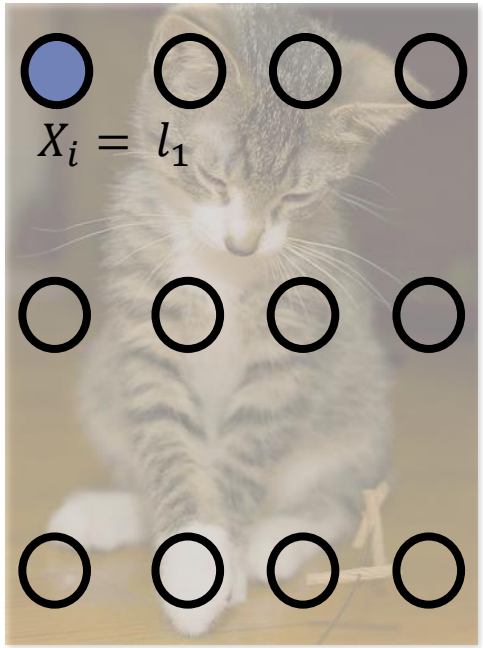
- Define a discrete random variable, X_i, for each pixel i
- Each X_i takes a value from the label set $\mathcal{L}$
- The random variables are connected to form a random field. The most probable assignment, conditioned on the image, is our semantic segmentation result.

The Best Assignment

$$P(X_1 = x_1, X_2 = x_2, \dots, X_N = x_n) = P(\boldsymbol{X} = \boldsymbol{x}\,|I)$$

$$P(\boldsymbol{X} = \boldsymbol{x}\,|I) = \frac{1}{Z}\exp(-E(\boldsymbol{x}\,|I)$$

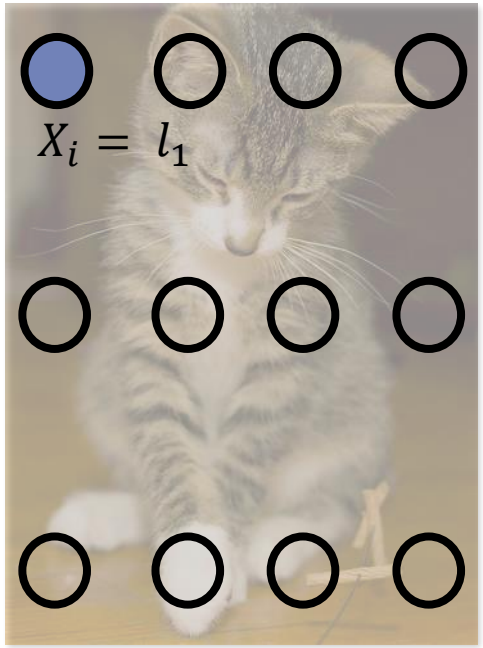
- Maximising the probability, is equivalent to minimising the energy of the CRF.



Energies

- Unary
  - Your final label does not agree with the initial classifier

    → you pay a penalty
  - In our case, the initial classifier is FCN [1]
- Pairwise
- Detection
- Superpixels

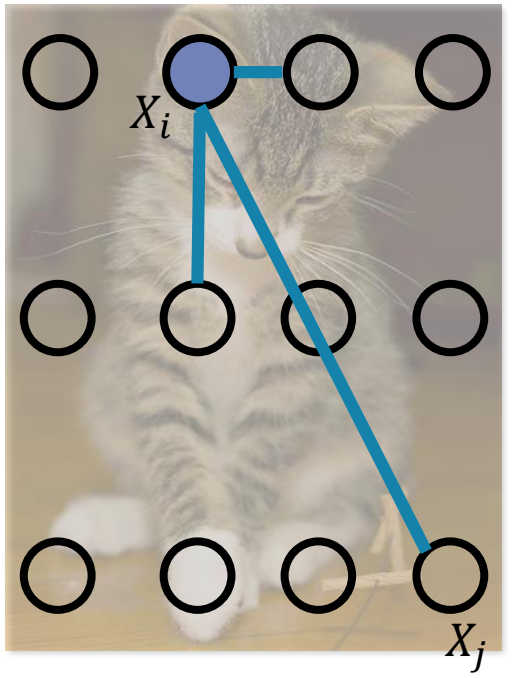


[1] Long *et al.* CVPR 2015

Energies

- Unary
- Pairwise
  - You assign different labels to two very similar pixels

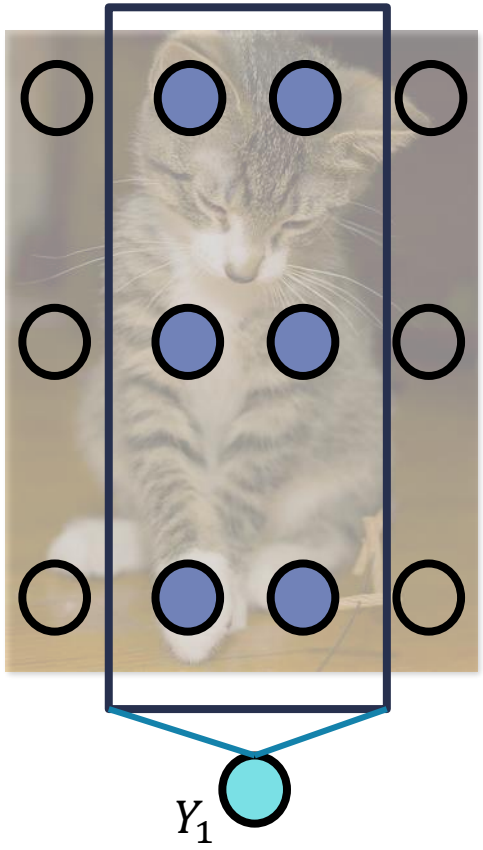
  → you pay a penalty
  - How do you measure similarity?
  - DenseCRF [1]
  - $\Psi_{i,j}^{P}(x_i, x_j) = w_1 \exp\left(-\frac{|p_i - p_j|^2}{2\sigma_\alpha^2} - \frac{|I_i - I_j|^2}{2\sigma_\beta^2}\right) + w_2 \exp\left(-\frac{|p_i - p_j|^2}{2\sigma_\gamma^2}\right)$
- Detection
- Superpixels



[1] Krahenbuhl and Koltun, NIPS, 2011

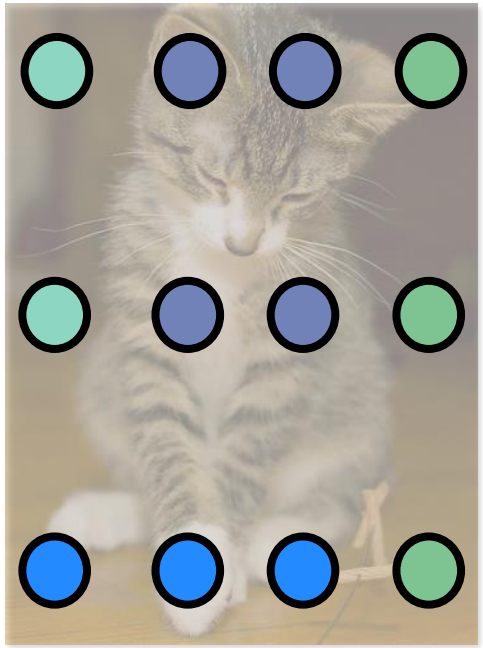
Energies

- Unary
- Pairwise
- Detection (Higher order)
  - Inference with pairwise potentials cannot help if unaries are poor
  - Cues from *object detectors* can help in this regard
  - Object detectors can "fire" over regions which have poor/incorrect unaries
  - Want our potential to be robust to false-positive detections
  - Introduce additional latent, Y variables which model whether the detection hypothesis is accepted or not
  - $$\psi_d^{Det}(\boldsymbol{X_d} = \boldsymbol{x_d}, Y_d = y_d) = \begin{cases} w_{det} \frac{s_d}{n_d} \sum_{i=1}^{n_d} \left[x_d^{(i)} = l_d \right] & if\, y_d = 0 \\ w_{det} \frac{s_d}{n_d} \sum_{i=1}^{n_d} \left[x_d^{(i)} \neq l_d \right] & if\, y_d = 1 \end{cases}$$
- Superpixels



Energies

- Unary
- Pairwise
- Detections
- Superpixels
  - Enforce consistency over entire regions obtained by superpixels
  - P^n-Potts model type energy [1,2]
  - Low cost if all the random variables within a superpixel are assigned the same label. High cost otherwise
  - Reduces spurious noise in segmentations
  - $$\psi_s^{SP}(\boldsymbol{X_s} = \boldsymbol{x_s}) = \begin{cases} w_{low}(l) & \text{if all } x_s^{(i)} = l \\ w_{high} & \text{otherwise} \end{cases}$$



Each colour represents a different superpixel

[1] Kohli et al. *CVPR* 2007. [2] Ladicky et al. *ICCV* 2009

Inference

- We have an energy that we want to minimise.
- Minimum energy is the highest probability
- $E(x) = \underbrace{\sum_i \psi_i^U(x_i)}_{\text{Unaries from CNN}} + \underbrace{\sum_{i<j} \psi_{i,j}^P(x_i, x_j)}_{\text{Pairwise [1]}} + \underbrace{\sum_d \psi_d^{Det}(\boldsymbol{x}_d)}_{\text{Detection potentials}} + \underbrace{\sum_s \psi_s^{SP}(\boldsymbol{x}_s)}_{\text{Superpixel potentials}}$
- During the forward pass of our network, we want to implement this energy minimisation / MAP estimation procedure

[1] Krahenbuhl and Koltun, NIPS, 2011

Mean Field Inference

- Our labelling is the most-likely assignment
  - $\arg\max_x \mathrm{P}(\boldsymbol{X} = x|\boldsymbol{I})$
- Approximate real distribution, P, with simpler one Q
  - $Q(\boldsymbol{X}) = \prod_i Q_i(X_i)$
- Minimise KL-Divergence between Q and P
  - $KL(Q||P) = -\sum_{x \in x} Q(x)\ \log(Q(x)) + Q(x)\ \log(P(x))$

Higher Order CRFs in Deep Neural Nets

- Mean-field: common inference algorithm for the MAP (maximum a posteriori) estimate of a CRF.
  - Iterative inference algorithm that is also differentiable

`Initialise`

$$Q_i = \frac{1}{Z_i} \exp(U_i(l))$$

`while not converged do`

$$Q^{t+1}(V_i = l) = \frac{1}{Z_i} \exp\left(- \sum_{c \in C} \sum_{\{\boldsymbol{v}_c \mid v_i = l\}} Q^t(\boldsymbol{v_{c-i}}) \psi(\boldsymbol{v}_c ; \theta) \right)$$

`end`

Higher Order CRFs in Deep Neural Nets

- Mean-field: Iterative inference algorithm that is also differentiable
  - We can formulate it as a recurrent neural network and optimise our network end-to-end

```
Initialise
```

$$Q_i = \frac{1}{Z_i} \exp(U_i(l))$$

```
while not converged do
```

$$Q^{t+1}(V_i = l) = \frac{1}{Z_i} \exp\left(- \sum_{c \in C} \sum_{\{\boldsymbol{v_c} \mid v_i = l\}} Q^t(\boldsymbol{v_{c-i}}) \psi(\boldsymbol{v_c}; \theta) \right)$$

```
end
```

$$Q^T = \text{update}\left(\ldots \text{ update}\left(\text{update}\left(\text{update}(Q^1) \right) \right) \ \ldots \right)$$

Higher Order CRFs in Deep Neural Nets

- Mean-field: Iterative inference algorithm that is also differentiable
  - We can formulate it as a recurrent neural network and optimise our network end-to-end

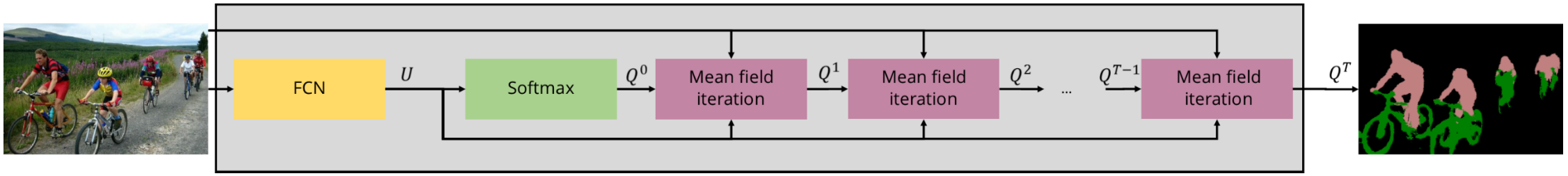
```
Initialise
```

$$Q_i = \frac{1}{Z_i}\exp(U_i(l))$$

```
for t = 1 … T
```

$$Q^{t+1}(V_i = l) = \frac{1}{Z_i}\exp\left(-\sum_{c\in C}\sum_{\{\boldsymbol{v_c}\mid v_i=l\}} Q^t(\boldsymbol{v_{c-i}})\psi(\boldsymbol{v_c};\theta)\right)$$

```
end
```



Results

- On PASCAL VOC 2012 reduced validation set

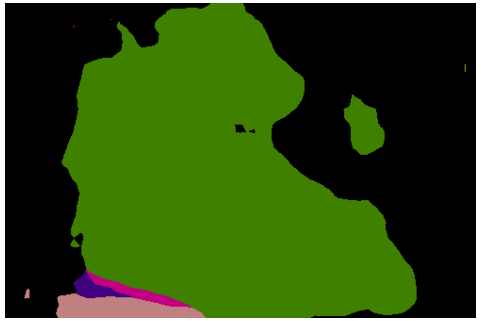
Method	Mean IoU [%]
Unary - FCN [1]	68.3
Pairwise [2]	72.9
Pairwise + Superpixels	74.0
Pairwise + Detections	74.9
Pairwise + Superpixels + Detections	75.8

[1] Long *et al.* CVPR 2015. [2] Zheng, Jayasumana *et al*. ICCV 2015

Results

FCN

$$E(\boldsymbol{x}) = \sum_i \psi_i^U(x_i)$$

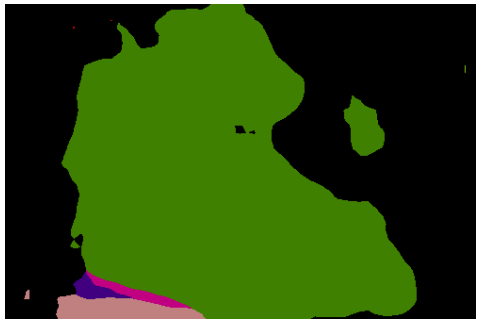


Results

Pairwise

$$E(\boldsymbol{x}) = \sum_i \psi_i^U(x_i) + \sum_{i<j} \psi_{i,j}^P(x_i, x_j)$$

FCN

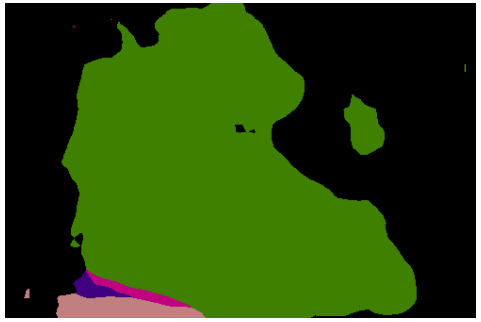


Results

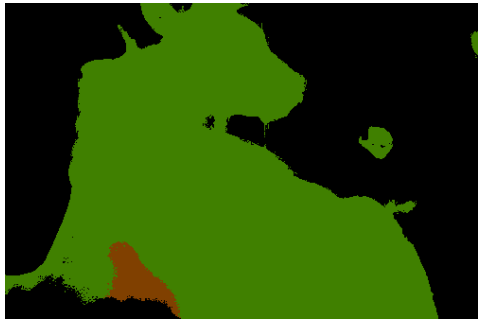
Superpixels

$$E(\boldsymbol{x}) = \sum_i \psi_i^U(x_i) + \sum_{i<j} \psi_{i,j}^P(x_i, x_j) + \sum_s \psi_s^{SP}(\boldsymbol{x}_s)$$

FCN



Pairwise

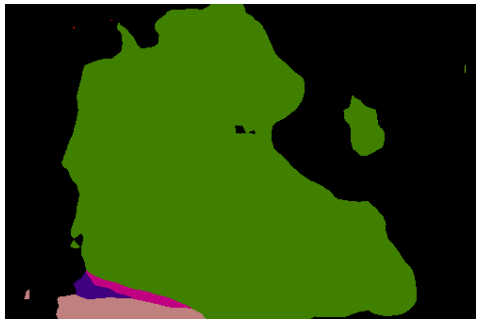


Results

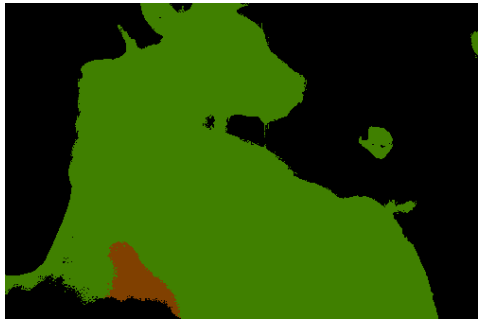
Detections

$$E(\boldsymbol{x}) = \sum_i \psi_i^U(x_i) + \sum_{i<j} \psi_{i,j}^P(x_i, x_j) + \sum_d \psi_d^{Det}(\boldsymbol{x}_d, y_d)$$

FCN



Pairwise

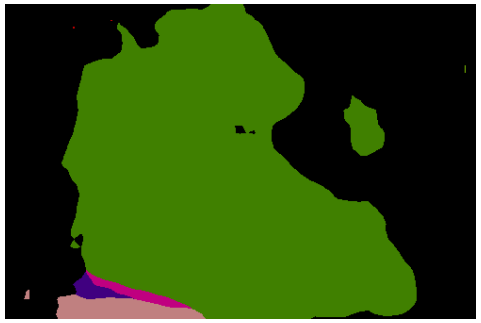


Superpixels

Results

$$E(\boldsymbol{x}) = \sum_i \psi_i^U(x_i) + \sum_{i<j} \psi_{i,j}^P(x_i, x_j) + \sum_s \psi_s^{SP}(\boldsymbol{x}_s) + \sum_d \psi_d^{Det}(\boldsymbol{x}_d, y_d)$$

FCN



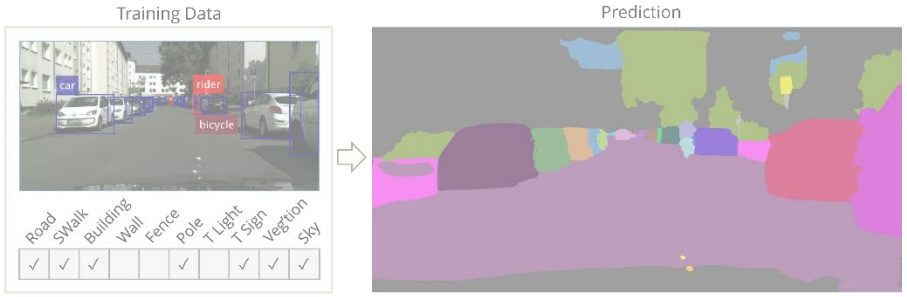
Pairwise

Superpixels

Detections

All

Outline

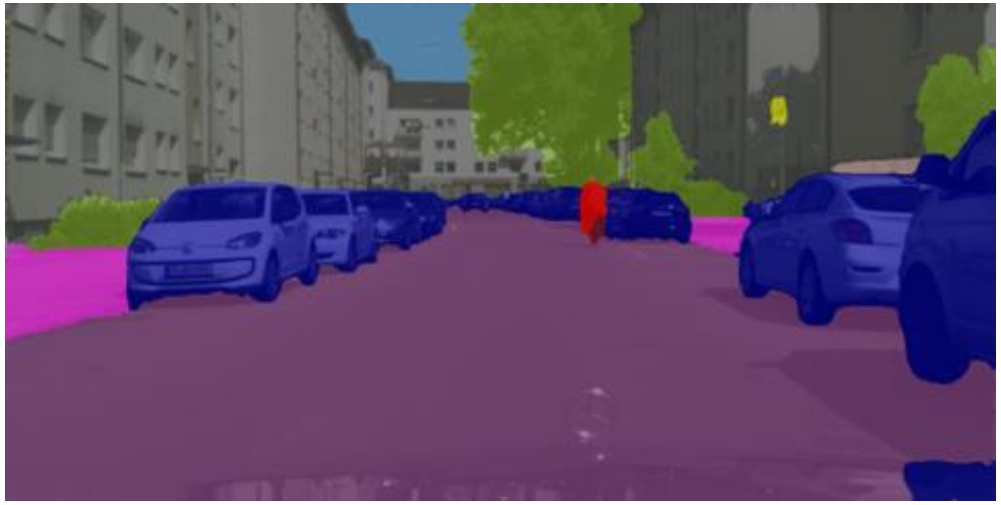


Higher Order CRFs in Deep Neural Networks (ECCV 2016)

Pixelwise Instance Segmentation with a Dynamically Instantiated Network (CVPR 2017)

Weakly and Semi-Supervised Panoptic Segmentation (ECCV 2018)

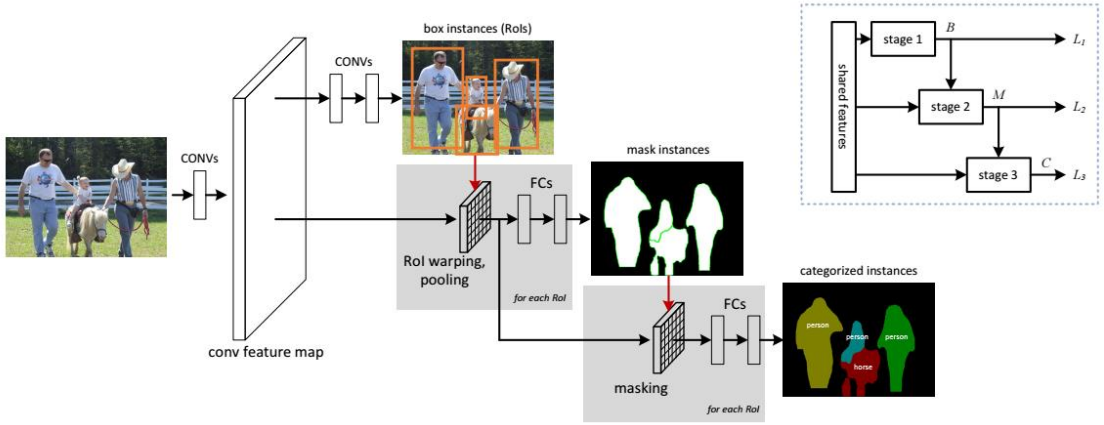
From Semantic to Instance Segmentation



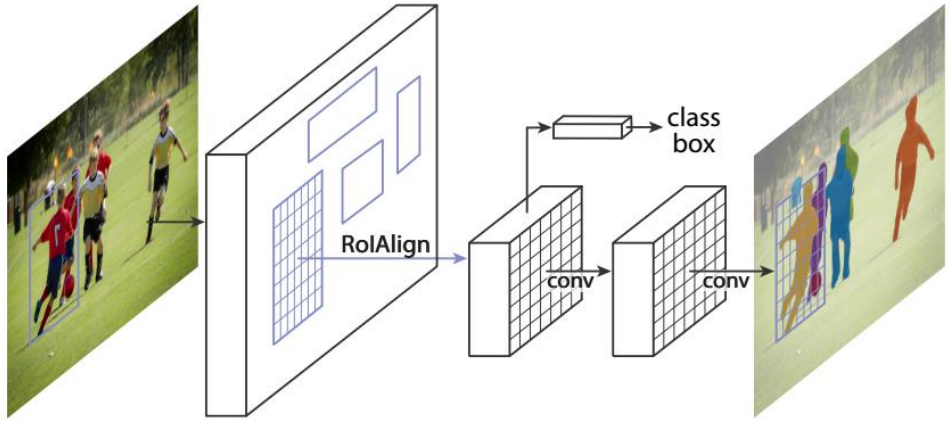
- Variable number of instances per image
- Instance ids do not have any semantic meaning. Permutation invariant
- No longer a straightforward mapping of images to labels.

Detection based approaches

- Most Instance Segmentation methods are based on Object Detection networks which are refined to produce segments instead of bounding boxes.



Multi-task Network Cascades (MNC). Dai *et al*. 2016



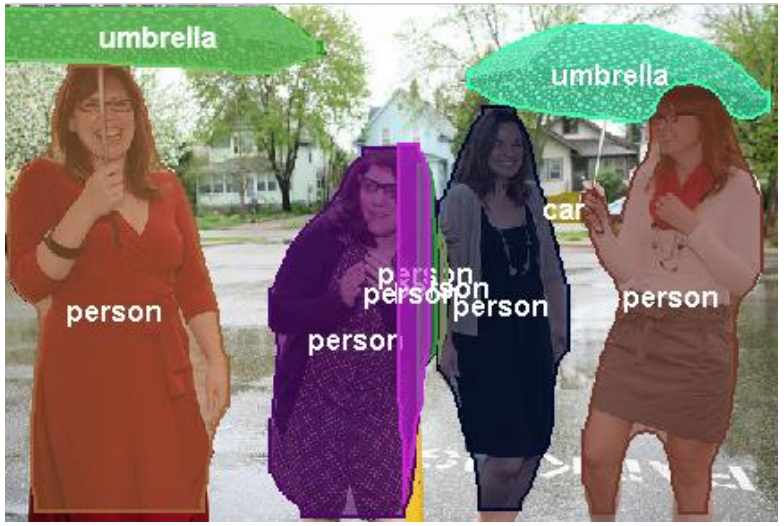
Mask R-CNN. He *et al*. 2017

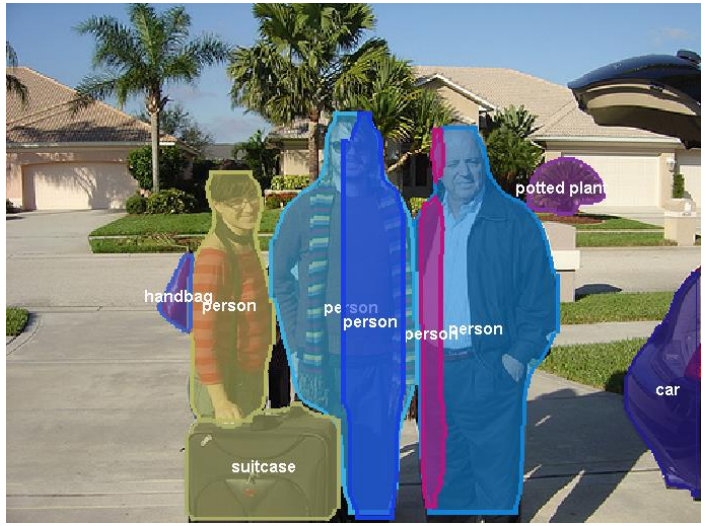
Hariharan *et al.* 2014, 2015. Chen *et al* 2015. Dai *et al*. 2015. Liu *et al*. 2016. Li *et al*. 2016. Qi *et al.* 2017

Detection based approaches

- Process independent object proposals
- More proposals than actual objects in the image
- Pixels can be ambiguous – can be assigned to multiple proposals
- Limited by quality of initial proposals
  - Cannot segment instances outside the detected box
- Detectors produce a *ranked list* of proposed instances and scores
  - Some methods cannot produce actual segmentation maps
  - Others require post-processing

Detection based approaches





Results of FCIS [1]. Winner of 2017 COCO challenge (around the time this paper was written). Full set here

[1] Qi *et al.* CVPR 2017.

Approach

- Based on a Semantic Segmentation network
- Outputs a variable number of instances depending on the image
- Reasons about the entire image holistically
  - Pixel can only belong to a single instance
  - Therefore has to reason about occlusions
- Produces segmentation maps directly
  - Requires no post-processing
  - Segmentations are more precise

Sounds like Panoptic Segmentation

- It is, this work was before the name was coined
- Panoptic segmentation
  - No overlapping instances
  - Must handle “stuff” classes

Panoptic Segmentation

Alexander Kirillov[1,2] Kaiming He[1] Ross Girshick[1] Carsten Rother[2] Piotr Dollár[1]
[1]Facebook AI Research (FAIR) [2]HCI/IWR, Heidelberg University, Germany

Abstract

We propose and study a novel panoptic segmentation *(PS) task. Panoptic segmentation unifies the typically distinct tasks of* semantic segmentation *(assign a class label to each pixel) and* instance segmentation *(detect and segment each object instance). The proposed task requires generating a* coherent *scene segmentation that is rich and complete, an important step toward real-world vision systems. While early work in computer vision addressed related image/scene parsing tasks, these are not currently popular, possibly due to lack of appropriate metrics or associated recognition challenges. To address this, we first propose a novel* panoptic quality *(PQ) metric that captures performance for all classes (stuff and things) in an interpretable and unified manner. Using the proposed metric, we perform a rigorous study of both human and machine performance for PS on three existing datasets, revealing interesting insights about the task. Second, we are working to introduce panoptic segmentation tracks at upcoming recognition challenges. The aim of our work is to revive the interest of the community in a more unified view of image segmentation.*

1. Introduction

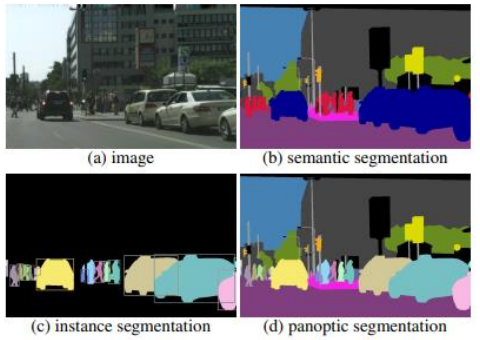
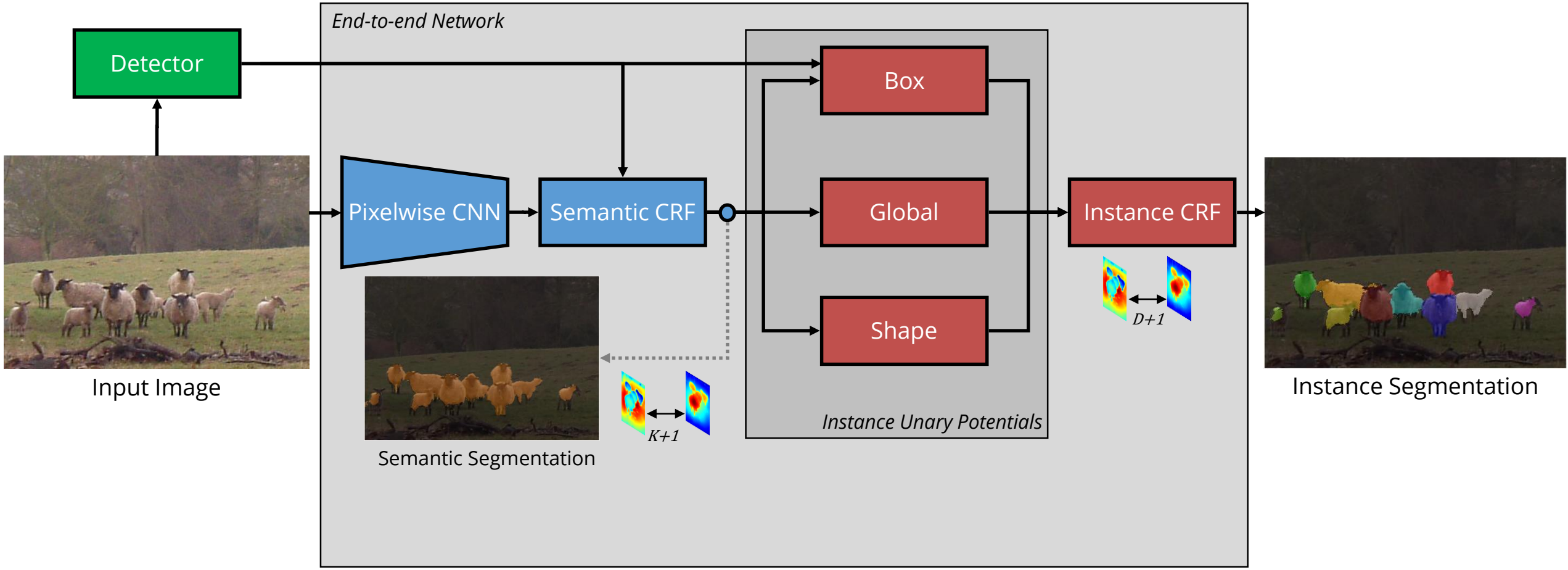


Figure 1: For a given (a) image, we show *ground truth* for: (b) semantic segmentation (per-pixel class labels), (c) instance segmentation (per-object mask and class label), and (d) the proposed *panoptic segmentation* task (per-pixel class+instance labels). The PS task: (1) encompasses both stuff and thing classes, (2) uses a simple but general format, and (3) introduces a uniform evaluation metric for all classes. Panoptic segmentation generalizes both semantic and instance segmentation and we expect the unified task will present novel challenges and enable innovative new methods.

Overview


End-to-end Network
Detector
Box
Pixelwise CNN
Semantic CRF
Global
Instance CRF
Shape
Instance Unary Potentials
D+1
K+1
Input Image
Semantic Segmentation
Instance Segmentation

Approach

- Based on initial semantic segmentation
- Conditional Random Field (CRF) defined over instances.
- Associates each pixel in semantic segmentation with a detection to get instance segmentation.
- Three different unary potentials consider:
  - Object detections for the image
    - Make assumption that $\#Object\ Detections = \#Possible\ Instances$
  - Initial semantic segmentation
  - Object shapes
- Two conceptually different models – but trained end-to-end.

Approach

$$E(\mathbf{V} = \mathbf{v}) = \sum_i U(v_i) + \sum_{i<j} P(v_i, v_j).$$

$$U(v_i) = -\ln[w_1 \psi_{Box}(v_i) + w_2 \psi_{Global}(v_i) + w_3 \psi_{Shape}(v_i)].$$

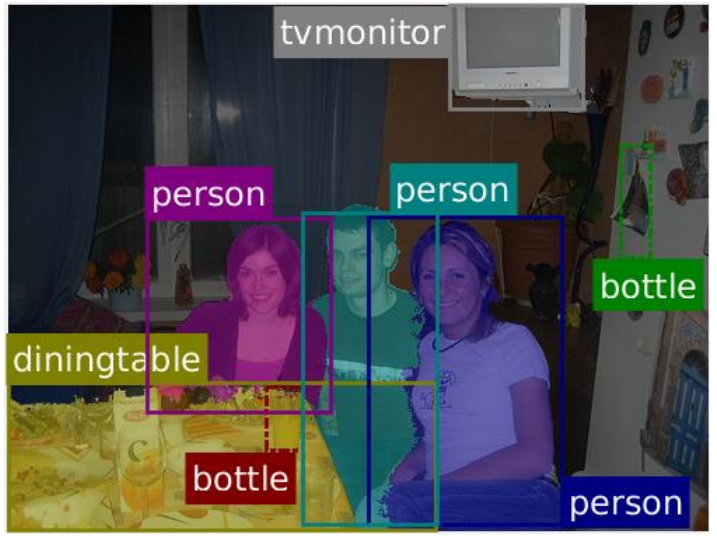
Pairwise terms are the same Gaussian kernels as DenseCRF [1]

$V_i \in \{0, 1, \ldots, D\}$ where D is the number of detections and changes for every image.

[1] Krahenbuhl and Koltun. NIPS 2011.

Box Term

- Based on intuition that if segmented pixel lies within bounding box and has the same class, then it belongs to the instance represented by that class
- Works well when we have good initial semantic segmentation

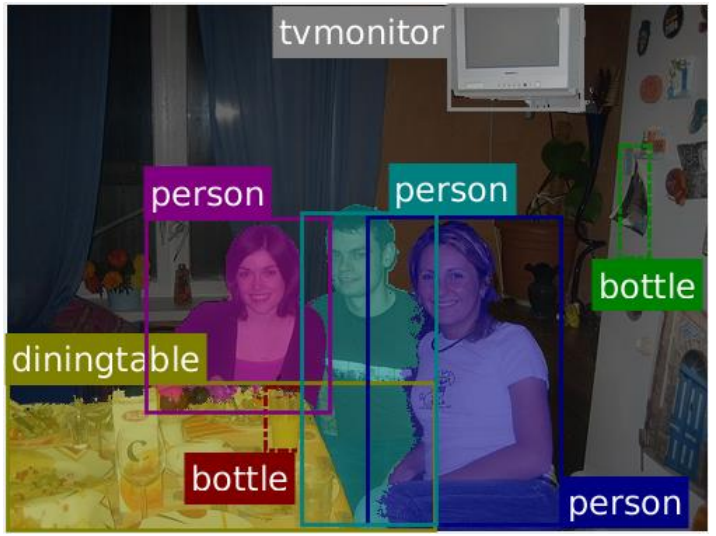


Box Term

$$\psi_{Box}(V_i = k) = \begin{cases} Q_i(l_k)s_k & \text{if } i \in B_k \\ 0 & \text{otherwise} \end{cases}$$

$Q_i(l_k)$ - probability of i^{th} pixel taking on label of k^{th} detection.

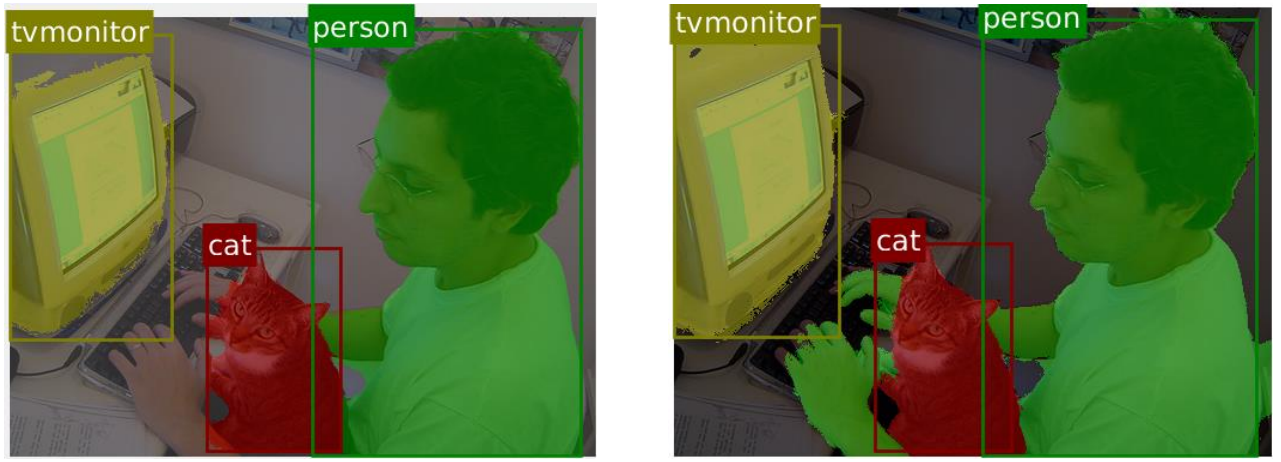
s_k - score of k^{th} detection.



Global Term

- If we have d possible instances of an object class, and no further localisation information, each instances is equally probable.
- Global term can handle poorly localised bounding boxes

$$\psi_{Global}(V_i = k) = Q_i(l_k).$$



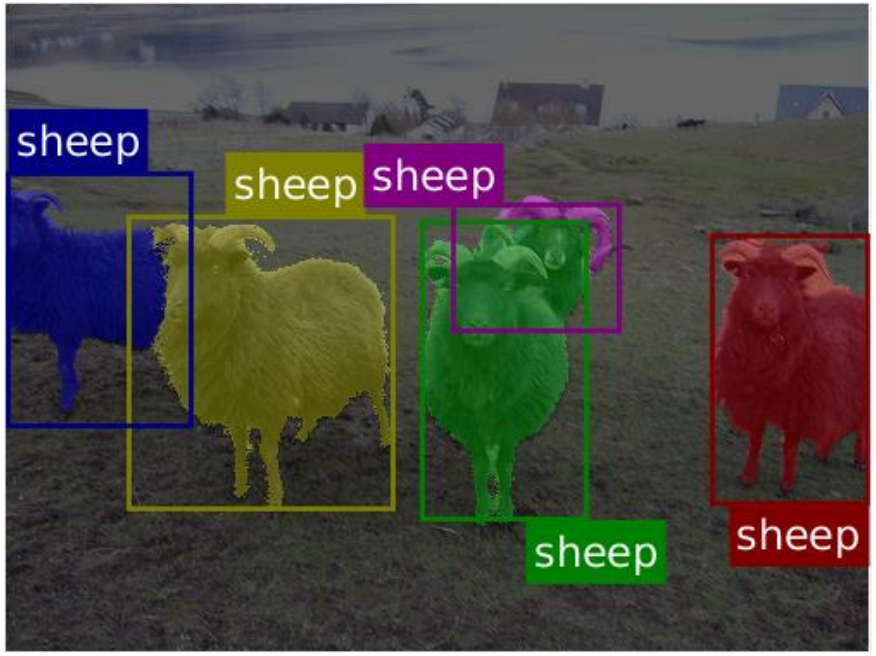
Global Term

- Gradients depend on entire semantic segmentation map, and are hence more stable
- Semantic Segmentation performance also improves with end-to-end training

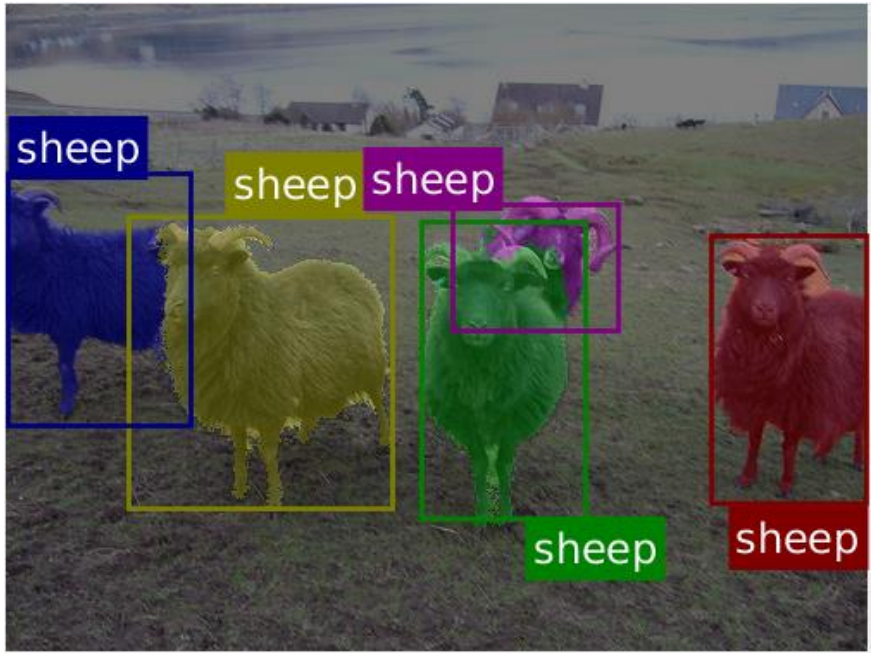
$$\psi_{Global}(V_i = k) = Q_i(l_k).$$

Shape Term

- Intuitively, shape priors help us reason about occluded objects which look the same



Without shape priors



With shape priors

Shape Term

- Match shape templates to objects
- Shape templates are weights within the network
- Update shape templates through backpropogation
- Layer is effectively a convolution and max over channels
- Initialised with old shape prior methods

$$t^* = \arg\max_{t \in \tilde{\mathcal{T}}} \frac{\sum \mathbf{Q}_{B_k}(l_k) \odot t}{\left\| \mathbf{Q}_{B_k}(l_k) \right\| \|t\|}$$

$$\psi(\mathbf{V}_{B_k} = k) = \mathbf{Q}_{B_k}(l_k) \odot t^*.$$

Inference of CRF

- Our predicted instance segmentation is the MAP of the CRF.
- Minimise the CRF energy

$$\Pr(\mathbf{V} = \mathbf{v}) = \frac{1}{Z} \exp\left(-E(\mathbf{v})\right)$$

$$E(\mathbf{v}) = \sum_i U(v_i) + \sum_{i<j} P(v_i, v_j).$$

- Use approximate mean-field inference which can be unrolled as part of the neural network.
- Enables end-to-end training

Extension

- Incorporate detector into network itself and train jointly.
- UPSNet a good example
  - Similar “Box” and “Shape” term in “Panoptic Head”

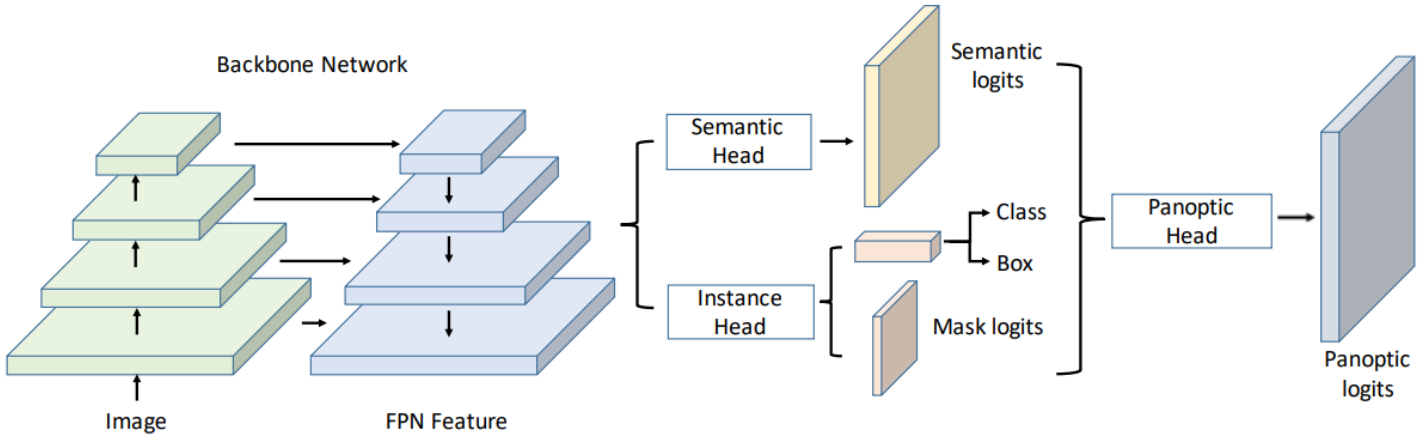


Figure 1: Overall architecture of our UPSNet.

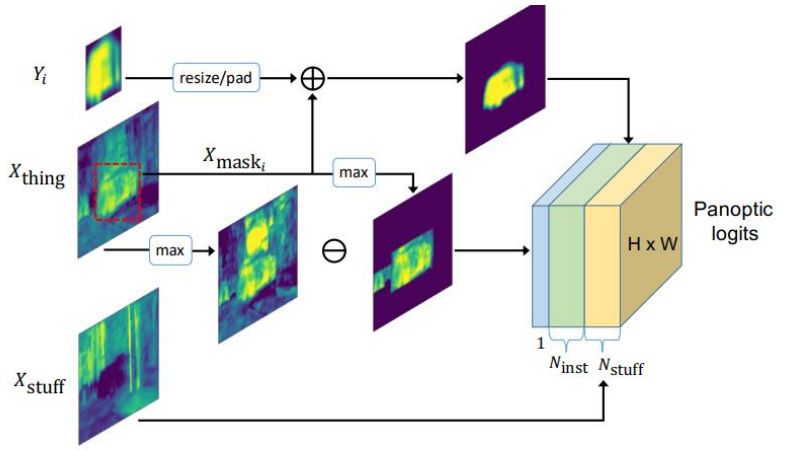
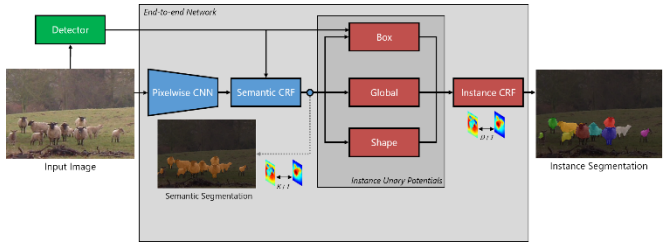


Figure 3: Architecture of our panoptic segmentation head.

[1] Xiong *et al.* CVPR 2019.

Loss Function



(a) Original ground truth, $\mathcal{G}$ (b) Prediction, $\mathcal{P}$ (c) "Matched" ground truth, $\mathcal{G}^*$

- "Match" prediction to ground truth, since permutations of an instance labelling are actually the same result
- Once matched, use any loss function. Log likelihood worked the best.

Results - VOC

Table 1. The effect of the different CRF unary potentials, and end-to-end training with them, on the VOC 2012 Validation Set.

	AP^r 0.5	AP^r 0.7	AP^r 0.9	AP^r_{vol}	match IoU
Box Term (piecewise)	60.0	47.3	21.2	54.9	42.6
Box+Global (piecewise)	59.1	46.1	23.4	54.6	43.0
Box+Global+Shape (piecewise)	59.5	46.4	23.3	55.2	44.8
Box Term (end-to-end)	60.7	47.4	24.6	56.2	46.9
Box+Global (end-to-end)	60.9	48.1	**25.5**	56.7	47.1
Box+Global+Shape (end-to-end)	**61.7**	**48.6**	25.1	**57.5**	**48.3**

Table 2. Comparison of Instance Segmentation performance to recent methods on the VOC 2012 Validation Set

Method	AP^r 0.5	0.6	0.7	0.8	0.9	AP^r_{vol}
SDS [16]	43.8	34.5	21.3	8.7	0.9	–
Chen *et al.* [7]	46.3	38.2	27.0	13.5	2.6	–
PFN [26]	58.7	51.3	42.5	31.2	15.7	52.3
Arnab *et al.* [3]	58.3	52.4	45.4	34.9	20.1	53.1
MPA 1-scale [31]	60.3	54.6	45.9	34.3	17.3	54.5
MPA 3-scale [31]	**62.1**	**56.6**	47.4	36.1	18.5	56.5
Ours	61.7	55.5	**48.6**	**39.5**	**25.1**	**57.5**

Results - VOC

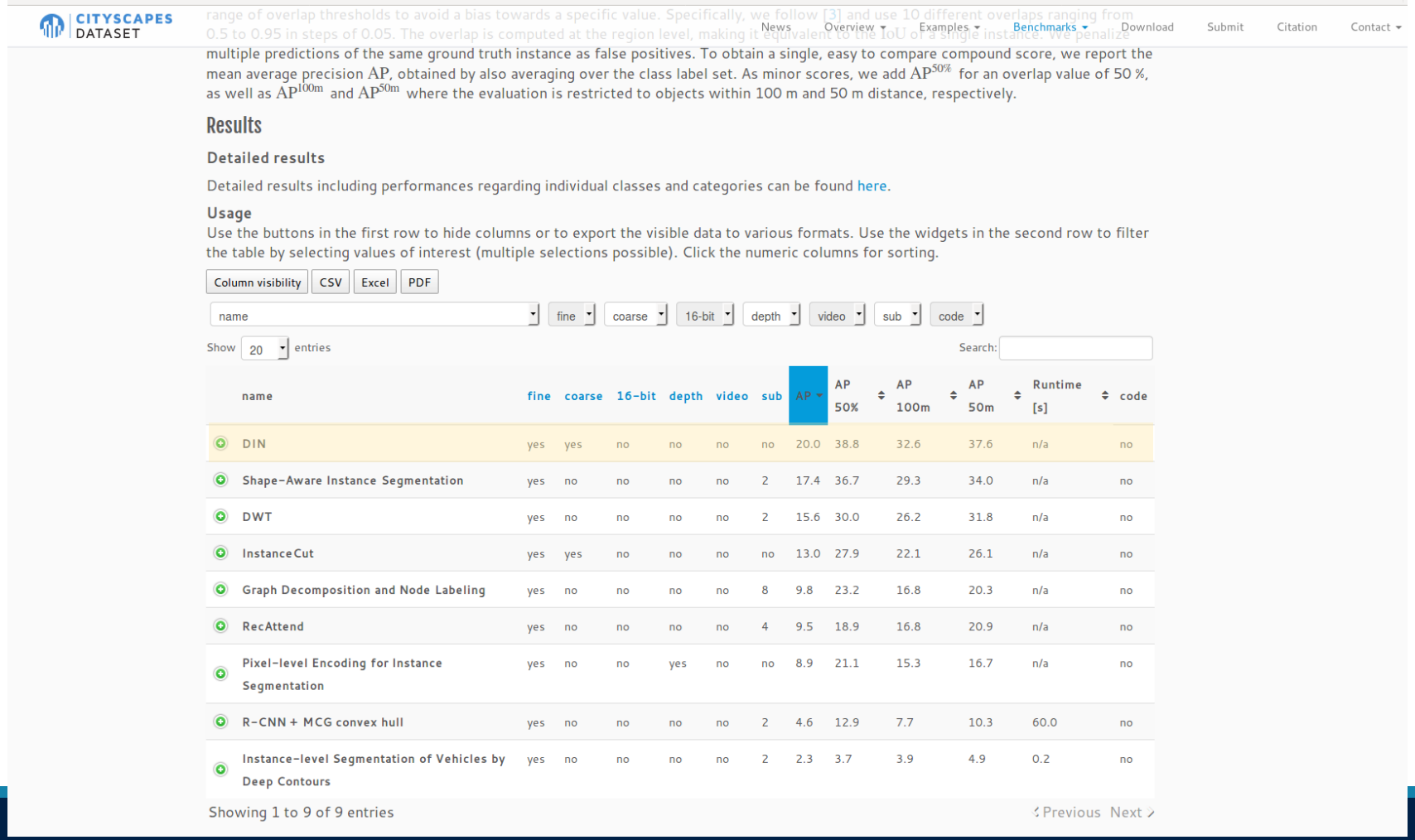
Table 1. The effect of the different CRF unary potentials, and end-to-end training with them, on the VOC 2012 Validation Set.

	AP^r 0.5	0.7	0.9	AP^r_{vol}	match IoU
Box Term (piecewise)	60.0	47.3	21.2	54.9	42.6
Box+Global (piecewise)	59.1	46.1	23.4	54.6	43.0
Box+Global+Shape (piecewise)	59.5	46.4	23.3	55.2	44.8
Box Term (end-to-end)	60.7	47.4	24.6	56.2	46.9
Box+Global (end-to-end)	60.9	48.1	**25.5**	56.7	47.1
Box+Global+Shape (end-to-end)	**61.7**	**48.6**	25.1	**57.5**	**48.3**

Table 2. Comparison of Instance Segmentation performance to recent methods on the VOC 2012 Validation Set

Method	AP^r 0.5	0.6	0.7	0.8	0.9	AP^r_{vol}
SDS [16]	43.8	34.5	21.3	8.7	0.9	–
Chen *et al.* [7]	46.3	38.2	27.0	13.5	2.6	–
PFN [26]	58.7	51.3	42.5	31.2	15.7	52.3
Arnab *et al.* [3]	58.3	52.4	45.4	34.9	20.1	53.1
MPA 1-scale [31]	60.3	54.6	45.9	34.3	17.3	54.5
MPA 3-scale [31]	**62.1**	**56.6**	47.4	36.1	18.5	56.5
Ours	61.7	55.5	**48.6**	**39.5**	**25.1**	**57.5**

Results - Cityscapes



CITYSCAPES DATASET

News Overview Examples Benchmarks Download Submit Citation Contact

range of overlap thresholds to avoid a bias towards a specific value. Specifically, we follow [3] and use 10 different overlaps ranging from 0.5 to 0.95 in steps of 0.05. The overlap is computed at the region level, making it equivalent to the IoU of a single instance. We penalize multiple predictions of the same ground truth instance as false positives. To obtain a single, easy to compare compound score, we report the mean average precision AP, obtained by also averaging over the class label set. As minor scores, we add $AP^{50\%}$ for an overlap value of 50 %, as well as AP^{100m} and AP^{50m} where the evaluation is restricted to objects within 100 m and 50 m distance, respectively.

Results

Detailed results

Detailed results including performances regarding individual classes and categories can be found here.

Usage

Use the buttons in the first row to hide columns or to export the visible data to various formats. Use the widgets in the second row to filter the table by selecting values of interest (multiple selections possible). Click the numeric columns for sorting.

Column visibility | CSV | Excel | PDF

name | fine | coarse | 16-bit | depth | video | sub | code

Show 20 entries Search:

name	fine	coarse	16-bit	depth	video	sub	AP	AP 50%	AP 100m	AP 50m	Runtime [s]	code
DIN	yes	yes	no	no	no	no	20.0	38.8	32.6	37.6	n/a	no
Shape-Aware Instance Segmentation	yes	no	no	no	no	2	17.4	36.7	29.3	34.0	n/a	no
DWT	yes	no	no	no	no	2	15.6	30.0	26.2	31.8	n/a	no
InstanceCut	yes	yes	no	no	no	no	13.0	27.9	22.1	26.1	n/a	no
Graph Decomposition and Node Labeling	yes	no	no	no	no	8	9.8	23.2	16.8	20.3	n/a	no
RecAttend	yes	no	no	no	no	4	9.5	18.9	16.8	20.9	n/a	no
Pixel-level Encoding for Instance Segmentation	yes	no	no	yes	no	no	8.9	21.1	15.3	16.7	n/a	no
R-CNN + MCG convex hull	yes	no	no	no	no	2	4.6	12.9	7.7	10.3	60.0	no
Instance-level Segmentation of Vehicles by Deep Contours	yes	no	no	no	no	2	2.3	3.7	3.9	4.9	0.2	no

Showing 1 to 9 of 9 entries Previous Next

Results - Cityscapes

Method	Validation AP^r_{vol} th.	Validation AP^r_{vol} st.	Validation AP^r_{vol} all	Validation PQ th.	Validation PQ st.	Validation PQ all	Validation IoU th.	Validation IoU st.	Validation IoU all	Test AP^r_{vol} th.
Ours (full, PSPNet init.)	28.6	52.6	42.5	42.5	62.1	53.8	80.1	79.5	79.8	23.4
Pixel Encoding [68]	9.9	–	–	–	–	–	–	–	–	8.9
RecAttend [69]	–	–	–	–	–	–	–	–	–	9.5
InstanceCut [30]	–	–	–	–	–	–	–	–	–	13.0
DWT [28]	21.2	–	–	–	–	–	–	–	–	19.4
SGN [31]	29.2	–	–	–	–	–	–	–	–	25.0

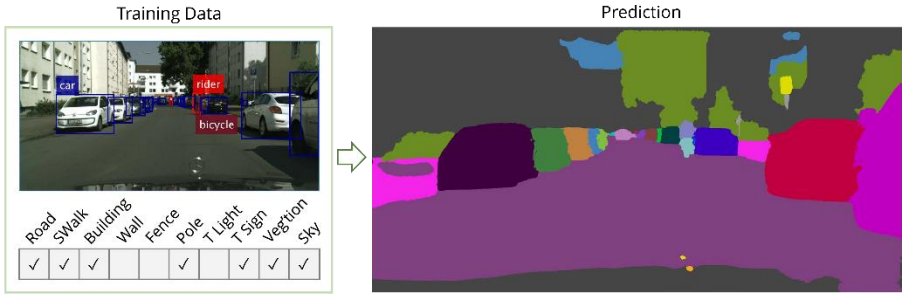
Comparison to FCIS

Ours (VOC)

FCIS [1] (COCO)

[1] Qi *et al.* CVPR 2017.

Outline



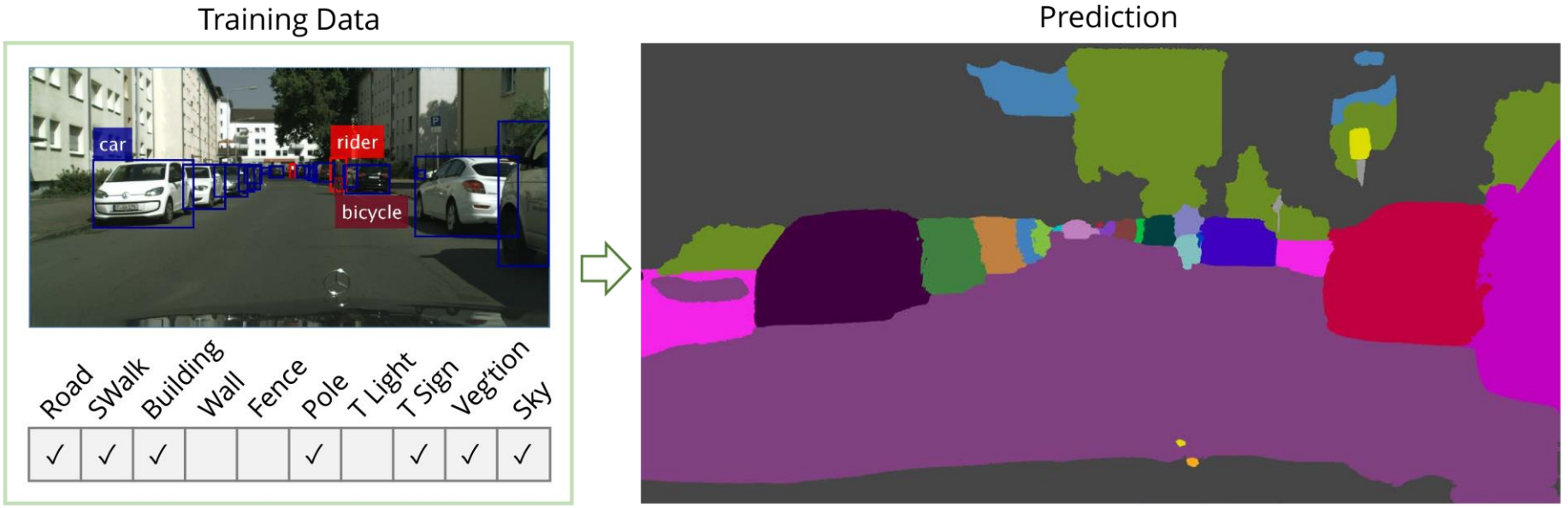
Higher Order CRFs in Deep Neural Networks (ECCV 2016)

Pixelwise Instance Segmentation with a Dynamically Instantiated Network (CVPR 2017)

Weakly and Semi-Supervised Panoptic Segmentation (ECCV 2018)

Weaker Supervision

- One Cityscapes image takes 90 minutes to annotate.
- Use bounding-box annotations for “things” and image-level tags for “stuff”
- 35x less annotation time according to [1] and [2] on Cityscapes.



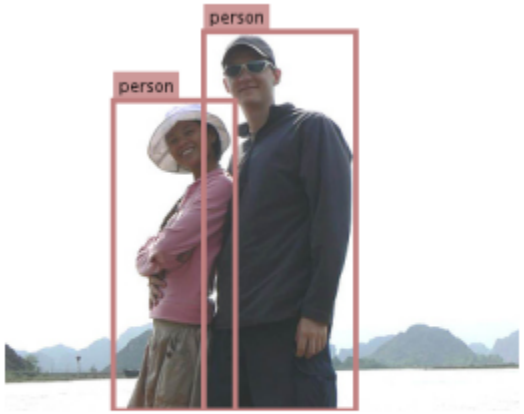
[1] Papadopoulos *et al.* ICCV 2017. [2] Papadopoulos *et al.* ECCV 2014

Approach

- Method similar to Expectation Maximisation (EM)
- Alternate between
  - Approximate the unknown, pixel-level ground truth with the trained model.
  - Training the model with the approximated ground truth.
- Initialisation of the approximate ground truth is critical.

Initialisation from Bounding Boxes

- Unsupervised foreground-background segmentation algorithms provide good priors.
- Perform both GrabCut [1] and MCG [2].
- Estimated foreground is assigned detected class label, but only if the two methods agree. Otherwise, the pixel is marked as “ignore”



(a) Input image

(b) Semantic segmentation approximate ground truth

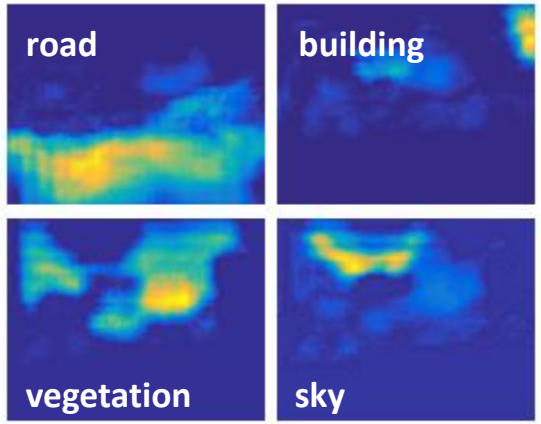
(c) Instance segmentation approximate ground truth

[1] Rother *et al.* ACM TOG 2004. [2] Arbelaez *et al.* CVPR 2014. [3] Khoreva *et al.* CVPR 2017

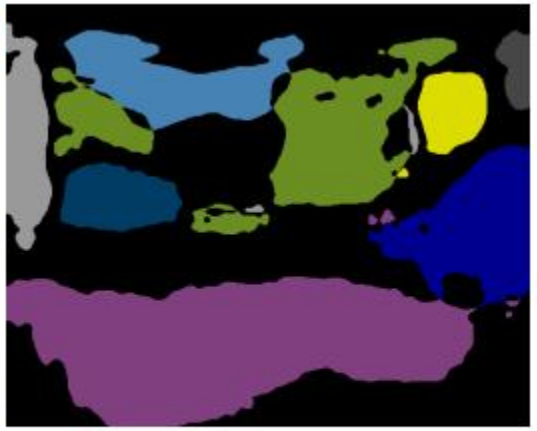
Initialisation from Image-level Tags

- First train network for multilabel binary classification of tags in image.
- Then perform Grad-CAM [1] to obtain pixels which contribute most to the classification.
- These masks are thresholded to obtain the approximate ground truth.

Input image



Localisation heatmaps for road, building, vegetation and sky



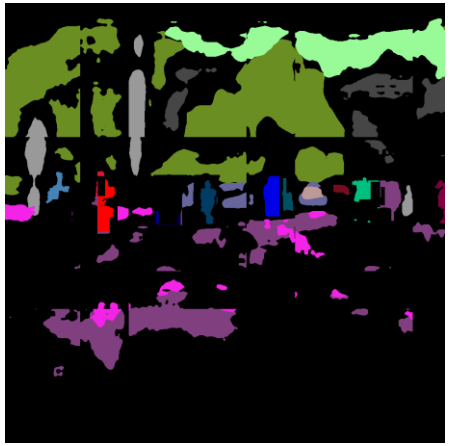
Approximate ground truth generated from image tags

[1] Selvaraju *et al.* ICCV 2017.

Iterative training

- Cues from bounding boxes and image-level tags are combined to produce initial approximate ground truth.
- Thereafter, network's own predictions are used as the approximate ground truth.
- Predictions below a confidence threshold, and not within a bounding box are marked as "ignore".

Input image



Iteration 0

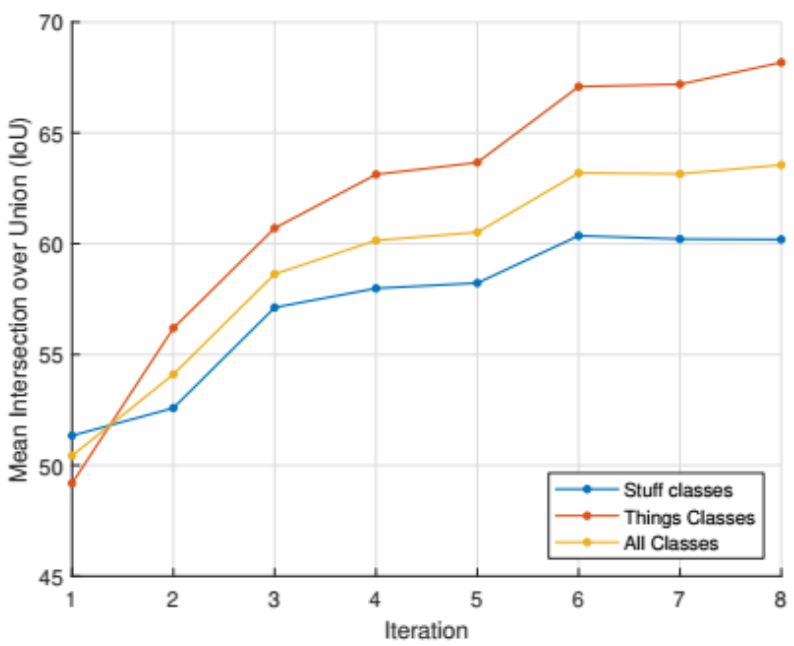
Iteration 2

Iteration 5

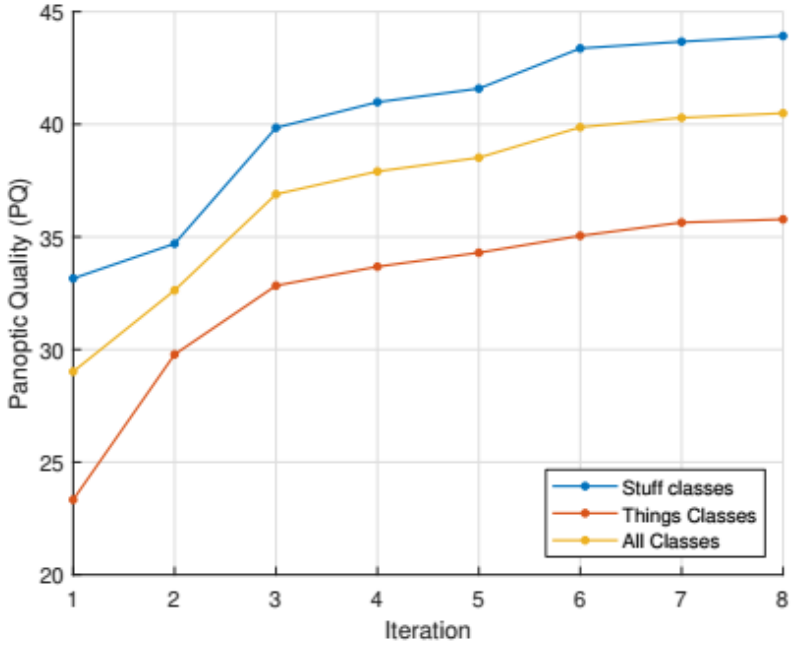
Ground truth

Iterative training

- Steady improvement in network performance with more iterations of ground truth generation.



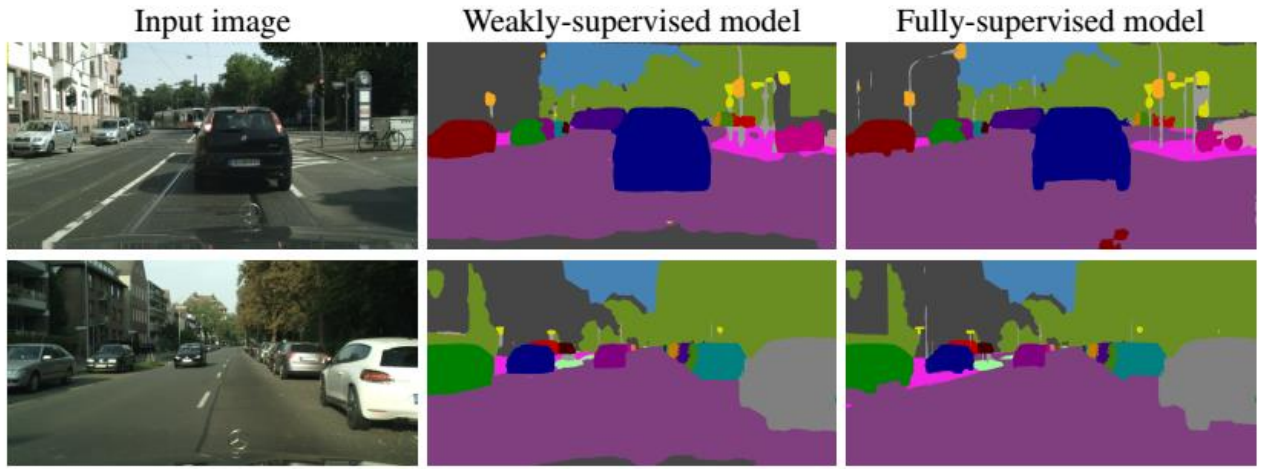
(a) Semantic segmentation (IoU)



(b) Instance segmentation (PQ)

Results - Cityscapes

Method	Validation AP^r_{vol} th.	AP^r_{vol} st.	AP^r_{vol} all	PQ th.	PQ st.	PQ all	IoU th.	IoU st.	IoU all	Test AP^r_{vol} th.
Ours (weak, ImageNet init.)	17.0	33.1	26.3	35.8	43.9	40.5	68.2	60.2	63.6	12.8
Ours (full, ImageNet init.)	24.3	42.6	34.9	39.6	52.9	47.3	70.4	72.4	71.6	18.8



Results – Pascal VOC

Method	AP^r 0.5	0.6	0.7	0.8	0.9	AP^r_{vol}	PQ
Weakly supervised without COCO							
SDI [43]	44.8	–	–	–	–	–	–
Ours	**60.5**	**55.2**	**47.8**	**37.6**	**21.6**	**55.6**	**59.0**
Fully supervised without COCO							
SDS [38]	43.8	34.5	21.3	8.7	0.9	–	–
Chen *et al.* [64]	46.3	38.2	27.0	13.5	2.6	–	–
PFN [65]	58.7	51.3	42.5	31.2	15.7	52.3	–
Ours (fully supervised)	**63.6**	**59.5**	**53.8**	**44.7**	**30.2**	**59.2**	**62.7**
Weakly supervised with COCO							
SDI [43]	46.4	–	–	–	–	–	–
Ours	**60.9**	**55.9**	**48.0**	**37.2**	**21.7**	**55.5**	**59.5**
Fully supervised with COCO							
Arnab *et al.* [17]	58.3	52.4	45.4	34.9	20.1	53.1	–
MPA [27]	62.1	56.6	47.4	36.1	18.5	56.5	–
SGN [31]	61.4	55.9	49.9	42.1	26.9	–	–
Ours (fully supervised)	**63.9**	**59.3**	**54.3**	**45.4**	**30.2**	**59.5**	**63.1**

Results – Pascal VOC

Method	AP^r 0.5	0.6	0.7	0.8	0.9	AP^r_{vol}	PQ
Weakly supervised without COCO							
SDI [43]	44.8	–	–	–	–	–	–
Ours	**60.5**	**55.2**	**47.8**	**37.6**	**21.6**	**55.6**	**59.0**
Fully supervised without COCO							
SDS [38]	43.8	34.5	21.3	8.7	0.9	–	–
Chen *et al.* [64]	46.3	38.2	27.0	13.5	2.6	–	–
PFN [65]	58.7	51.3	42.5	31.2	15.7	52.3	–
Ours (fully supervised)	**63.6**	**59.5**	**53.8**	**44.7**	**30.2**	**59.2**	**62.7**
Weakly supervised with COCO							
SDI [43]	46.4	–	–	–	–	–	–
Ours	**60.9**	**55.9**	**48.0**	**37.2**	**21.7**	**55.5**	**59.5**
Fully supervised with COCO							
Arnab *et al.* [17]	58.3	52.4	45.4	34.9	20.1	53.1	–
MPA [27]	62.1	56.6	47.4	36.1	18.5	56.5	–
SGN [31]	61.4	55.9	49.9	42.1	26.9	–	–
Ours (fully supervised)	**63.9**	**59.3**	**54.3**	**45.4**	**30.2**	**59.5**	**63.1**

Results – Semi-Supervised

- Common to mix Pascal VOC (10 000 images) and COCO (60 000 images) in training set.
- Using fully- or weakly-supervised COCO data does not make much difference.
- COCO data is of much lower quality than VOC.
- Conclusion: Rather label a few images well (Pascal VOC) than many images poorly (COCO)

Dataset		IoU	AP^r_{vol}	PQ
VOC	COCO			
Weak	Weak	75.7	55.5	59.5
Weak	Full	75.8	56.1	59.8
Full	Weak	77.5	58.9	62.7
Full	Full	79.0	59.5	63.1

COCO annotation

VOC annotation

Results – Semi-Supervised

- Common to mix Pascal VOC (10 000 images) and COCO (60 000 images) in training set.
- Using fully- or weakly-supervised COCO data does not make much difference.
- COCO data is of much lower quality than VOC.
- Conclusion: Rather label a few images well (Pascal VOC) than many images poorly (COCO)

Dataset		IoU	AP^r_{vol}	PQ
VOC	COCO			
Weak	Weak	75.7	55.5	59.5
Weak	Full	75.8	56.1	59.8
Full	Weak	77.5	58.9	62.7
Full	Full	79.0	59.5	63.1

COCO annotation

VOC annotation

Results – Semi-Supervised

- Common to mix Pascal VOC (10 000 images) and COCO (60 000 images) in training set.
- Using fully- or weakly-supervised COCO data does not make much difference.
- COCO data is of much lower quality than VOC.
- Conclusion: Rather label a few images well (Pascal VOC) than many images poorly (COCO)

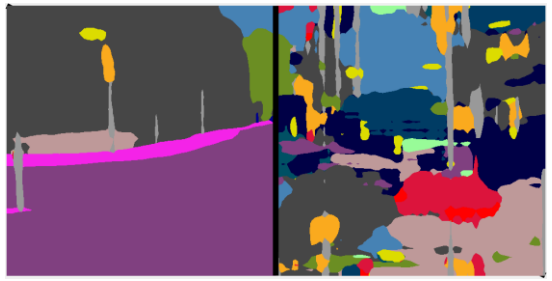
Dataset		IoU	AP^r_{vol}	PQ
VOC	COCO			
Weak	Weak	75.7	55.5	59.5
Weak	Full	75.8	56.1	59.8
Full	Weak	77.5	58.9	62.7
Full	Full	79.0	59.5	63.1

COCO annotation

VOC annotation

Outline

Exploiting Temporal Context for 3D Pose Estimation in the Wild (CVPR 2019) – DeepMind internship



On the Robustness of Semantic Segmentation Models to Adversarial Attacks (CVPR 2018)

Introduction

- Monocular 3D human pose estimation is an inherently ill-posed problem
- Metric ground-truth for real-world data is prohibitively difficult to collect
- Common datasets are from motion capture in controlled labs
- Models trained on these datasets generalise poorly to “in the wild”

Human 3.6M (mocap dataset)

“In-the-wild” data

Generalising to the real-world

[1] Kanazawa *et al.* End-to-end recovery of human shape and pose. *CVPR* 2018.

Temporal Consistency

- Temporal dimension of ordinary video encodes valuable information
- Multiple views of person observed
  - Body shape and bone lengths remain constant
  - Joint positions in both 2D and 3D vary slowly

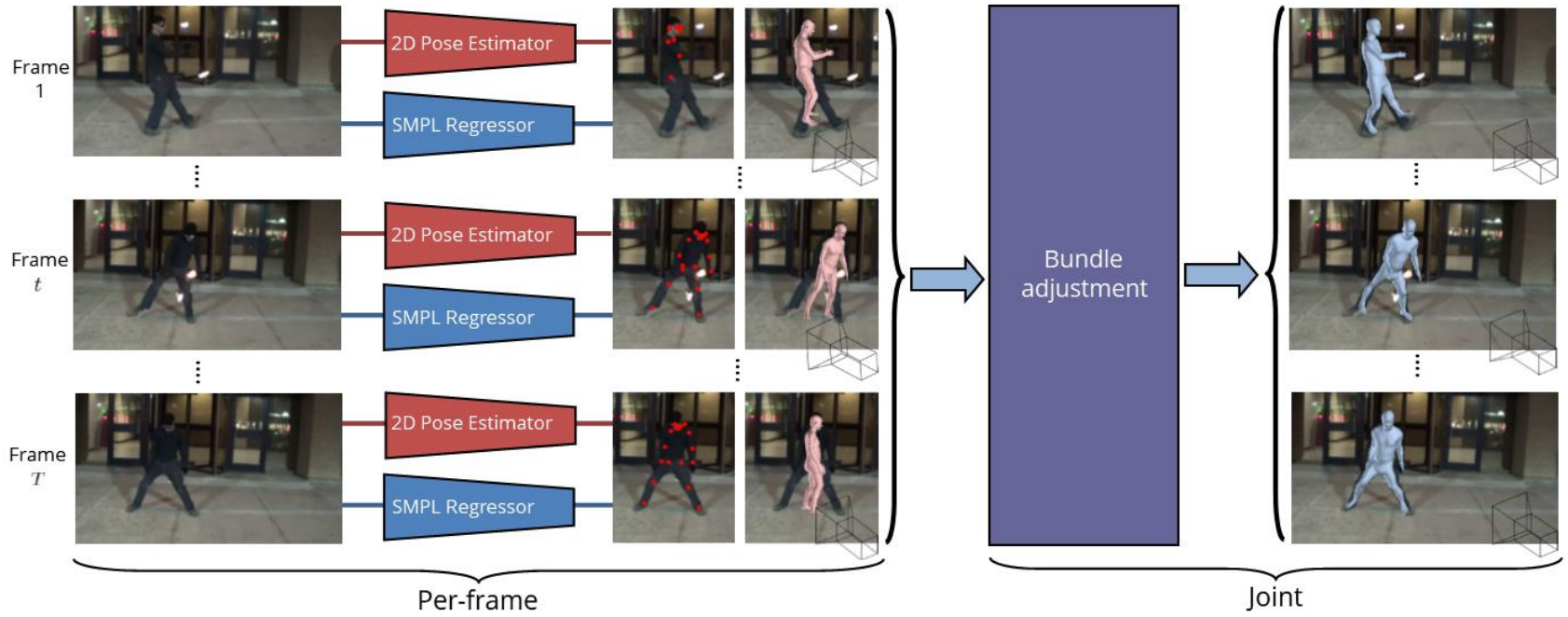
Our approach

- Propose a form of bundle adjustment
  - Take into account temporal information and multi-view geometry of the video
- Apply our method to about 107 000 YouTube videos in the Kinetics dataset.
- Automatically create a new “in-the-wild” dataset from this
- Improve performance of per-frame model using this dataset.

Bundle Adjustment

- Objective function consists of reprojection, temporal and prior terms

$$E(\beta, \theta, \mathbf{\Omega}) = E_R(\beta, \theta, \mathbf{\Omega}) + E_T(\beta, \theta, \mathbf{\Omega}) + E_P(\theta, \beta)$$

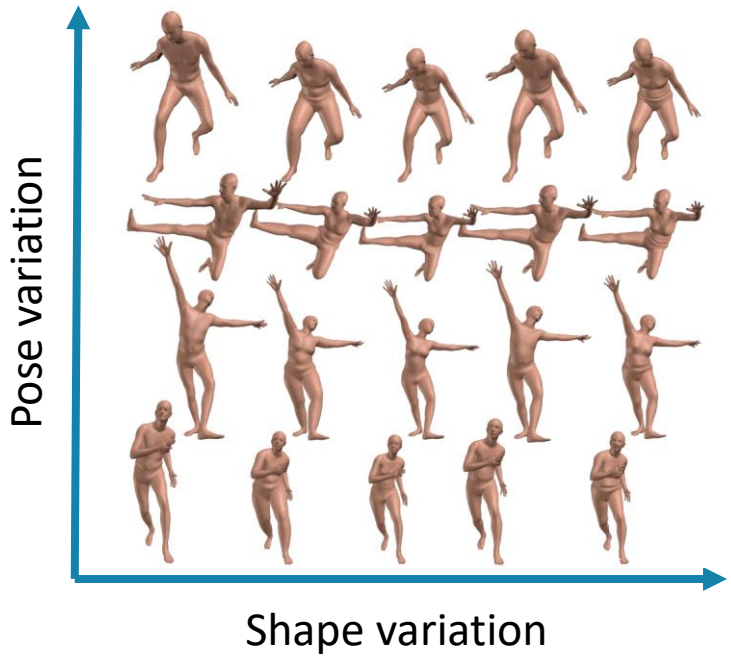


Bundle Adjustment

- Objective function consists of reprojection, temporal and prior terms

$$E(\beta, \theta, \mathbf{\Omega}) = E_R(\beta, \theta, \mathbf{\Omega}) + E_T(\beta, \theta, \mathbf{\Omega}) + E_P(\theta, \beta)$$

- Use the SMPL body model [1] to parameterise the 3D pose
  - Pose parameters: $\theta^t \in \mathbb{R}^{23\times 3}$
  - Shape parameters: $\beta \in \mathbb{R}^{10}$
  - Shape remains constant throughout the video.



[1] Loper *et al.* SMPL: A Skinned Multi-Person Linear Model. *ECCV* 2016.

Bundle Adjustment

- Objective function consists of reprojection, temporal and prior terms

$$E(\beta, \theta, \mathbf{\Omega}) = E_R(\beta, \theta, \mathbf{\Omega}) + E_T(\beta, \theta, \mathbf{\Omega}) + E_P(\theta, \beta)$$

- Use the SMPL body model [1] to parameterise the 3D pose
- 3D joints (and mesh vertices) are obtained from the differentiable SMPL function
- 3D joints, $\mathbf{X}^t = \mathrm{SMPL}(\beta, \theta^t)$.
- Assume scaled orthographic projection, Π, with camera parameters $\mathbf{\Omega}^t = \{s^t, u^t\}$.
- We can project this onto 2D using the camera parameters.
- 2D joints, $\mathbf{x}^t = s^t \Pi(R\mathbf{X}^t) + u^t$.

Reprojection Error

- Encourage 3D joint to project onto predicted 2D keypoints.

$$E_R(\beta, \theta, \mathbf{\Omega}) = \lambda_R \sum_t^T \sum_i^J w_i \rho(\mathbf{x}_i^t - \mathbf{x}_{det,i}^t).$$

- We use 2D human detector of [1].
- Use robust Huber error function
- And weight each reprojection term by the keypoint detector's confidence.

Input keypoints

Predicted joint projection

Predicted 3D mesh

[1] Papanderou *et al.* Towards accurate multi-person pose in the wild. *CVPR* 2017.

Temporal Error

- Encourage smooth motion of predicted:
  - 3D joints, $\mathbf{X}$
  - 2D joint projection, $\mathbf{x}$
  - camera parameters, $\mathbf{\Omega}$

$$E_T(\beta, \theta, \mathbf{\Omega}) = \sum_{t=2}^{T} \sum_{i=1}^{J} \lambda_1 \rho(\mathbf{X}_i^t - \mathbf{X}_i^{t-1}) + \lambda_2 \rho(\mathbf{x}_i^t - \mathbf{x}_i^{t-1}) + \lambda_3 \rho(\mathbf{\Omega}^t - \mathbf{\Omega}^{t-1})$$

3D Prior

- Many 3D poses (some not humanly possible) that project correctly onto 2D and temporally smooth.
- One term encourages solution to stay close to the initialisation, the other the commonly used GMM pose prior [1].

No prior

Prior

[1] Bogo *et al.* Keep it SMPL … *ECCV* 2016.

3D Prior

- Many 3D poses (some not humanly possible) that project correctly onto 2D and temporally smooth.
- One term encourages solution to stay close to the initialisation, the other the commonly used GMM pose prior [1].

$$E_P(\beta, \theta) = \sum_t^T E_J(\theta^t) + \lambda_I E_I(\theta^t, \beta)$$

$$E_J(\theta) = -\log\left(\sum_i g_i \mathcal{N}\left(\theta^t; \mu_i, \Sigma_i\right)\right)$$

$$E_I(\theta^t, \beta) = \sum_i^J \rho(\mathbf{X}_i^t - \tilde{\mathbf{X}}_i^t) + \lambda_\beta \rho(\beta - \tilde{\beta}^t).$$

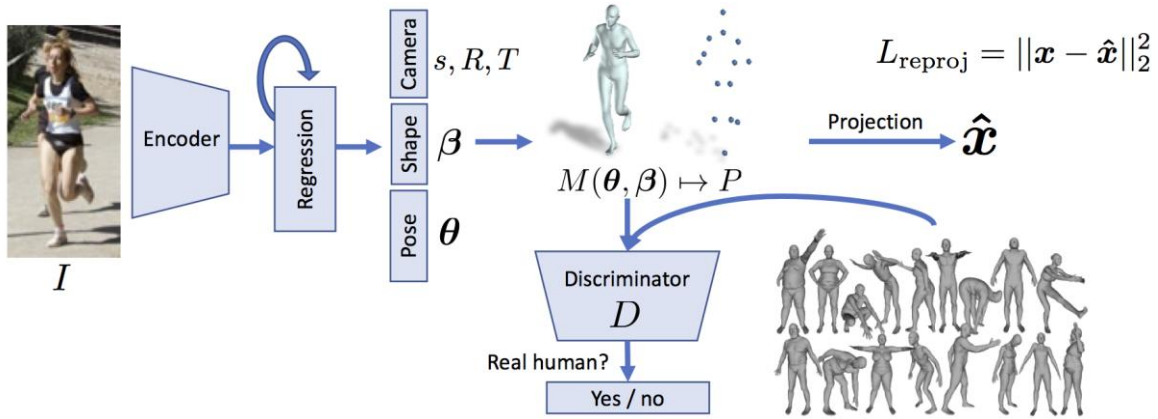
[1] Bogo *et al.* Keep it SMPL … *ECCV* 2016.

Bundle Adjustment

- Objective function consists of reprojection, temporal and prior terms

$$E(\beta, \theta, \mathbf{\Omega}) = E_R(\beta, \theta, \mathbf{\Omega}) + E_T(\beta, \theta, \mathbf{\Omega}) + E_P(\theta, \beta)$$

- Use the per-frame results of HMR [1] to initialise.
- Optimise with L-BFGS
  - Only optimising SMPL- and camera-parameters
  - $10 + 75F$ parameters where F is the number of frames in the video



[1] Kanazawa *et al.* End-to-end recovery of human shape and pose. CVPR 2018.

Scaling up to Kinetics

- Data very noisy.
- Initialisation from HMR and 2D pose detector often incorrect.
- Also need to deal with multiple people
  - Tracking person-of-interest not robust to detection failures.
- Modify reprojection loss instead

$$E_R(\beta, \theta^t, \mathbf{\Omega}^t) = \min\left(\min_{p \in P^t} \sum_i^J w_i h(\mathbf{x}_i^t - \mathbf{x}_{det,i}^{t,p}), \tau_R\right)$$

Scaling up to Kinetics

- Data very noisy.
- Initialisation from HMR and 2D pose detector often incorrect.
- Also need to deal with multiple people
- Modify reprojection loss instead

$$E_R(\beta, \theta^t, \mathbf{\Omega}^t) = \min\left(\min_{p \in P^t} \sum_i^J w_i h(\mathbf{x}_i^t - \mathbf{x}_{det,i}^{t,p}), \tau_R\right)$$

- "Inner min" means the loss is with respect to the best matching 2D pose estimate
- "Outer min" means that if our estimate is too far from 2D pose, we consider it an outlier and pay a constant penalty.

Exploiting temporal consistency

[1] Kanazawa *et al.* End-to-end recovery of human shape and pose. CVPR 2018.

Exploiting temporal consistency

Input

HMR mesh overlay

Bundle adjustment mesh overlay

[1] Kanazawa *et al.* End-to-end recovery of human shape and pose. CVPR 2018.

Ablation study on Human 3.6M

- Mocap dataset, has metric 3D ground truth
- Allows us to set hyperparameters.
- Each term in objective improves result.
- Ground truth keypoints provide substantial benefits
  - Occluded keypoints help a lot

Method	MPJPE (mm)	PA-MPJPE (mm)
HMR initialisation [20]	85.8	57.5
E_R	154.3	99.7
$E_R + E_P$	79.6	55.3
$E_R + E_P + E_T$	77.8	54.3
E_R (gt. keypoints)	89.2	64.5
$E_R + E_P$ (gt. keypoints)	66.5	45.7
$E_R + E_P + E_T$ (gt. keypoints)	63.3	41.6

Comparison on Human 3.6M

- Compare to other methods using the SMPL model
- Our whole-video approach also achieves state-of-the-art performance on Human 3.6M

Method	MPJPE (mm)	PA-MPJPE (mm)
Self-Sup [49]	–	98.4
Lassner *et al.* direct fitting [23]	–	93.9
SMPLify [7]	–	82.3
Lassner *et al.* optimisation [23]	–	80.7
Pavlakos *et al.* [36]	–	75.9
NBF [32]	–	59.9
MuVS (Note uses 4 cameras) [16]	–	58.4
HMR [20]	88.0	56.8
Ours	**77.8**	**54.3**

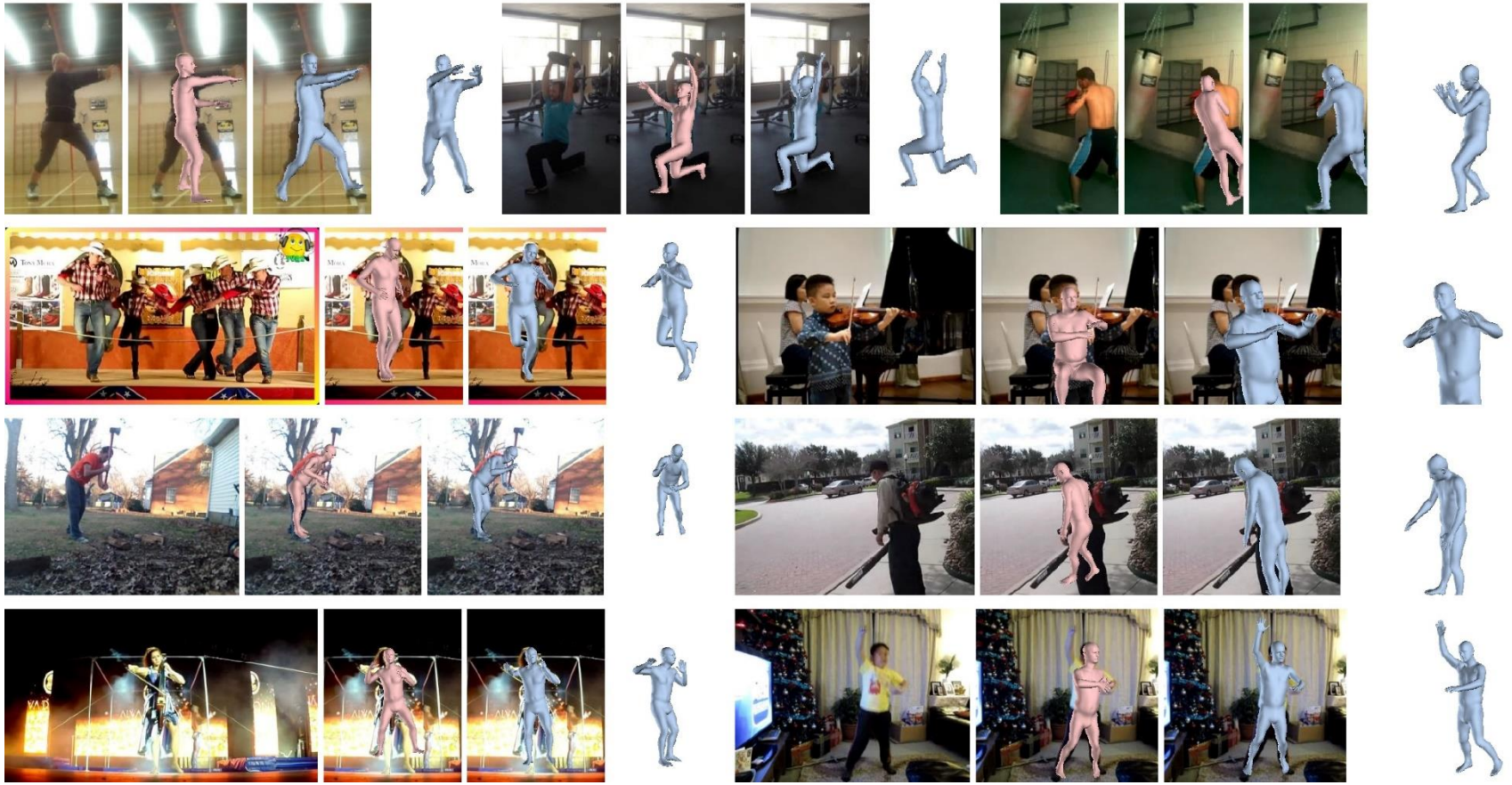
Scaling up to Kinetics

- Run our method on 106 589 YouTube videos in the Kinetics dataset.
- Thresholding the normalised loss, we obtain 16 720 videos containing 4.1 million frames.
- We keep the frames where our 2D projections match that of our 2D keypoint detector.
- Final dataset is 3.4 million frames.
- Available to public: https://github.com/deepmind/Temporal-3D-Pose-Kinetics

Example of video automatically filtered out

Scaling up to Kinetics

- Auto-generated dataset contains diversity in pose, scene, action and camera viewpoints not found in mocap.



HMR (per-frame model)

Ours

Effect of our new dataset

- Training with our new automatically-generated dataset improves the performance of HMR on two datasets.
  - 3DPW – “in-the-wild”
  - HumanEVA – mocap
- More improvement from 3.4 million additional frames, than 300 000 frames.

PA-MPJPE error (mm) on 3DPW and HumanEVA datasets

Dataset	Original data	Original + Kinetics 300K	Original + Kinetics 3M
3DPW	77.2	73.8	**72.2**
HumanEVA	85.7	83.5	**82.1**

[1] Kanazawa *et al.* End-to-end recovery of human shape and pose. CVPR 2018.

Conclusion

- Interpret mean-field inference of generic CRFs as a recurrent network
- Enables end-to-end training of CNN and CRF models
- Applications in structured prediction problems like semantic- and instance/panoptic-segmentation
- We can effectively labelled weakly labelled data as supervision for complex tasks.
- Weakly supervised panoptic segmentation
  - Up to 95% of fully-supervised performance, with annotation time reduced by as much as 35x.
  - If you have annotation budget, better to label a few images well rather than many images poorly.
- 3D human pose estimation
  - Can leverage temporal consistency and unlabelled YouTube videos to improve existing models

Collaborators

- Sadeep Jayasumana
- Shuai Zheng
- Qizhu Li
- Philip Torr
- Carl Doersch
- Andrew Zisserman

Questions?

A Arnab, S Jayasumana, S Zheng, P Torr. *Higher Order Conditional Random Fields in Deep Neural Networks*. ECCV, 2016

A Arnab and P Torr. *Pixelwise Instance Segmentation with a Dynamically Instantiated Network*. CVPR, 2017

A Arnab, O Miksik, P Torr. *On the Robustness of Semantic Segmentation Models to Adversarial Attacks*. CVPR, 2018

Q Li*, A Arnab*, P Torr. *Weakly- and Semi-Supervised Panoptic Segmentation*. ECCV, 2018

A Arnab*, C Doersch*, A Zisserman. *Exploiting Temporal Context for 3D Human Pose Estimation in the Wild.* CVPR, 2019.