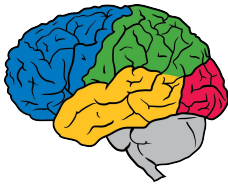
Google



Learning with Proxy Supervision for End-to-End Visual Learning

Jiri Cermak*
Czech Technical Univ., Prague

Anelia Angelova
Google Brain

* Work done while at Google Brain

Stop for pedestrians?

Stop for pedestrians?

Classical approach:

- Find pedestrians
- Find road
- Analyze 3D scene

- Make a decision

Make a complex decision, directly, whether to drive or not.
End-to-end: **From raw inputs to final decision**.

HOW?

End-to-end decision: Input image -> command

Pedestrians

End-to-end decision: Input image -> command

Pedestrians

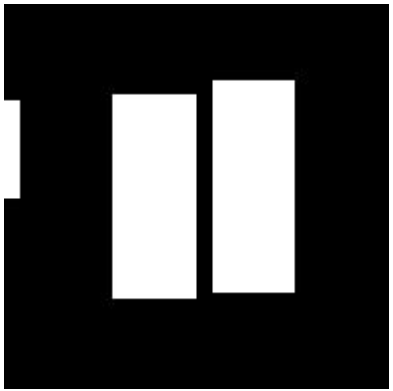
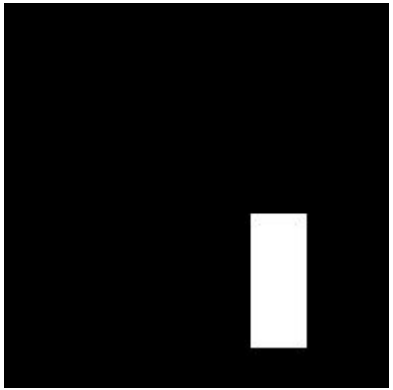
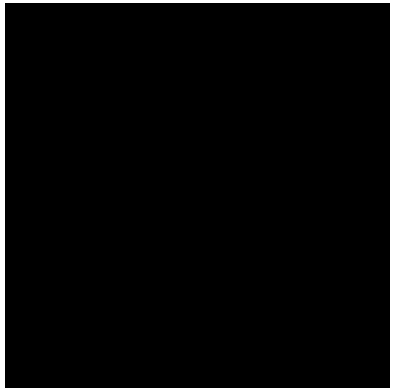
‘STOP’ decision

‘GO’ decision

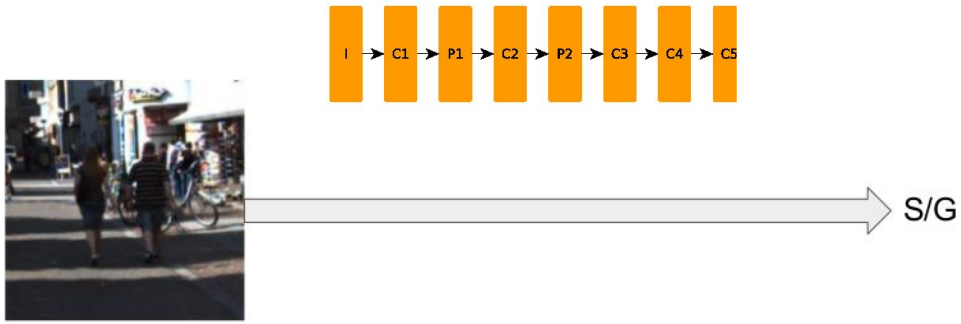
Proposal: Use Proxy Supervision

Ground truth locations

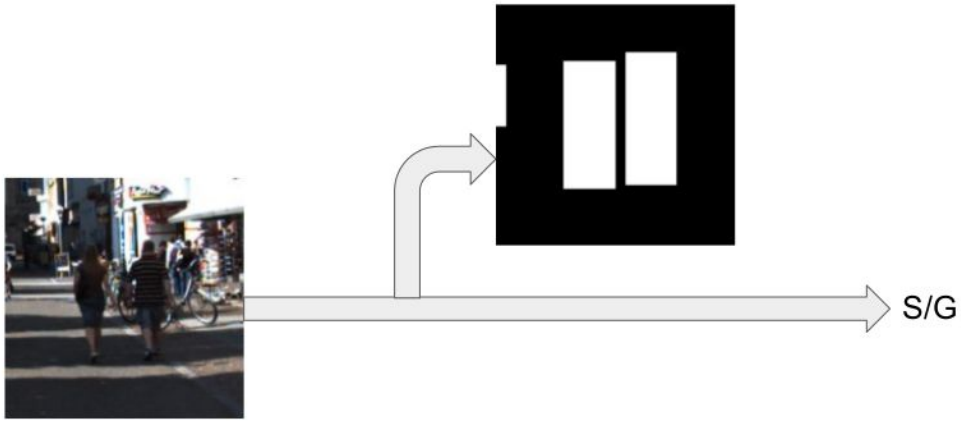
- e.g. of pedestrians

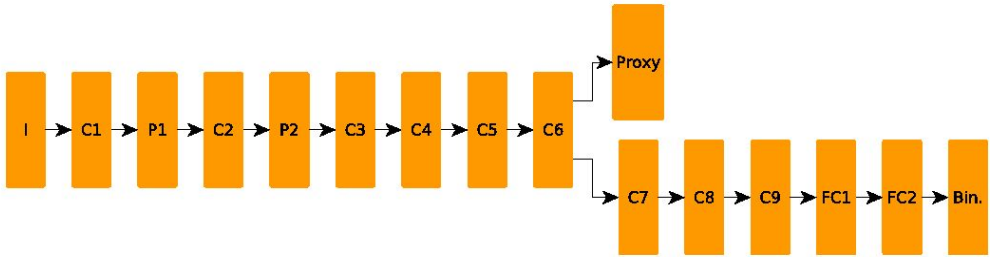





End-to-end decision: Input image -> command

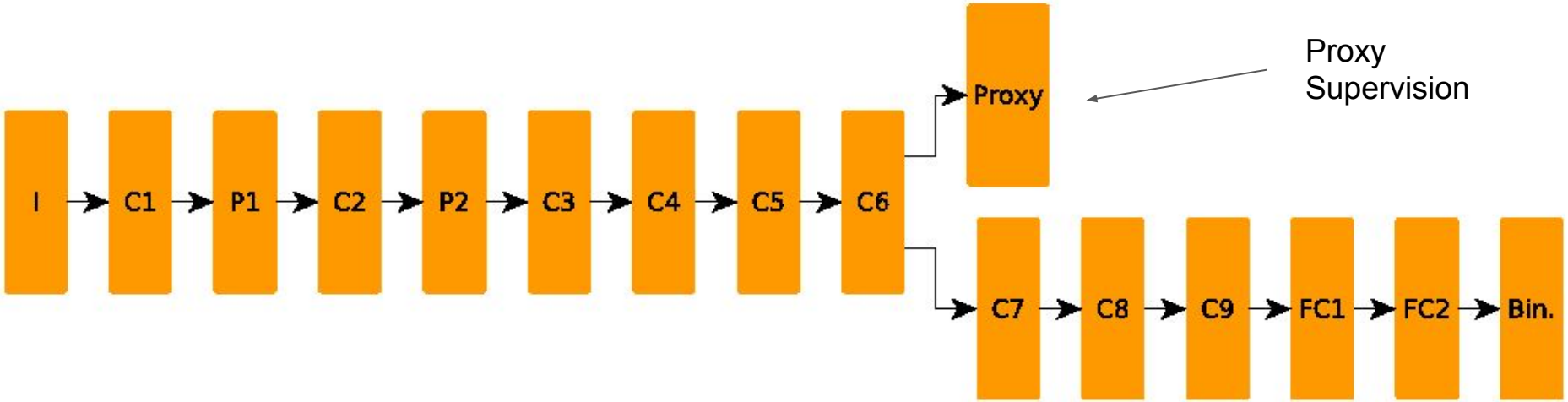


End-to-end decision: Input image -> command + Proxy





Architecture: Feedforward deep neural net



$min_w \; C_{binary}(w) + \lambda \, C_{proxy}(w), \quad \text{where } \lambda = 10^{\wedge -5}$

$C_{proxy}(w)$ = Pixel-wise cross entropy
$C_{binary}(w)$ = Binary cross entropy for s/g

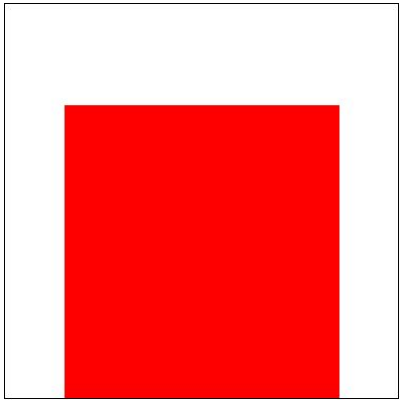
Experimental evaluation

Direct criterion

The Stop/Go decision is a ***nonlinear function*** of the provided bounding box information.

Image

Person bounding box



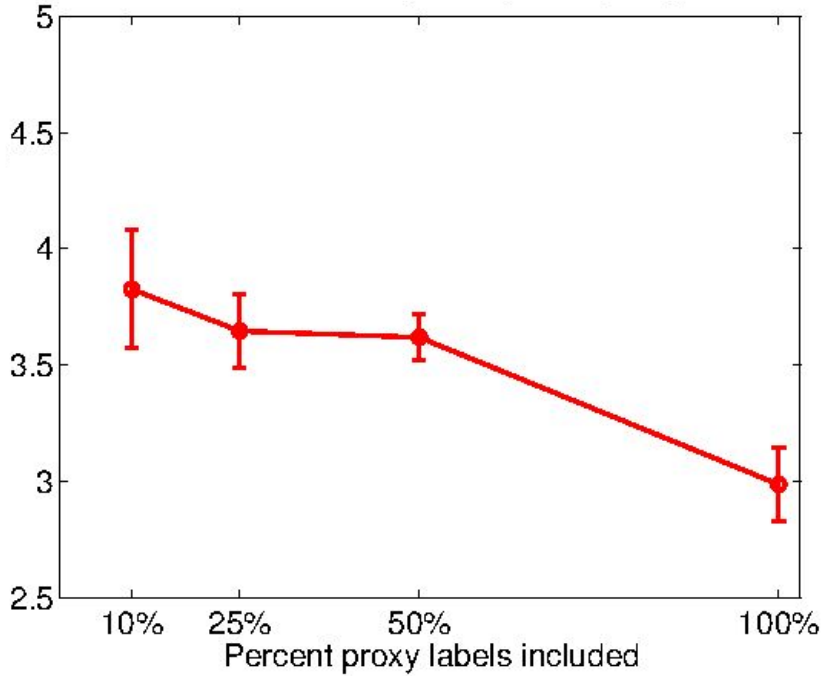
Dangerous zone mask

This is a ‘STOP’ example.

Criterion: If any bounding box (of person) overlaps with the “Danger zone” mask.

Results: learning with partial supervision

Classification error in %



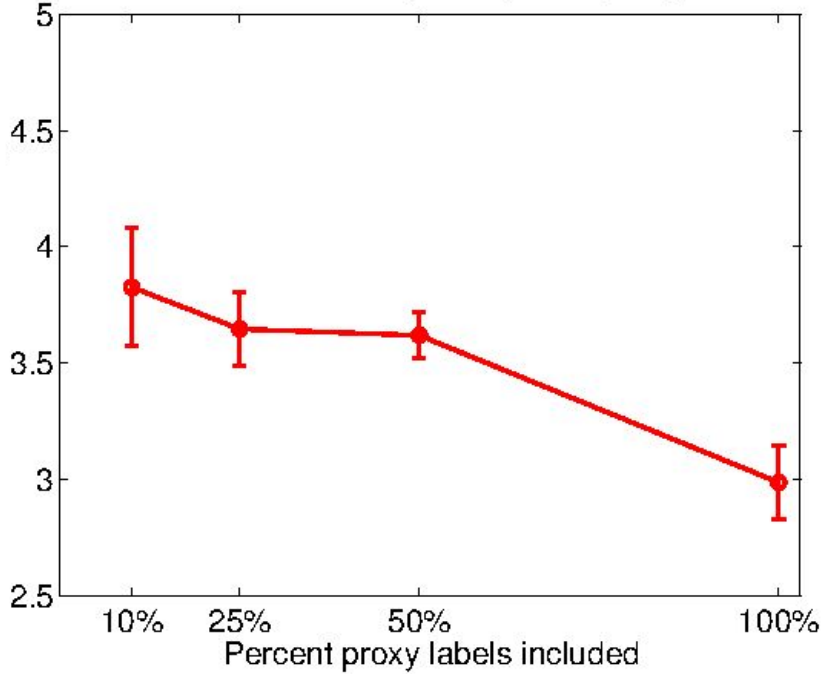
Training data

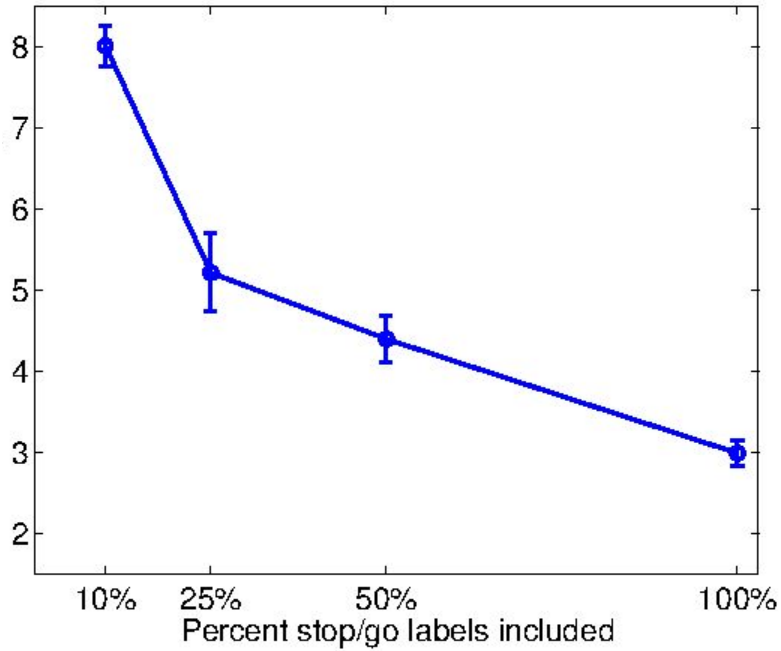
KITTI dataset: Geiger et al.'12
192x192 crops.

Classification error ~22% if no proxy supervision

Results: learning with partial supervision/labeling

Classification error in %





Even partial supervision (e.g. 10% labels) is helpful!

Results: What if the model is pre-trained?

Comparison to a pre-trained model. Classification error (in %)

	No Proxy	With Proxy
No pretraining	22.0 %	2.98 %
With pretraining	3.63 %	2.82 %

Proxy supervision helps more than pre-training.

Proxy supervision helps in addition to pre-training too.

Results: What if proxy supervision is provided by another category?

E.g. bounding boxes for cars

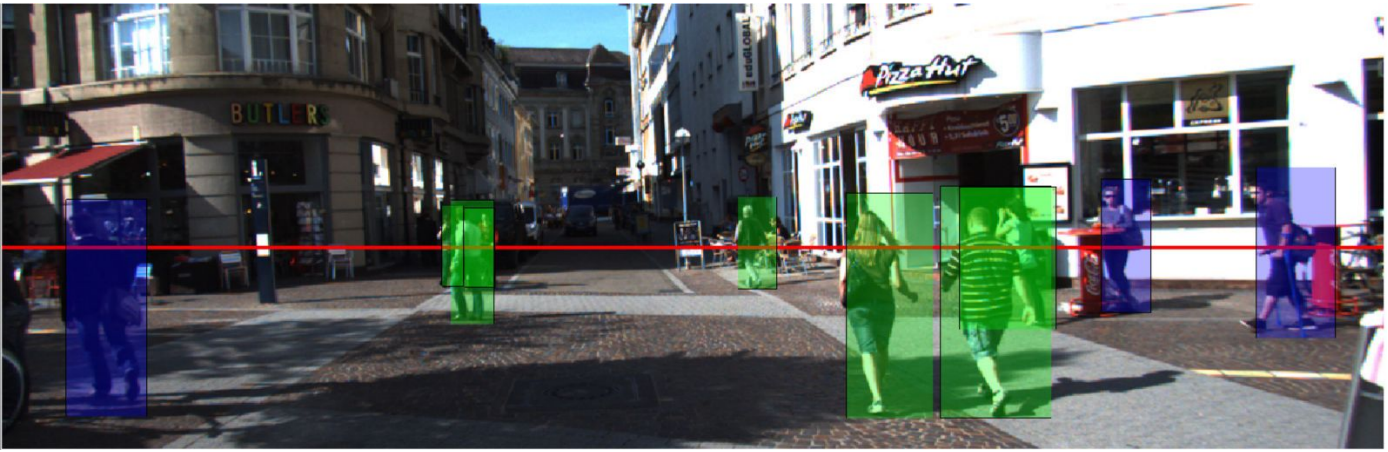
Classification error in %, with pretraining

No Proxy	Car proxy	Pedestrian proxy
3.63 %	3.22 %	2.82 %

Car proxy supervision helps a binary end-to-end decision for persons!

Manually labeled scenes: Labeling Tool

Initial, ground truth boxes.



Labeling:
The persons that may cause a stop decision.

Results on manually labeled data

Results on human labeled data. Full scene/image input. With pre-training.

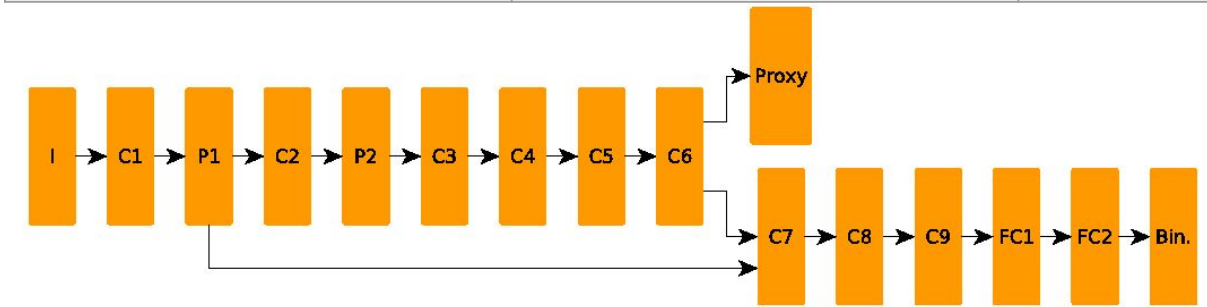
Classification error (in %)

	No proxy	With proxy
Main architecture	6.37 %	5.49 %

Results with a skip connection

Results on human labeled data. Full scene/image input. Classification error (in %)

	No proxy (w. pretr)	With proxy (w. pretr)
Main architecture + Skip	3.19 %	2.88 %



Skip connection

Conclusions

Here: Learning in end-to-end?

This paper:
HOW?

Main takeaway: If you need to learn a complex decision...

...and the input sensors are unfamiliar, unknown:

Learning direct complex decisions is better **with proxy supervision**.

Key findings:

- Proxy supervision **is helpful,** (even when pretraining)
- Proxy supervision of a **non-related task** is helpful.
- Even **partial proxy** supervision is beneficial.

Future work

Here: Automatic criterion and human assessed criterion.

Future: Use actual drives with getting feedback in next frames (did the car stop several frames ahead?).

Here: Stop/Go for persons.

Future: Generalize to many causes for stopping.

Supervision with respect to many other helpful auxiliary tasks, e.g. traffic signs, lights.

THANK YOU!