Transfer Learning for Visual Scene Understanding

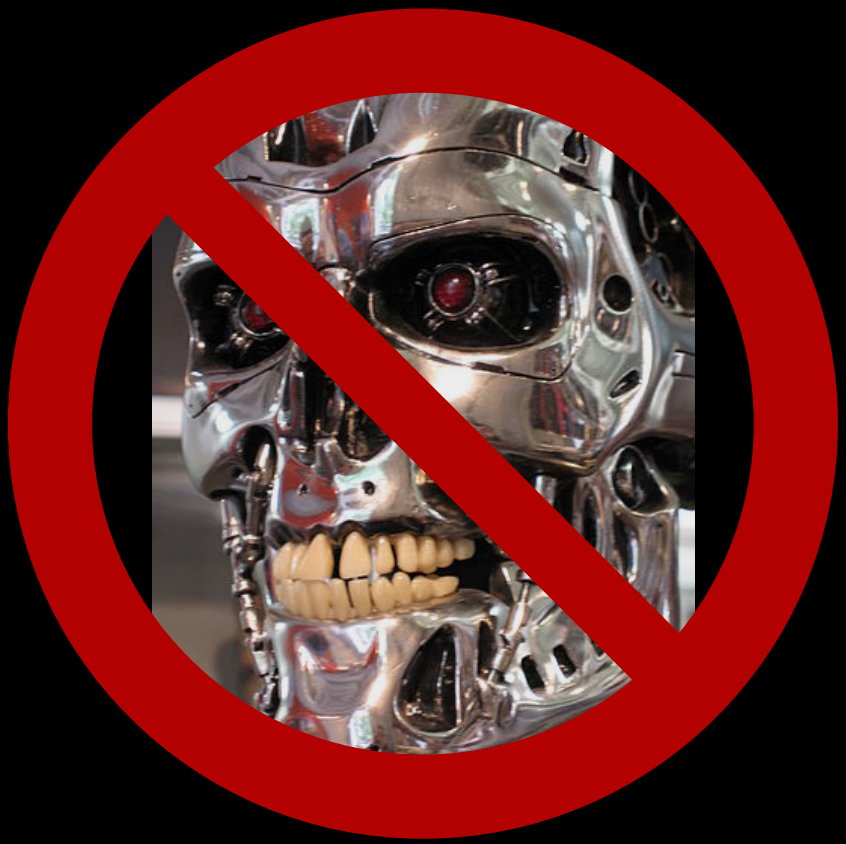
Christoph Lampert

IST AUSTRIA
Institute of Science and Technology

Qualcomm Augmented Reality Lecture Series
June 16, 2016

Ultimate goal

Automatic systems that learn and act autonomously



Medium term goal

Automatic systems that can analyze and interpret data

→

Image Understanding

→

"Three men sit at a table in a pub, drinking beer. One of them talks while the other two listen."

Image: British Broadcasting Corporation (BBC)

State of the art

Analyze individual aspects of visual data

Scene Classification

- indoors
- in a pub

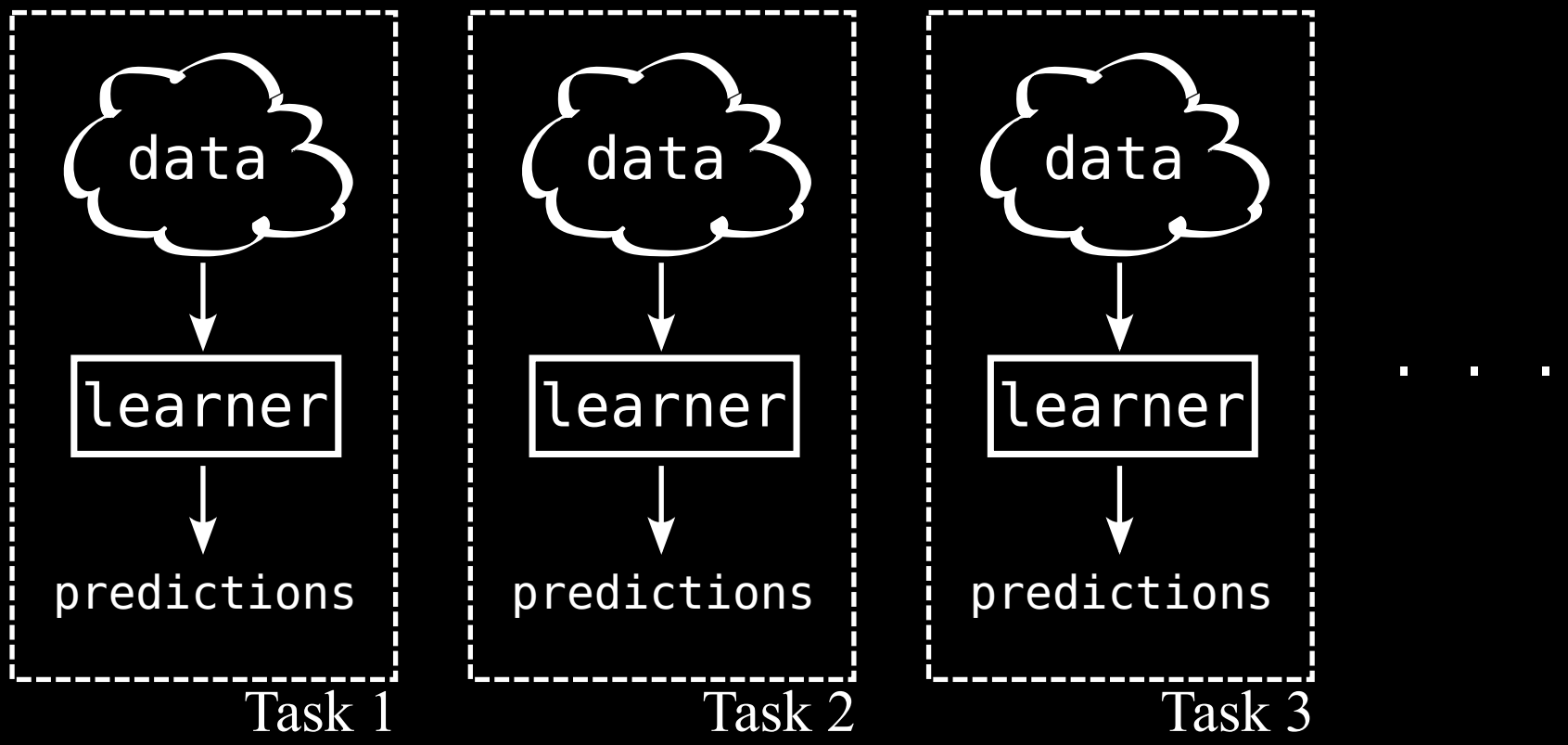
Action Classification

- drinking
- talking

Object Recognition

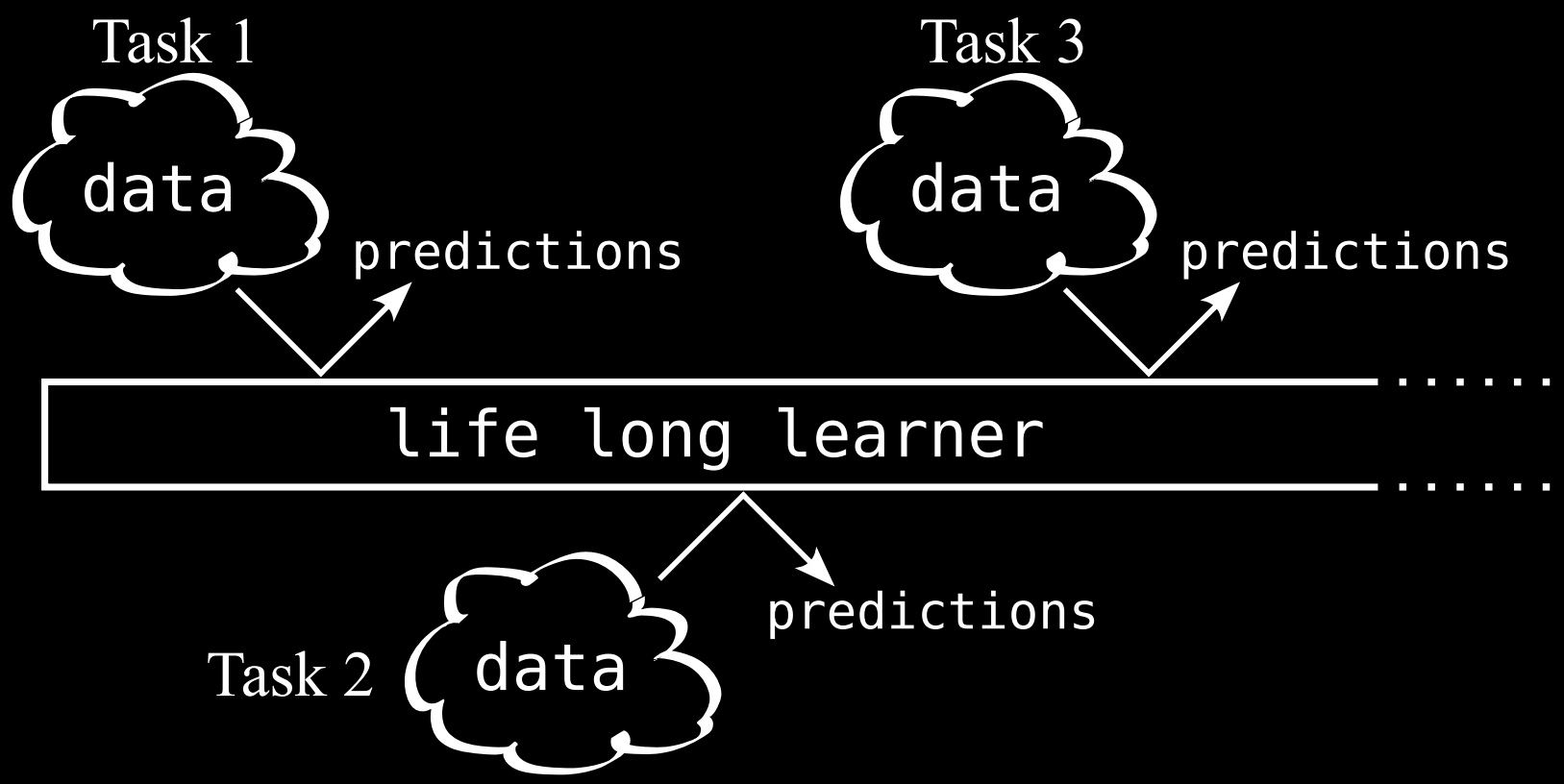
- three persons
- one table
- three glasses

State of the art



Tabula Rasa Learning

Future challenge: towards continuously improving systems



Lifelong Learning

Research topics

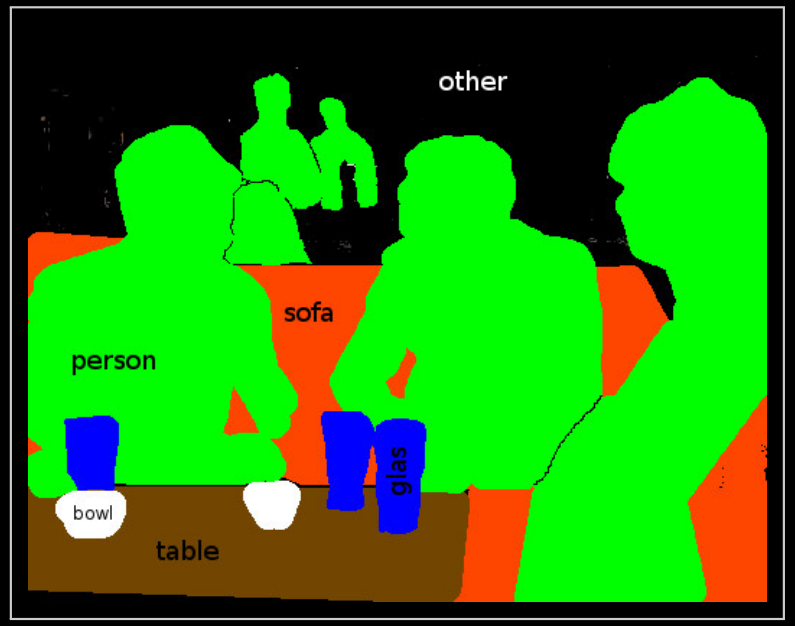
Machine Learning

- Multi-task learning
- Domain adaptation
- Learning to learn
- Learning with weak supervision

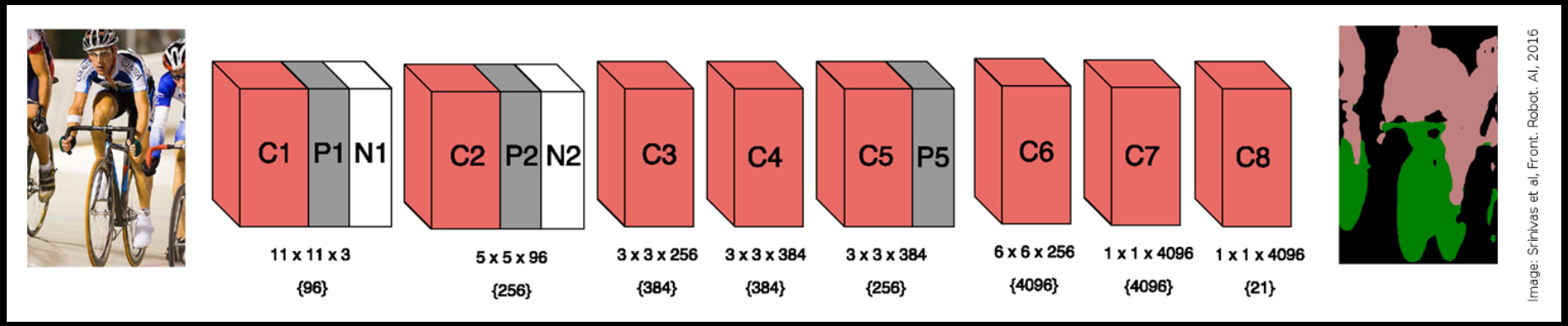
Computer Vision

- Object recognition
- Object localization
- Semantic segmentation
- Attribute-based classification

Semantic Image Segmentation



State-of-the-art: Convolutional Neural Networks (CNNs)



- deep neural network, all layers convolutional
- predict per-pixel output from per-pixel input
- trained from images with per-pixel ground truth

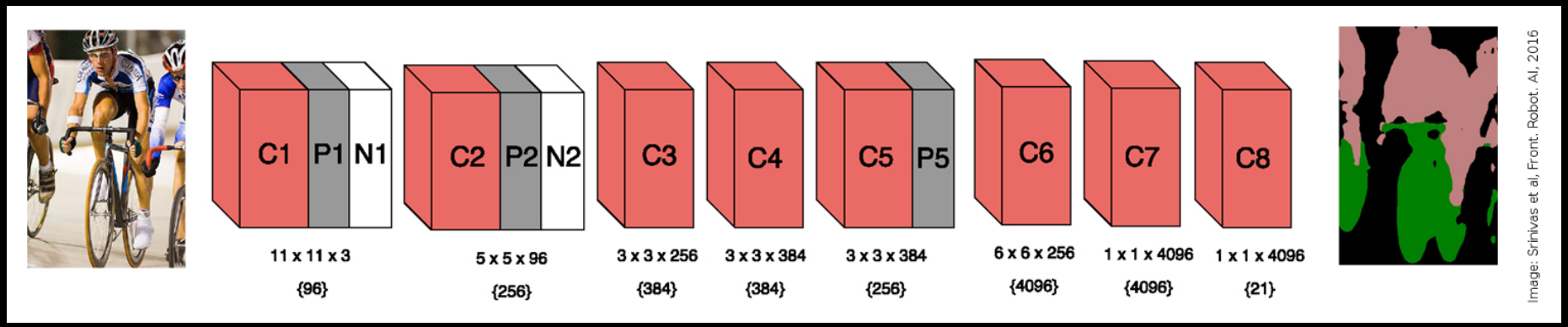

,

,

,

, . . .

State-of-the-art: Convolutional Neural Networks (CNNs)



- training set $\{(x^1, y^1), \ldots, (x^m, y^m)\}$
- images x^i, ground truth segmentation masks y^i
- CNN output $f_{u,c}(x;\theta)$ (probability of class c at location u)
- measure quality of one prediction by a loss function, e.g.

$$\text{loss}(f_{u,c}, y) = -\sum_c \sum_u y_{u,c} \log f_{u,c}(x)$$

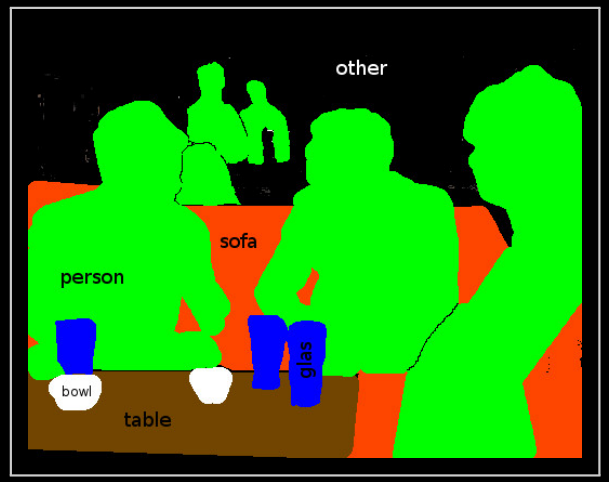
- learn CNN parameters by minimizing loss over training set

$$\min_\theta \sum_{i=1}^m \text{loss}(f_{u,c}(x^i), y^i)$$

Problem: creating per-pixel annotation cost a lot of time

Weakly-Supervised Semantic Segmentation

→



- train from images with per-image class labels (tags)

cat
sofa

,

horse

,

table
chair

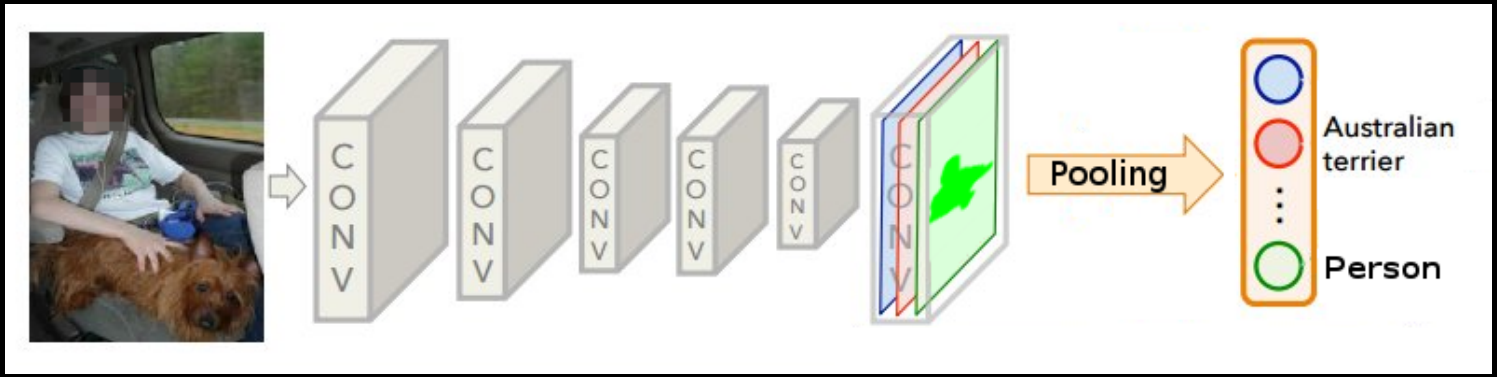
,

motorbike

, ...

- annotation is much weaker, but much easier to generate

Weakly-Supervised Semantic Segmentation



Training:

- training set $\{(x^1, T^1), \ldots, (x^m, T^m)\}$
- images x^i, tag annotation T^i, e.g. $T^i = \{\texttt{cat}, \texttt{dog}\}$
- how to measure quality of a predicted segmentation mask?
  1) pool per-pixel scores $f_{u;c}(x)$ into per-image scores, $G_c(x)$
  2) measure if correct classes were predicted, e.g.

$$\text{loss}(x, T) = -\sum_{c \in T} \log G_c(x) \; - \sum_{c \in \mathcal{C} \setminus T} \log(1 - G_c(x))$$

Problem: it doesn't work very well...

Image: adapted from [Zhou, Khosla, Lapedriza, Oliva, Torralba. *"Learning Deep Features for Discriminative Localization"* CVPR 2016],

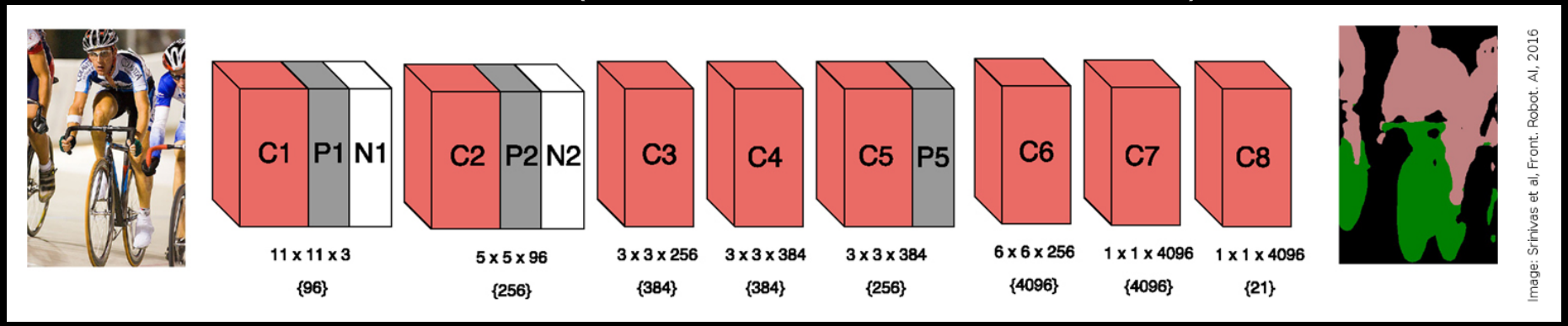
Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation

Alex Kolesnikov

A. Kolesnikov,, CHL, *"Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation"*, arXiv:1603.06098 [cs.CV]

SEC: Seed, Expand and Constrain

VGG-style deep network (16 layers, all convolutional):



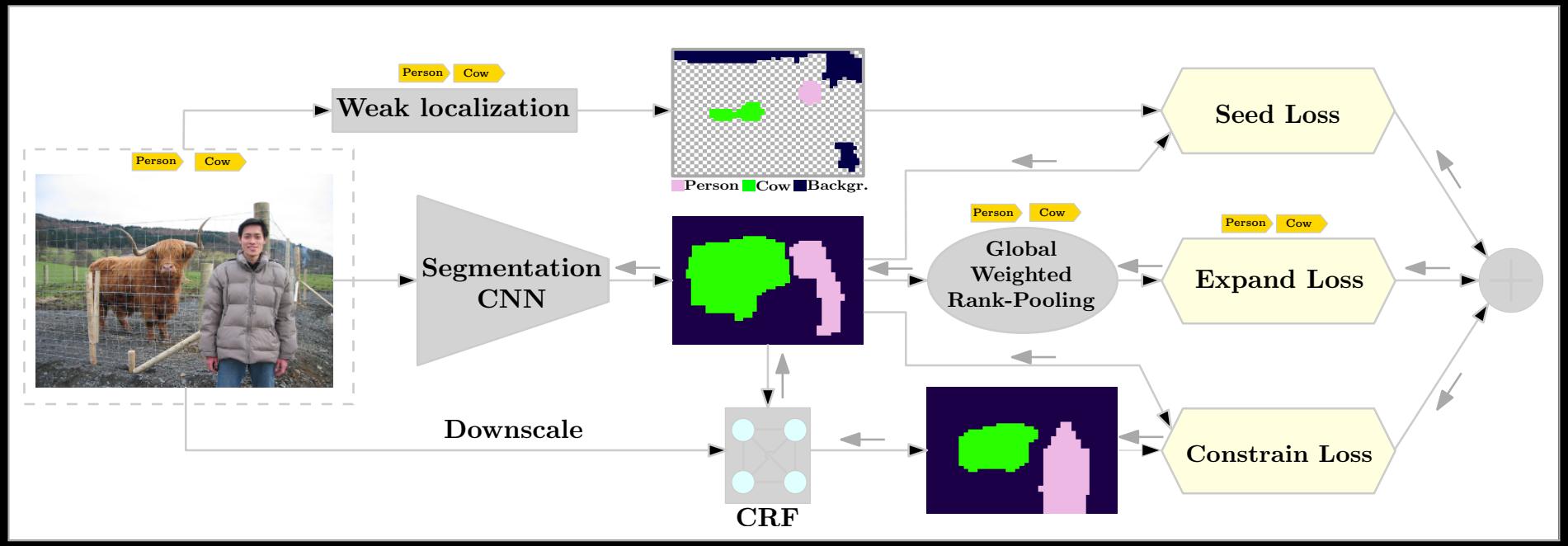
Required for training:

- pre-trained classification network
- image dataset to learn from (with per-image class labels)

Main contribution: new, three-part, loss function

$$L_{\text{seed}}(x, f(x;\theta), T) + L_{\text{exp.}}(f(x;\theta), T) + L_{\text{cnstr.}}(x, f(x;\theta))$$

SEC: Seed, Expand and Constrain



Seed loss:

- network should reproduce weak cues from classification network

Expand loss:

- network should produce reasonable object sizes

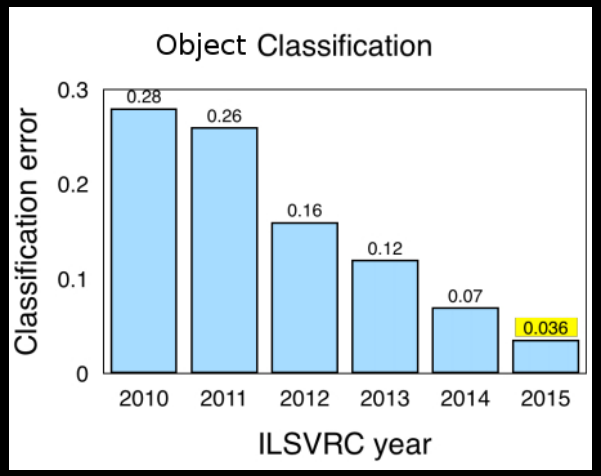
Constrain loss:

- network should respect boundaries (image gradients)

SEC: Seed, Expand and Constrain

Observation:

- Convolutional networks achieve very good results in full-image classification tasks



- If we know which part of the image caused the network to make its decision, we can find out where the object is.

Image: http://blog.clip.mn/2016/01/06/the-relevance-of-artificial-intelligence-to-digital-video-creation-consumption-and-monetization/

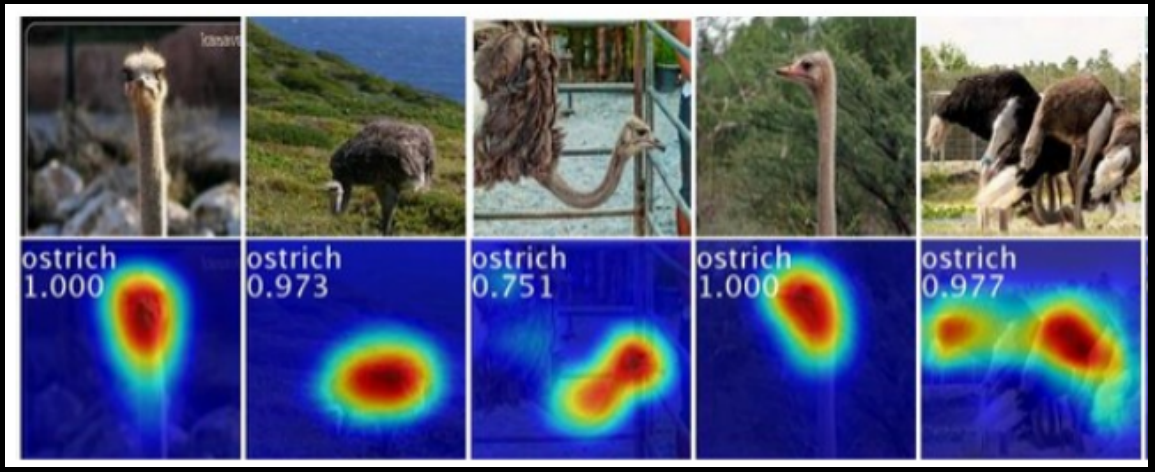
SEC: Seed, Expand and Constrain

Multiple possibilities:

1) gradient back-propagation from label to image
   "which change to the image affects the score the most?"

2) mask out different image regions and observe the score

3) use a network with spatial representation until the last layer

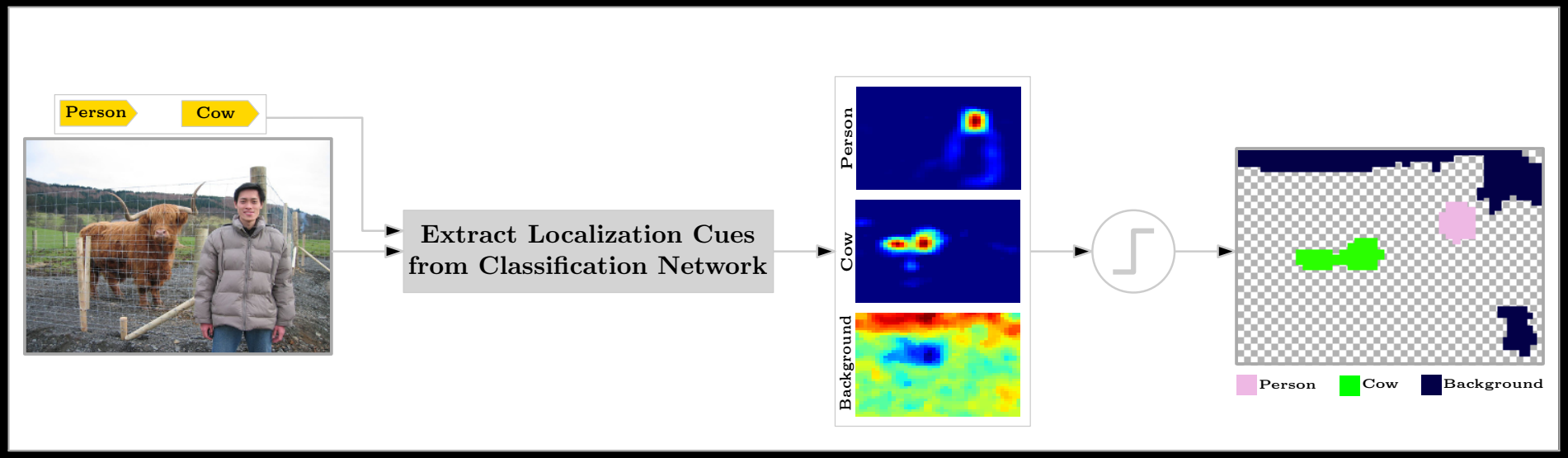


Images: adapted from [Zhou, Khosla, Lapedriza, Oliva, Torralba. *"Learning Deep Features for Discriminative Localization"* CVPR 2016],

SEC: Seed, Expand and Constrain

Observation:

- Heatmaps, $g_u(x)$, from classification network gives only rough localization, not segmentation mask

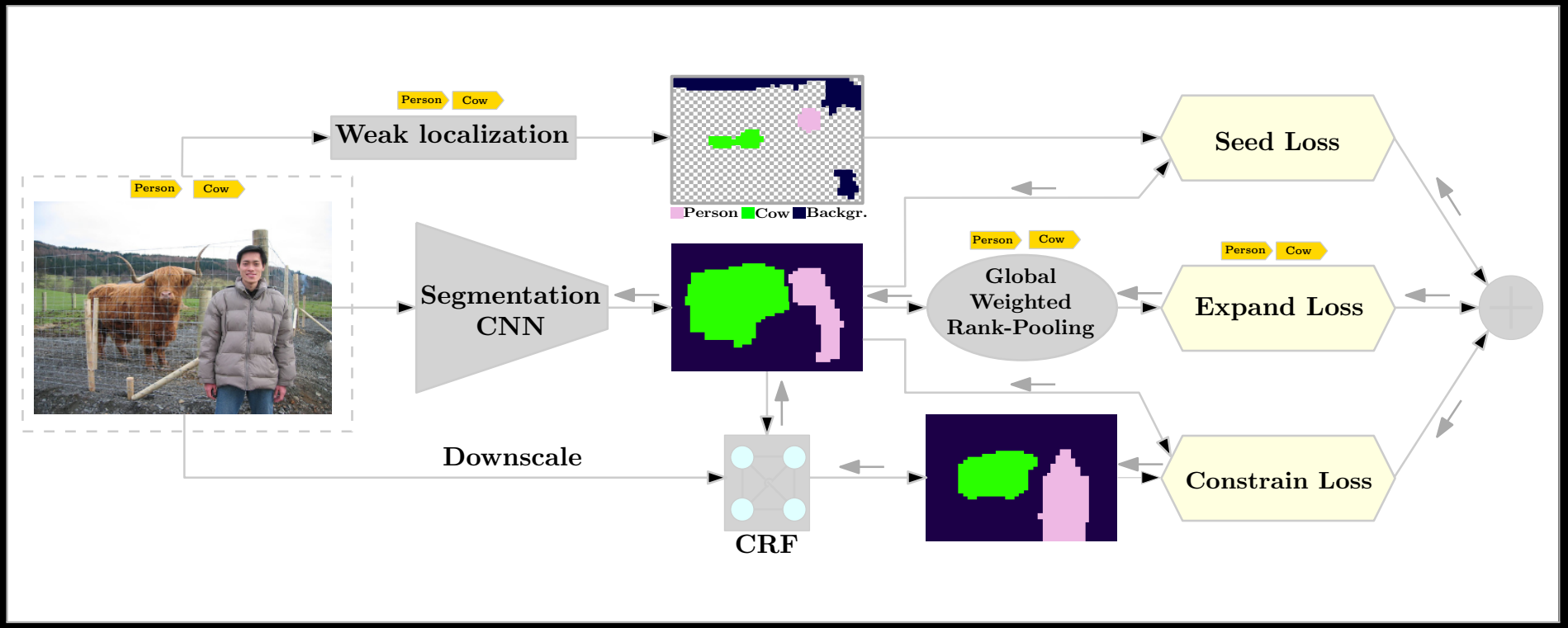


We trust only the most confident core areas $\rightarrow$ *seed regions*

$$S_c(x) = \{u : g_u(x) \geq 0.2\alpha\} \quad \text{for} \quad \alpha = \max_v g_v(x)$$

(can be precomputed)

SEC: Seed, Expand and Constrain



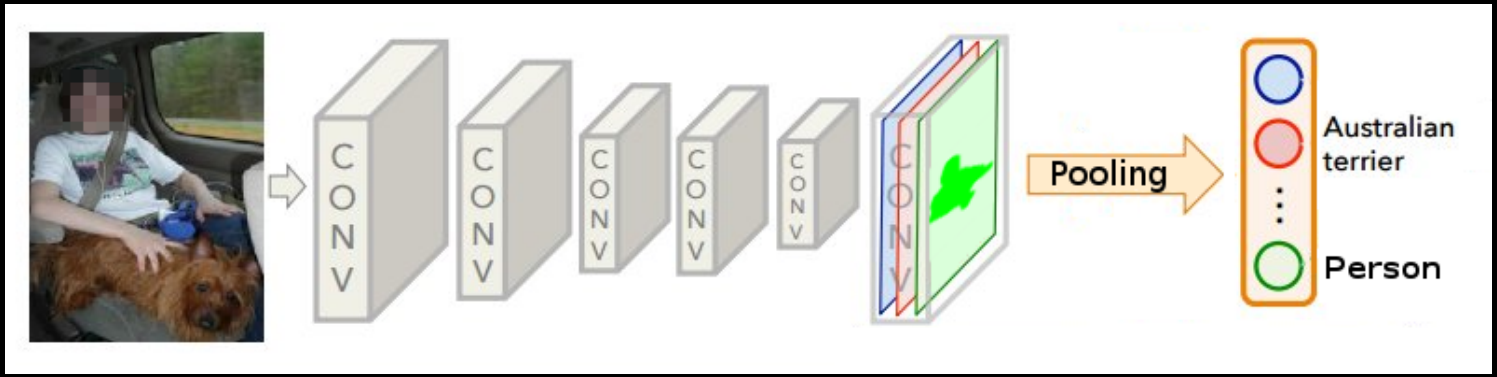
Seed loss:

$$L_{\text{seed}}(X, f(X), T) = \frac{-1}{\sum_{c \in T} |S_c|} \sum_{c \in T} \sum_{u \in S_c} \log f_{u,c}(X)$$

network should produce correct labels where weak cues tell it to

SEC: Seed, Expand and Constrain

Observation: Pooling method influences predicted object sizes



- max pooling†: class score is maximum of per-pixel scores
  all weight lies on a single pixel → bias towards small objects
- average pooling*: class score is average of pixel scores
  all pixels have the same weight → bias towards large objects
- ideal: all object pixels contribute, but none of the others

---

* [Zhou, Khosla, Lapedriza, Oliva, Torralba. *"Learning Deep Features for Discriminative Localization"* CVPR 2016],

† [Oquab, Bottou, Laptev, Sivic; *"Is Object Localization for Free? - Weakly-Supervised Learning With Convolutional Neural Networks"*, CVPR 2015]

SEC: Seed, Expand and Constrain

Proposed: global weighted rank pooling (GWRP):

- sort pixels by their activation score, $i_1, \ldots, i_n$

$$f_{i_1,c}(x) \geq f_{i_2,c}(x) \geq \cdots \geq f_{i_n,c}(x)$$

- pool by linear combination with decreasing weights:

$$G_c(x; d_c) = \frac{1}{Z(d_c)} \sum_{j=1}^{n} (d_c)^{j-1} f_{i_j,c}(x), \quad \text{for } Z(d_c) = \sum_{j=1}^{n} (d_c)^{j-1}.$$

- d_c: decay parameter for class c.

Generalizes other poolings:

- $d_c = 0$: max pooling $\qquad$ $d_c = 1$: average pooling
- $0 < d_c < 1$: put more emphasis on locations with high scores

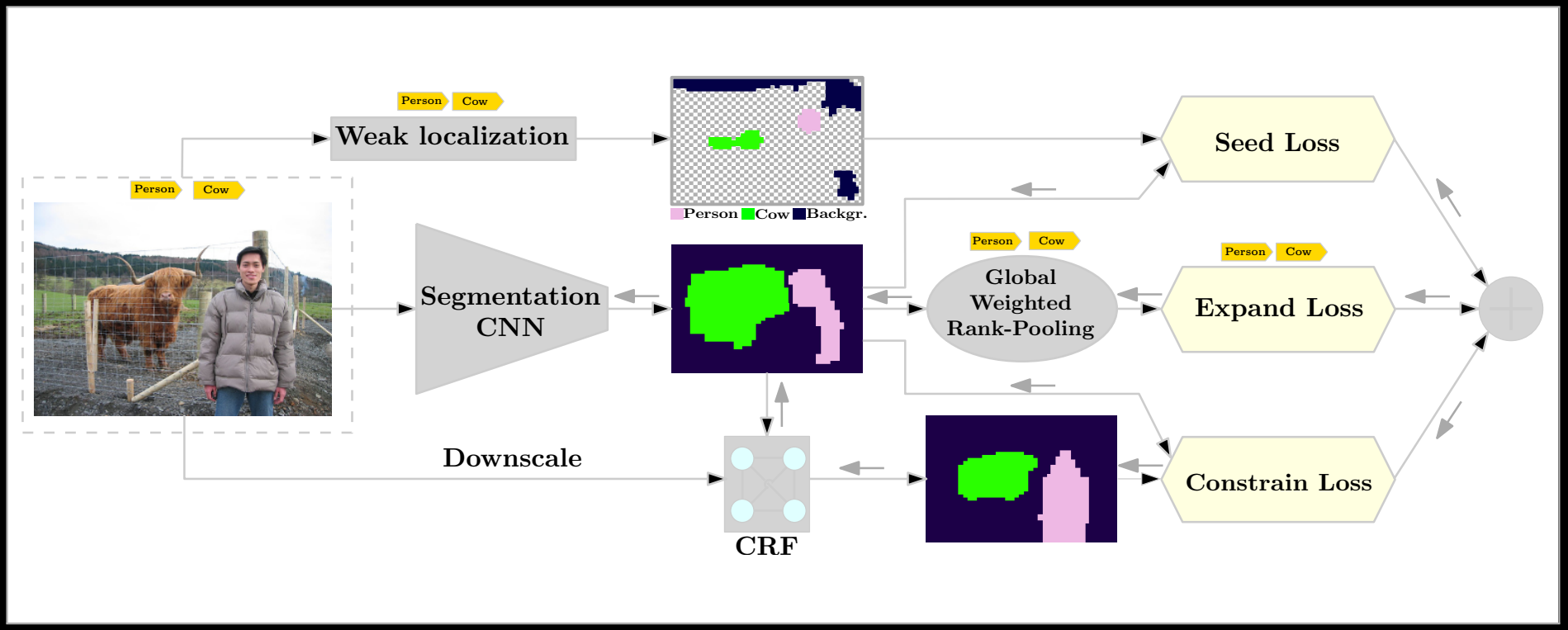
SEC: Seed, Expand and Constrain

Our choices: (for $n = 41 \times 41$)

- for classes c that are in the image: $d_+ = 0.996$
  $\rightarrow$ 50% of weight on top 10% pixels
- for classes c that are not in the image: $d_- = 1$
  $\rightarrow$ max-pooling, no pixel should have high object score
- for `background` class: $d_{\mathrm{bg}} = 0.999$
  $\rightarrow$ 50% of weight on top 30% pixels

Expresses our prior belief about object sizes.

SEC: Seed, Expand and Constrain



Loss function:

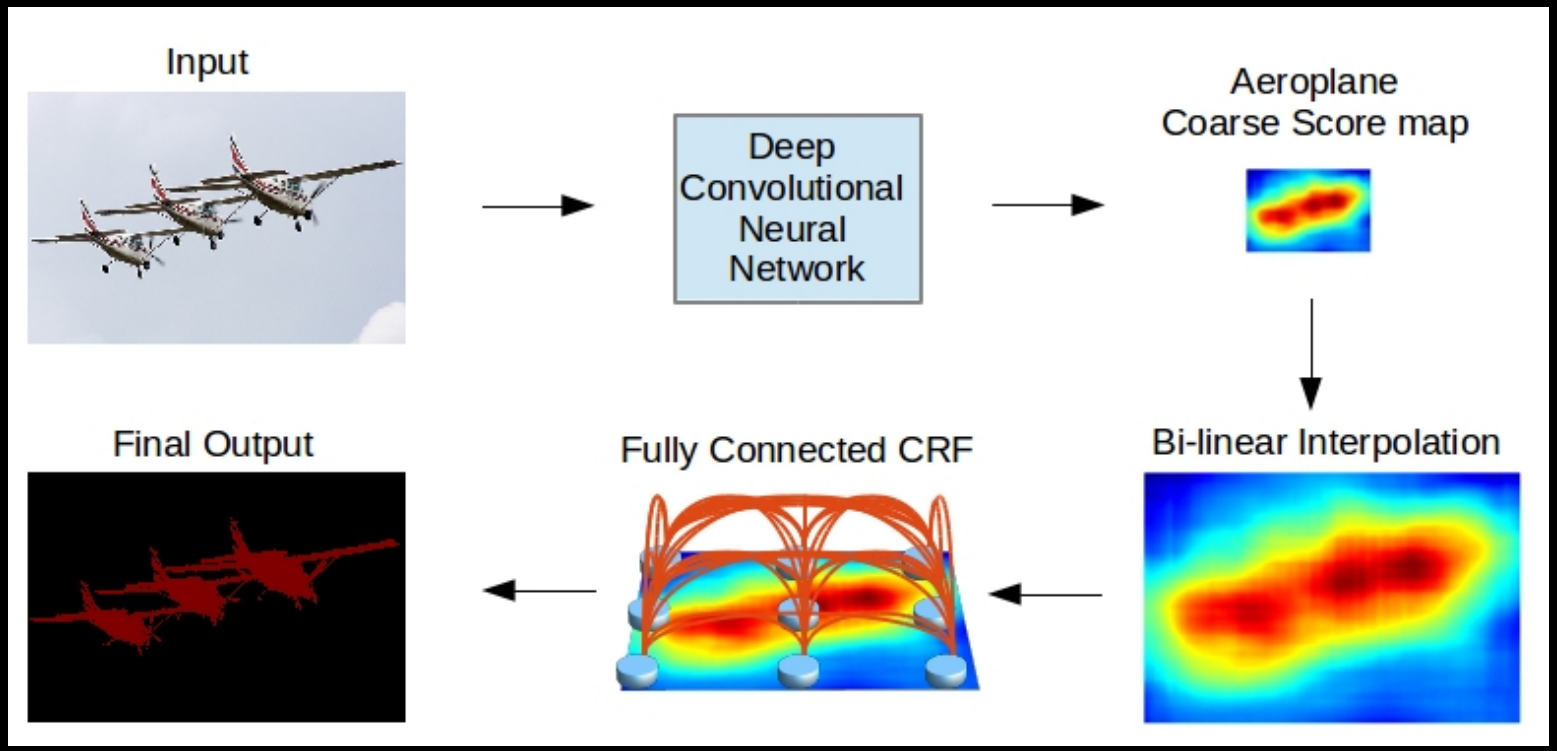
$$L_{\text{exp.}}(f(X), T) = -\frac{1}{|T|}\sum_{c\in T}\log G_c(X; d_+) - \log G_{c^{\text{bg}}}(X; d_{\text{bg}})$$

$$-\frac{1}{|\mathcal{C}\setminus T|}\sum_{c\in\mathcal{C}\setminus T}\log(1 - G_c(X; d_-))$$

SEC: Seed, Expand and Constrain

Observation:

- applying a fully-connected conditional random field (CRF) to the network output yields crisp segmentation



Can we make use of this property also at training time?

Images: [Chen, Papandreou, Kokkinos, Murphy, Yuille. *"Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs"*, ICLR 2015]

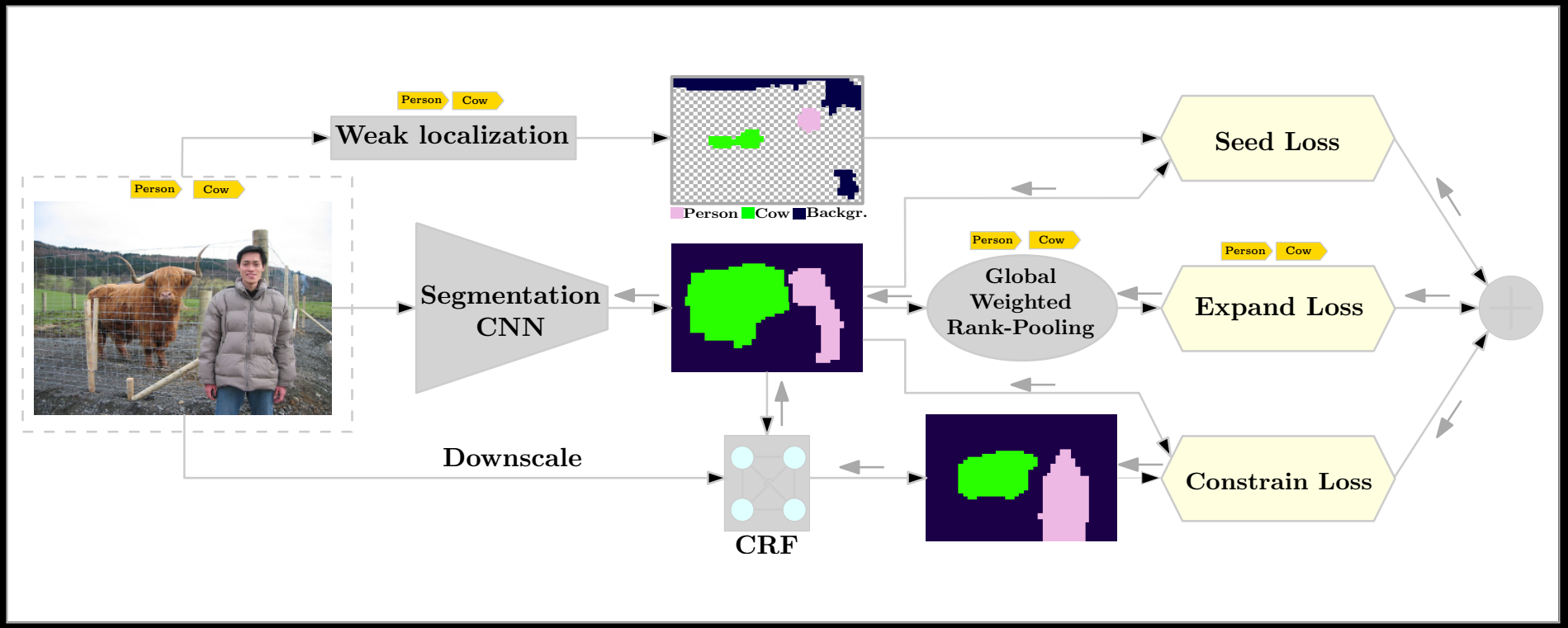
SEC: Seed, Expand and Constrain

Main idea:

- learn network such that per-pixel predictions look like CRF predictions (in particular: follow image boundaries)
- $Q_{u,c}$(x) is output of CRF with CNN outputs $f_{u,c}(x)$ as inputs
- measure difference between CNN output and CRF output (by Kullback-Leibler divergence)

$$\mathrm{KL}(p||q) = \sum_i p_i \, \frac{\log p_i}{\log q_i}$$

SEC: Seed, Expand and Constrain



Loss function:

$$L_{\text{cnstr.}}(x, f(x)) = \frac{1}{|I|} \sum_{u \in I} \sum_{c \in \mathcal{C}} Q_{u,c}(x) \log \frac{Q_{u,c}(x)}{f_{u,c}(x)}$$

Training

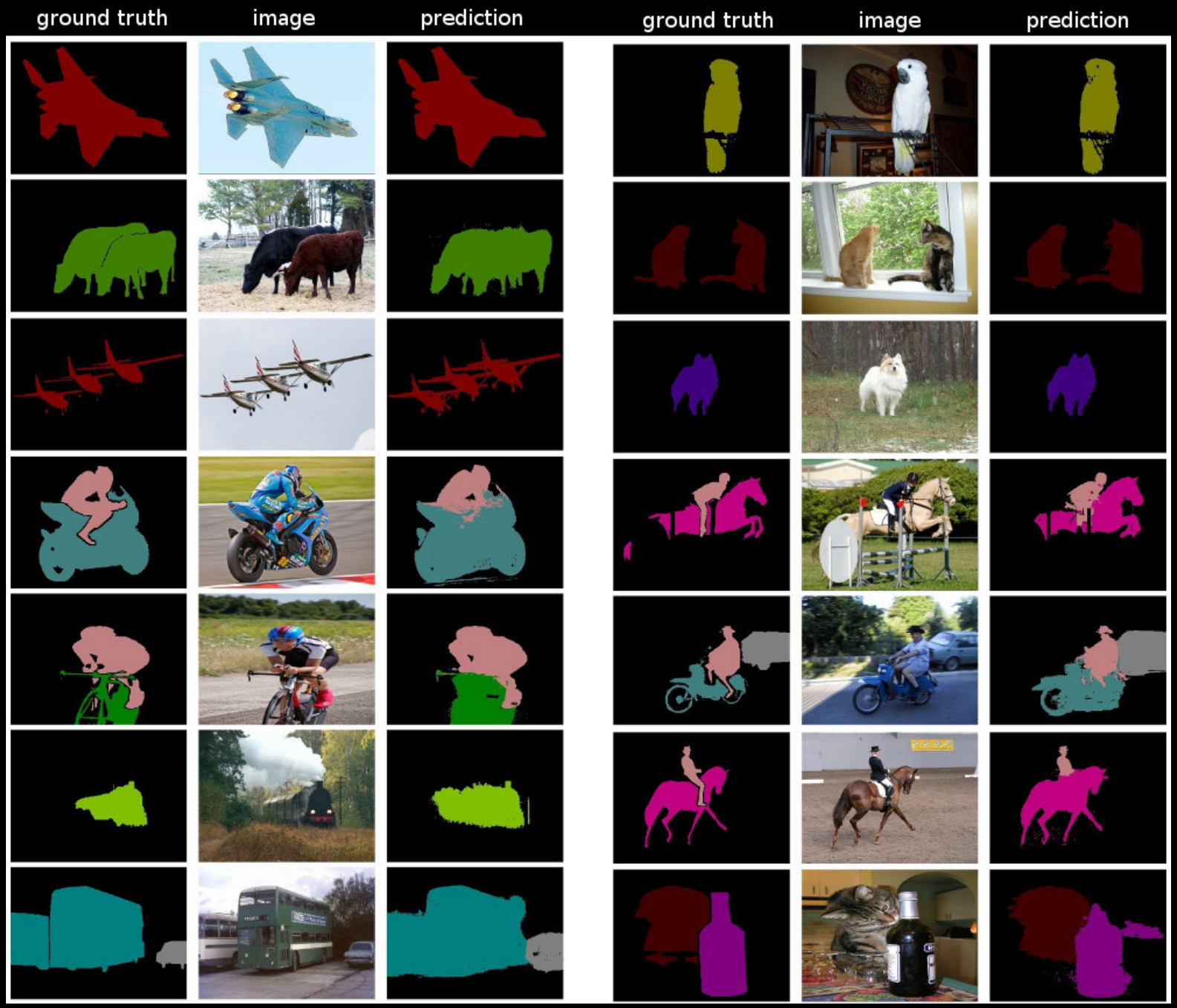
- continuous in all parameters ("end-to-end differentiable")
- gradients computated automatically using `theano`
- stochastic gradient descent (backpropagation)
  - minibatches (size 15)
  - 8000 iterations
  - dropout rate 0.5
  - weight decay 0.0005
  - initial learning rate 0.001, decreased every 2000 iterations
- 7-8 hours on `GeForce TITAN-X` GPU

Data

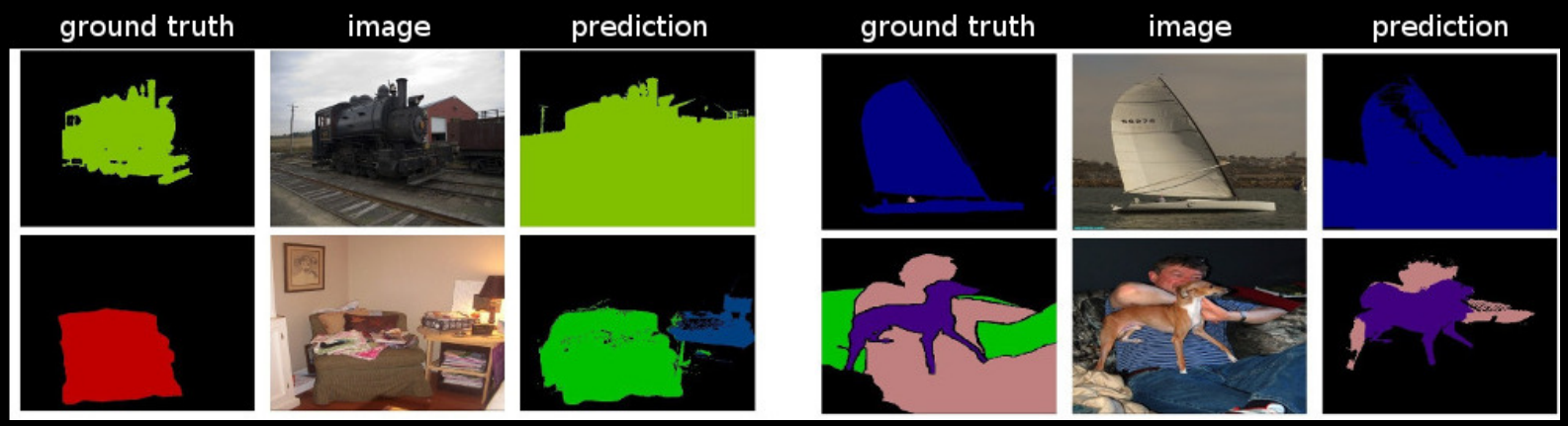
- PASCAL VOC 2012 challenge: 20 object classes + background
- training set: 10,582 weakly annotated images
- test set: 1456 images (no public labels, uses evaluation server)
- evaluation by *mean intersection-over-union*

PASCAL VOC 2012 *test* set	MIL-FCN	CCNN	MIL+ILP+SP-sppxl	SEC (proposed)
background		$\approx 71^{\dagger}$	74.7	83.0
aeroplane		24.2	38.8	55.6
bike		19.9	19.8	27.4
bird		26.3	27.5	61.1
boat		18.6	21.7	22.9
bottle		38.1	32.8	52.4
bus		51.7	40.0	70.2
car		42.9	50.1	58.8
cat		48.2	47.1	70.0
chair		15.6	7.2	22.1
cow		37.2	44.8	54.3
diningtable		18.3	15.8	27.9
dog		43.0	49.4	67.4
horse		38.2	47.3	59.4
motorbike		52.2	36.6	70.7
person		40.0	36.4	59.0
plant		33.8	24.3	38.7
sheep		36.0	44.5	58.6
sofa		21.6	21.0	38.1
train		33.4	31.5	37.6
tv/monitor		38.3	41.3	45.2
average	25.7	35.6	35.8	51.5

†: inferred from average



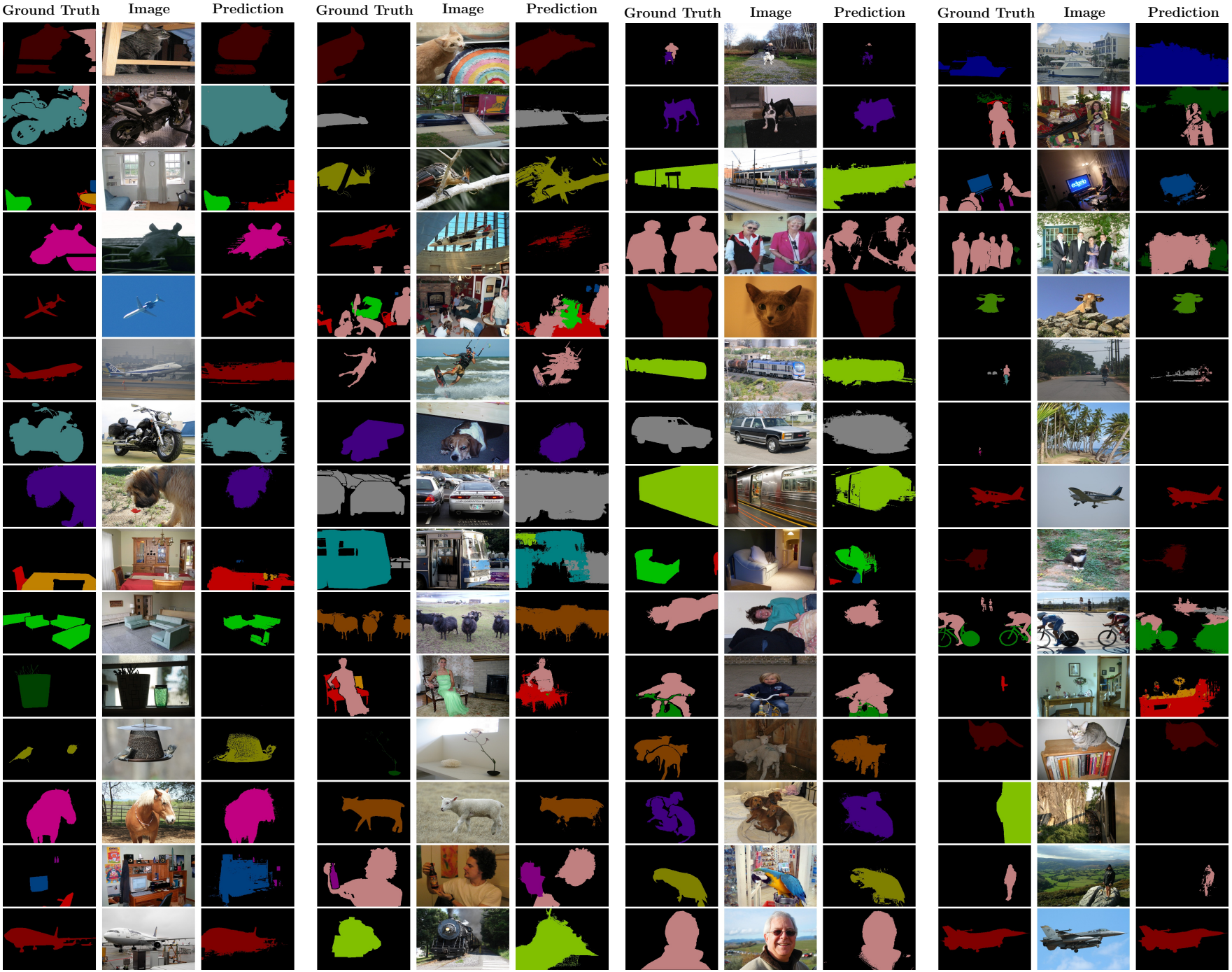
(PASCAL VOC 2012, `val` set)



(PASCAL VOC 2012, `val` set)

Failure cases (aka, future work):

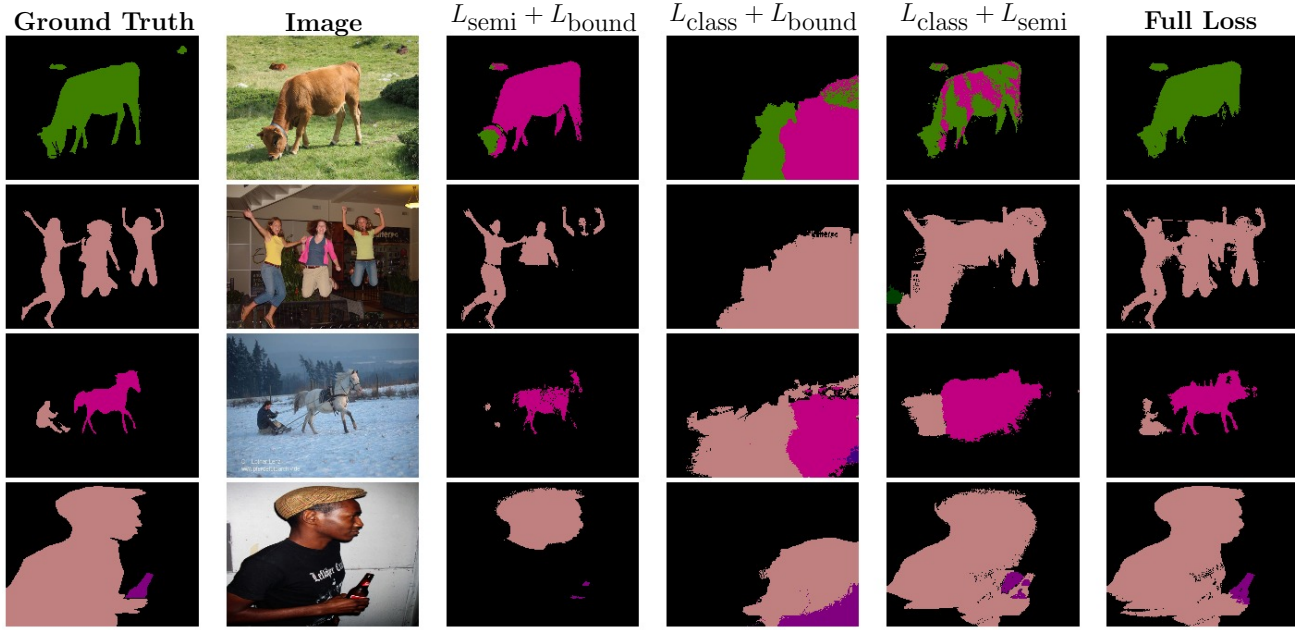
- consistently co-occurring distractors (trains+tracks, boat+water)
- confusion between objects (chair vs. sofa)
- disconnected object (usually due to occlusion)


Ground Truth
Image
Prediction
Ground Truth
Image
Prediction
Ground Truth
Image
Prediction
Ground Truth
Image
Prediction

Insights from Ablation Study

Most important term in loss function: localization seeds

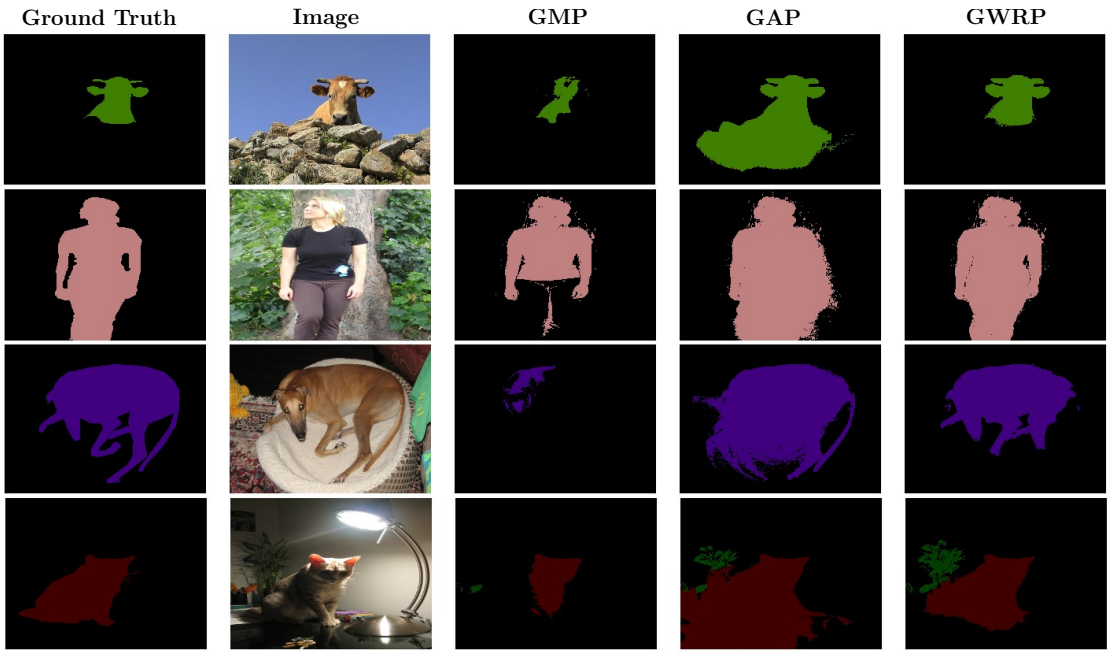
loss function	L_{expand}	L_{seed}	$L_{\text{cnstr.}}$	$L_{\text{seed}} + L_{\text{cnstr.}}$	$L_{\text{expand}} + L_{\text{cnstr.}}$	$L_{\text{seed}} + L_{\text{expand}}$	full SEC loss
mIoU	27.8	49.2	–	49.4	17.2	45.7	50.7



Insights from Ablation Study

Global Weighted Rank Pooling leads to better object sizes than max-pooling or average-pooling

Model	foreground fraction	mIoU
GMP	21.0	47.3
GAP	37.5	45.1
GWRP	26.7	**50.7**
ground truth	27.1	–



Summary

Transfer Learning for Visual Scene Understanding

- Transfer information between different learning tasks
  $\rightarrow$ less training data or less necessary annotation

Weakly-Supervised Image Segmentation

- annotation is weaker (image tags) than the desired system output (segmentation masks)
- requires transfer of information and/or prior knowledge

SEC: Seed, Expand and Constrain [arXiv:1603.06098 [cs.CV]]

- transfer: weak location cues from classification network
- prior knowledge: typical objects sizes
- prior knowledge: objects boundaries align with image gradients
- code and pretrained models will be online

Thanks to...

Our team at IST Austria:

Alex Kolesnikov

Georg Martius

Asya Pentina

Amélie Royer

Alex Zimin

Funding Sources:

IST AUSTRIA
Institute of Science and Technology

erc