Other Methods and Applications of Deep Learning

Yann Le Cun
The Courant Institute of Mathematical Sciences
New York University
http://yann.lecun.com

Denoising Auto-Encoders

- [Vincent & Bengio, ICML 2008]
- Idea: feed a “noisy” (corrupted) input to an auto-encoder, and train it to produce the uncorrupted version.
- Use the states of the hidden layer as features
- Stack multiple layers
- Very simple and effective technique!

Another way to Learn Deep Invariant Features: DrLIM

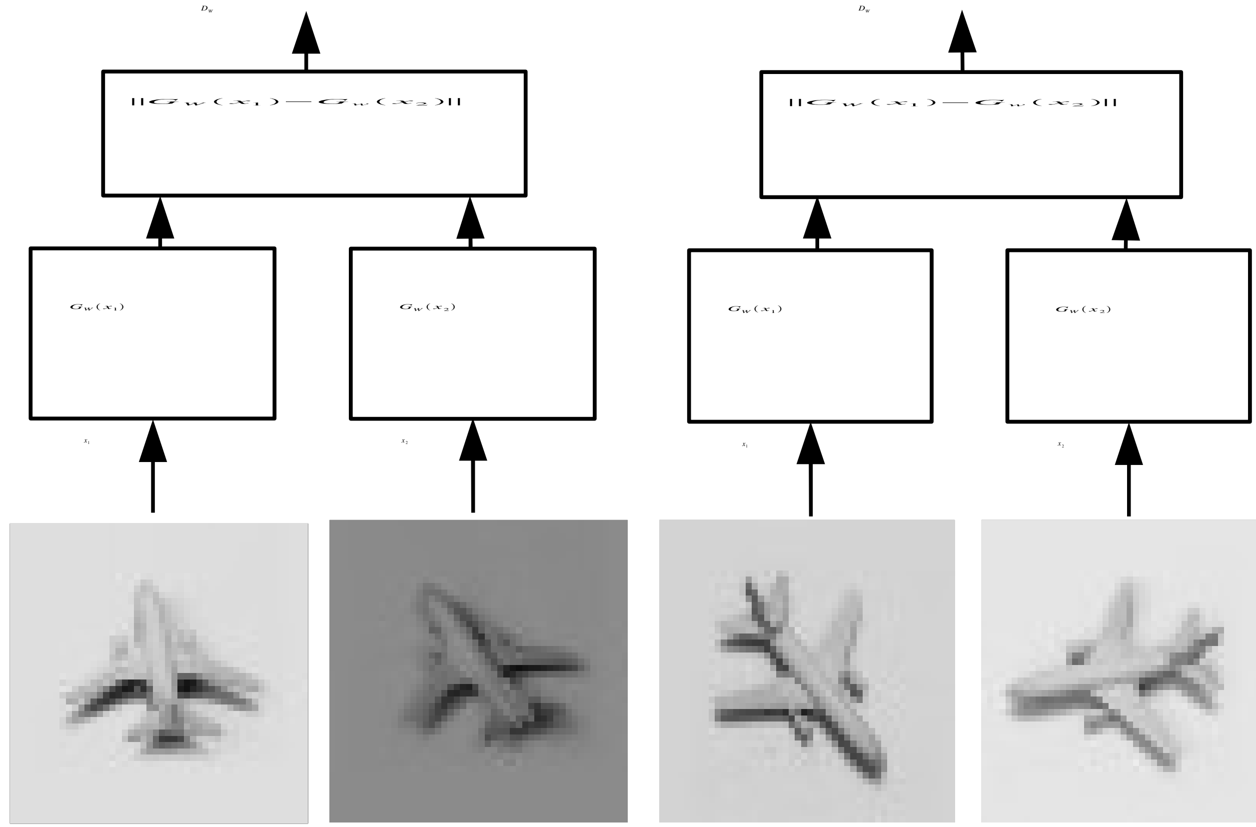
Hadsell, Chopra, LeCun CVPR 06], also [Weston & Collobert ICML 08 for language models]

Loss function:

- Outputs corresponding to input samples that are neighbors in the neigborhood graph should be nearby
- Outputs for input samples that are not neighbors should be far away from each other

Make this small

Make this large



Similar images (neighbors in the neighborhood graph)

Dissimilar images (non-neighbors in the neighborhood graph)

Application of Stacked Auto-Encoders to Text Retrieval

- **Ranzato et al. ICML 08**

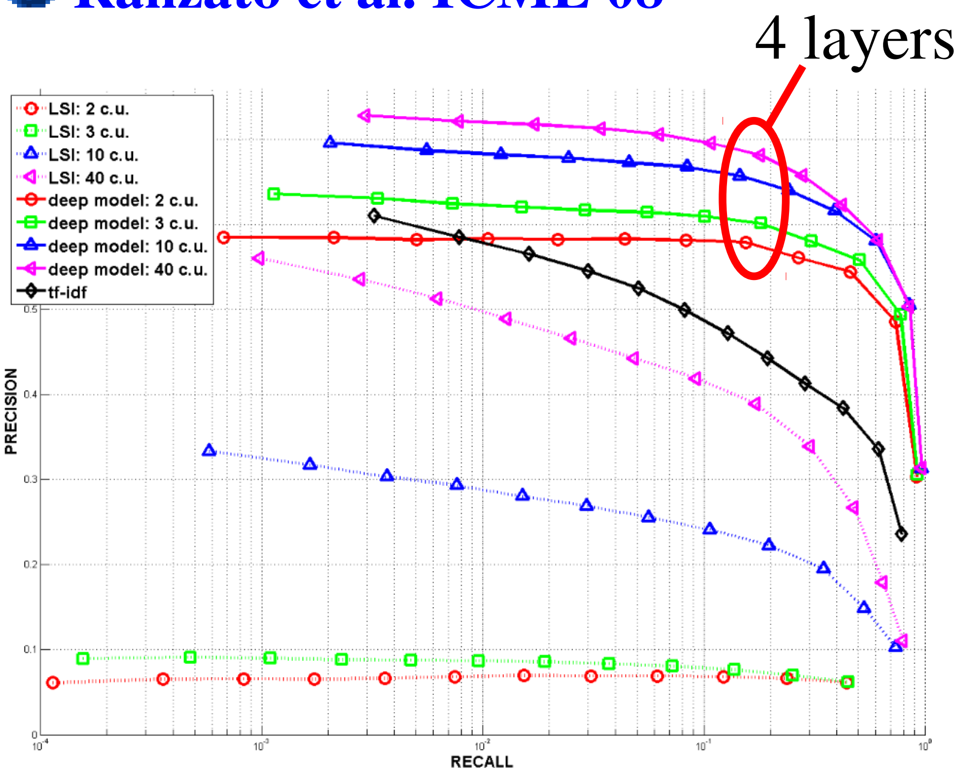


Figure 4. Precision-recall curves of the Reuters dataset comparing a linear model (LSI) to the non-linear deep model with the same number of code units (c.u.). Retrieval is done using the k most similar documents according to cosine similarity, with $k \in [1 \ldots 2047]$.

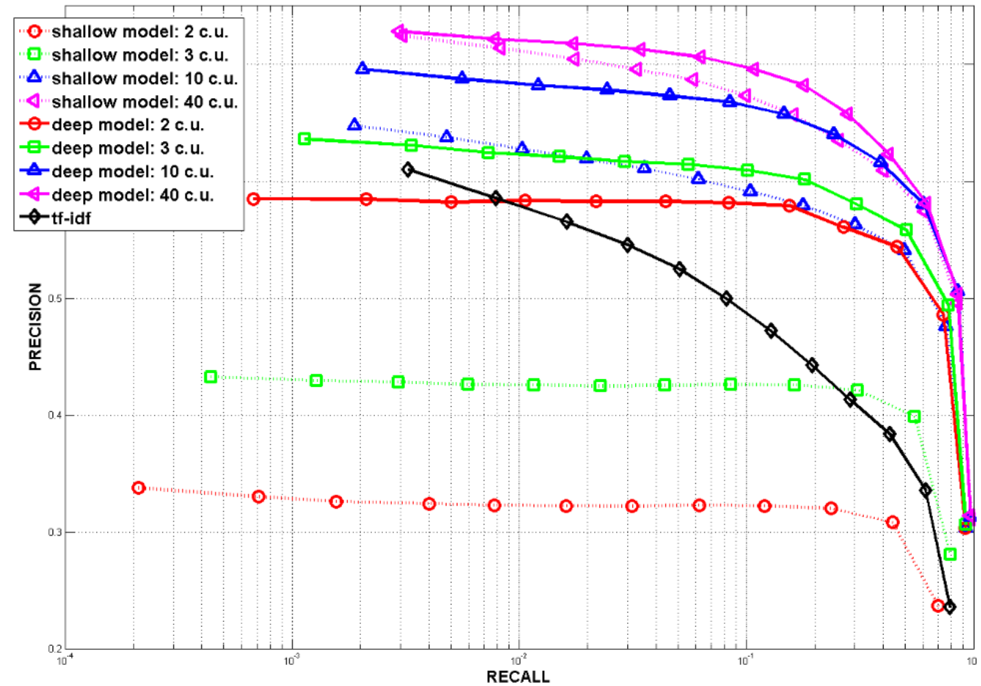


Figure 5. Precision-recall curves of the Reuters dataset comparing the model trained with only one layer (shallow architecture) to a deep model with the same number of code units. The deep model outperforms the shallow one overall when the features are extremely compact.

Application of Stacked Auto-Encoders to Text Retrieval

- **Ranzato et al. ICML 08**

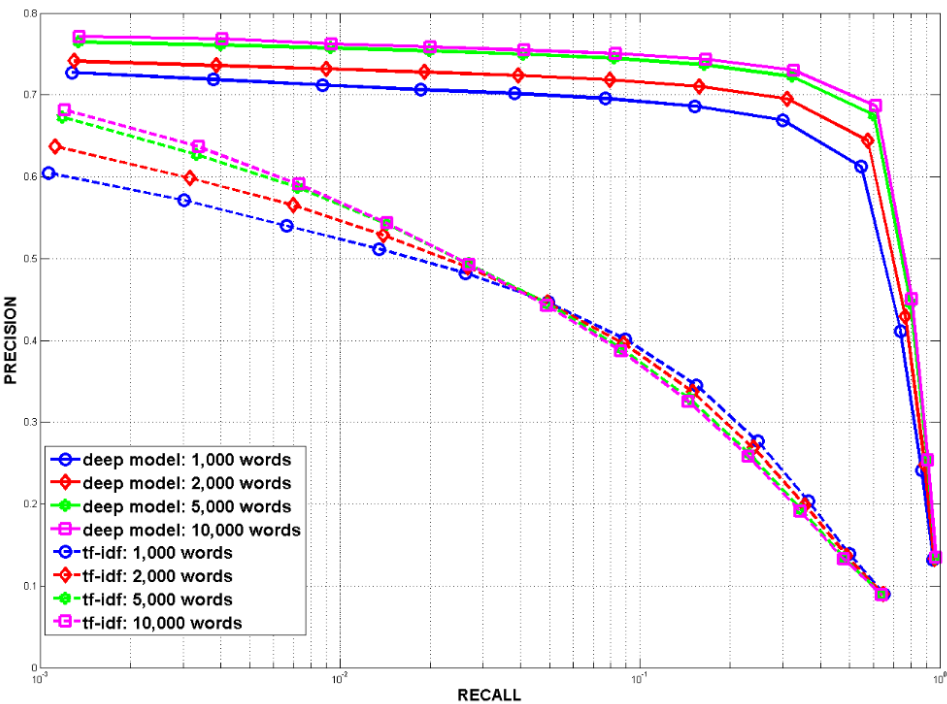


Figure 6. Precision-recall curves of the 20 Newsgroups dataset comparing the performance of the model (1 layer) trained on documents with various number of words in the dictionary (from 1000 to 10000).

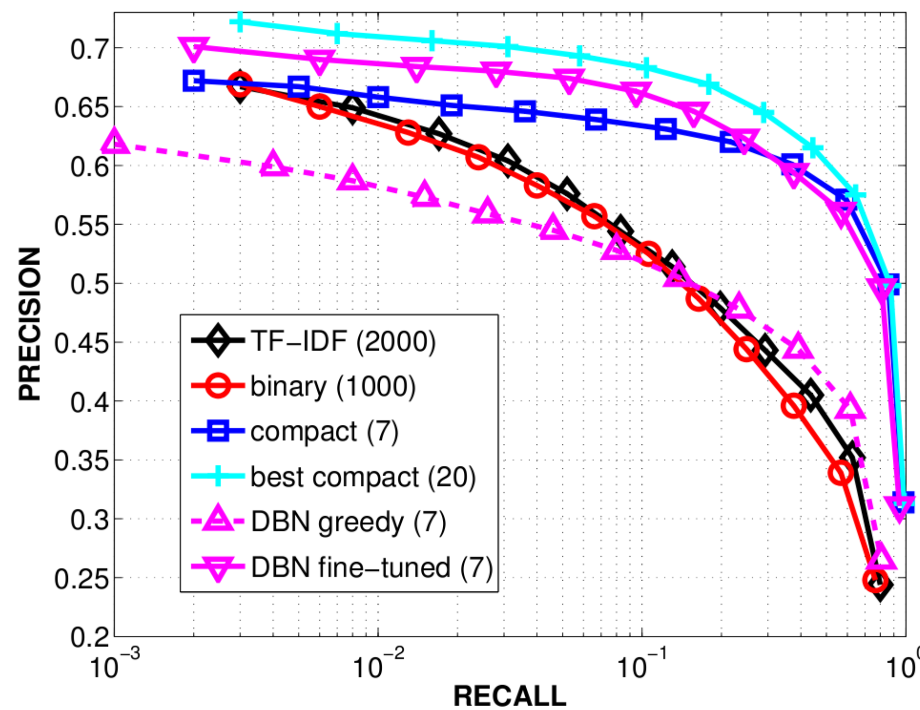


Figure 7. Precision-recall curves using very compact representations and high dimensional binary representations. Compact representations can achieve better performance using less memory and CPU time.

Application of Stacked Auto-Encoders to Text Retrieval

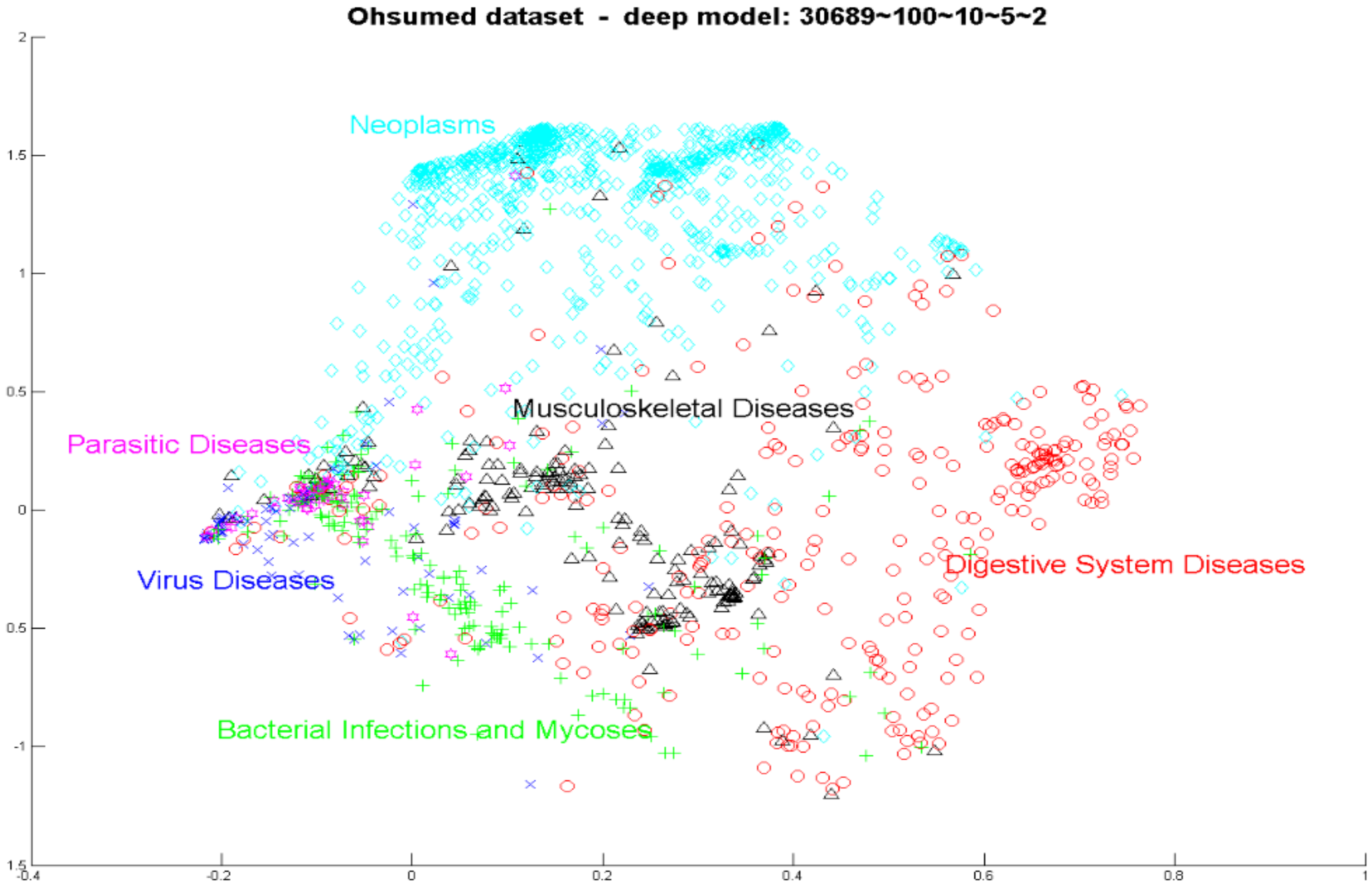


Figure 8. The two-dimensional codes produced by the deep model trained on the Ohsumed dataset (shown only the 6 most numerous classes). The codes have been computed by propagating test documents through the 4-layer network.

Learning Codes for Natural Language Processing

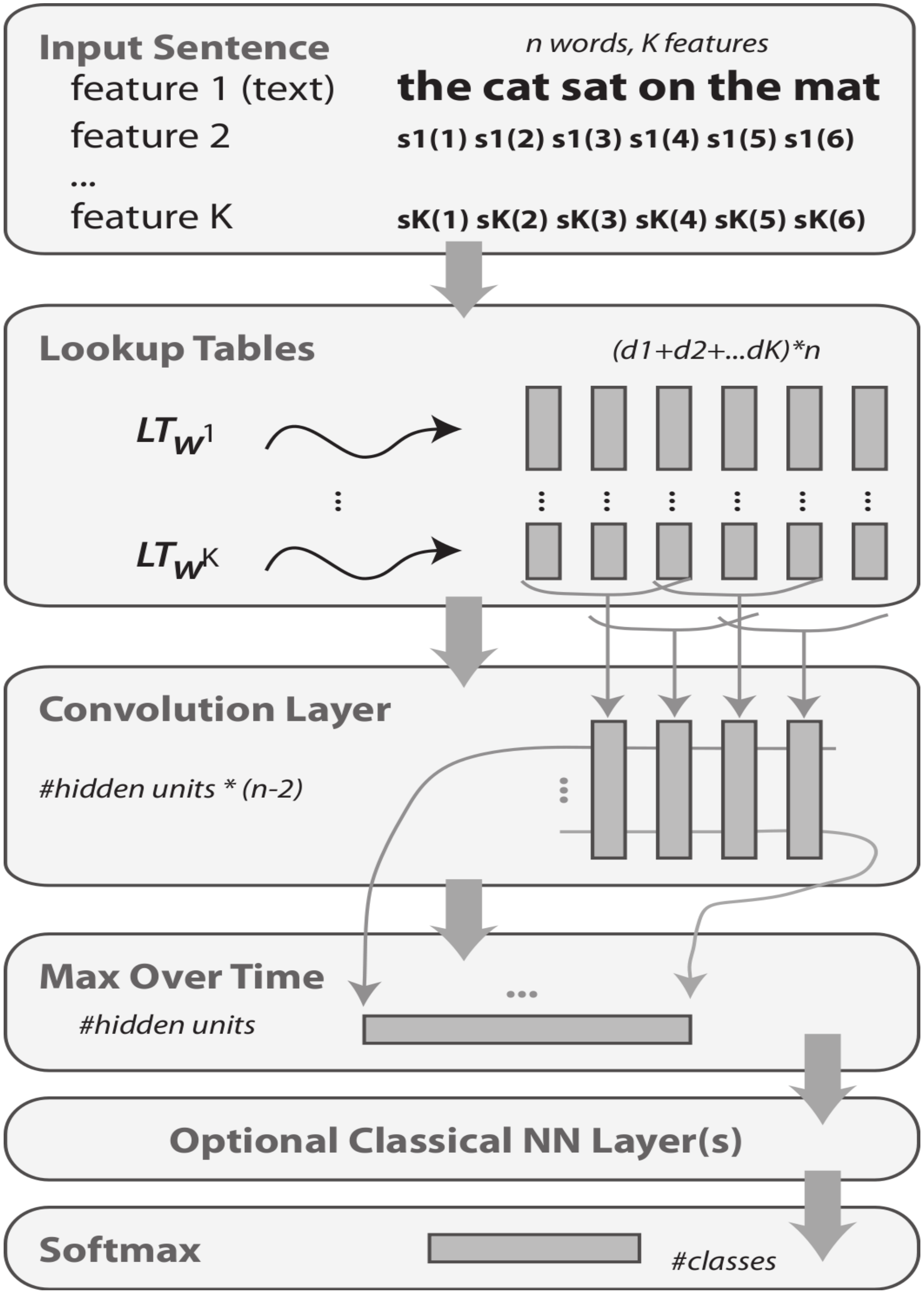
[Collobert & Weston ICML 2008, ACL 2008]

- **1D convolutional networks. Input is window of 11 words on a text, output is a single unit.**
  - Input is 1-of-N code, where N is the size of the lexicon
- **Positive examples come Wikipedia text**
- **Negative examples are generated by substituting the middle word by another random word**
- **The network is trained to produce 0 for positive examples and 1 for negative examples**
- **The first layer learns "semantic-syntactic codes" for all words**
- **The codes are used as input representation for various NLP tasks**

Learning Codes for NLP

[Collobert & Weston ICML 2008, ACL 2008]

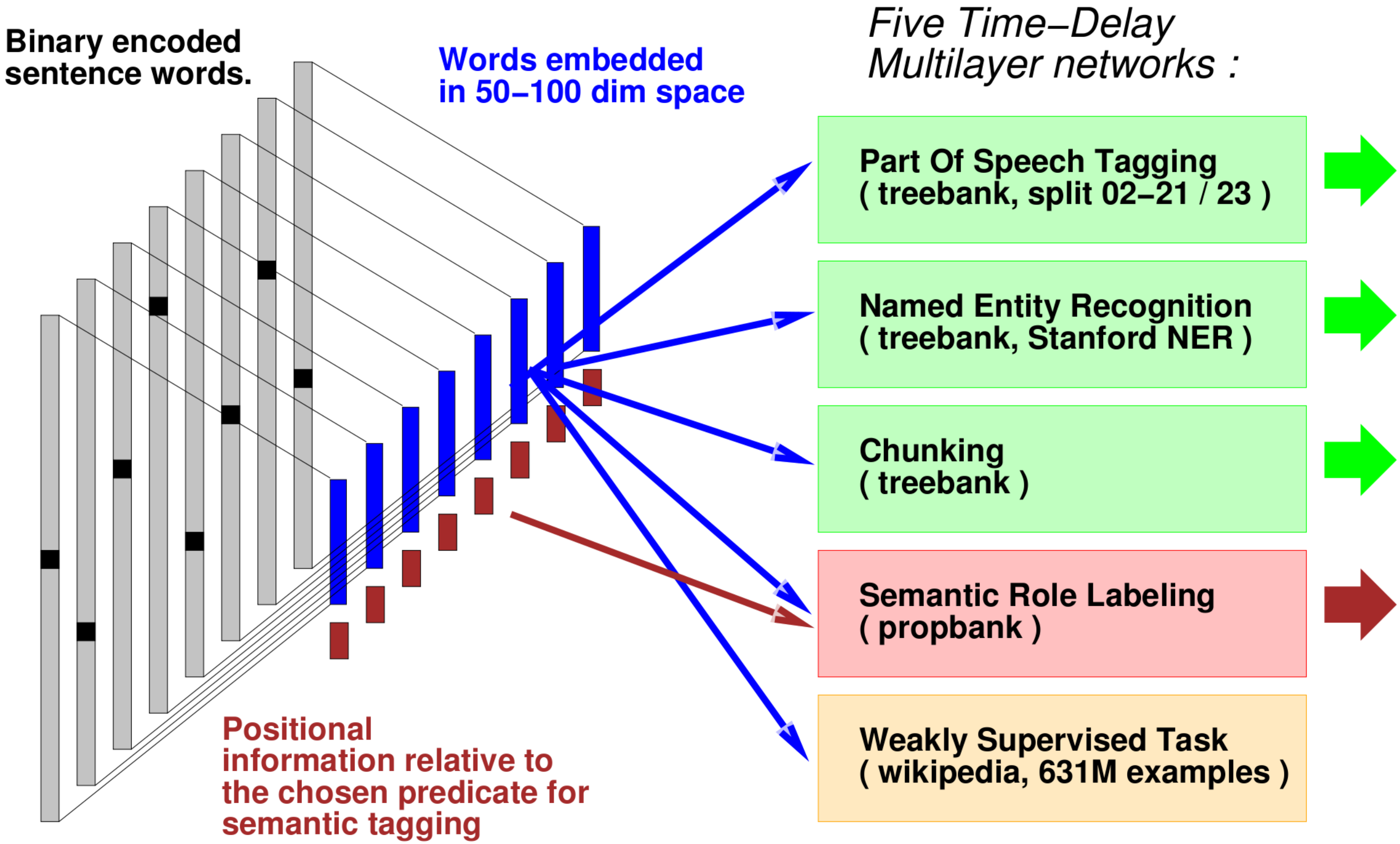
- **Convnet Architecture**



Learning Codes for Natural Language Processing

[Collobert & Weston ICML 2008, ACL 2008]

- Convnet on word window



Learning Codes for Natural Language Processing

[Collobert & Weston ICML 2008, ACL 2008]

Performance on various NLP tasks

Table 2.4: Deep NN architecture with no hand-engineered *features (with or without multitasking) vs. top systems that use hand-engineered features. The top systems for POS is [Toutanova et al., 2003], for CHUNK is [Ando and Zhang, 2005b], for NER is [Florian et al., 2003], and for SRL is [Punyakanok et al., 2004].*

Method	**POS** (% Err)	**CHUNK** (F1)	**NER** (F1)	**SRL** (% Err)
Top Systems	2.76	94.20	88.76	13.36
NN	3.15	88.82	81.61	16.40
NN + Multitask	2.78	94.18	88.88	13.82

Learning Codes for Natural Language Processing

[Collobert & Weston ICML 2008, ACL 2008]

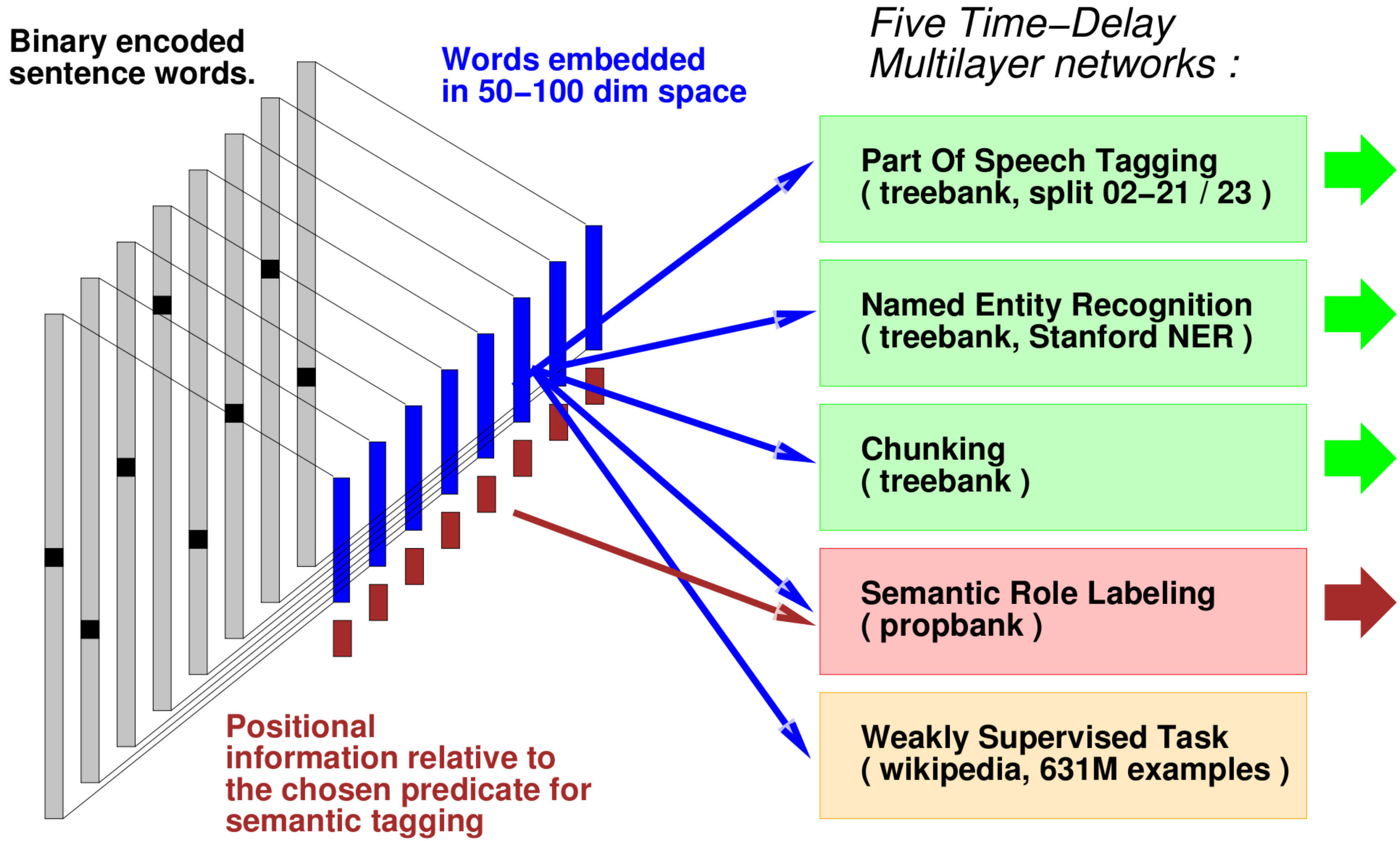
- **Nearest neighbor words to a given word in the feature space**

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	VLADIMIR	NYU
SPAIN	CHRIST	PLAYSTATION	YELLOWISH	SMASHED	VIKTOR	MSU
ITALY	GOD	DREAMCAST	GREENISH	RIPPED	ALEKSANDR	CALTECH
RUSSIA	RESURRECTION	PS2	BROWNISH	BRUSHED	MIKHAIL	BERKLEE
POLAND	PRAYER	SNES	BLUISH	HURLED	ALFRED	JUILLARD
ENGLAND	YAHWEH	WII	CREAMY	GRABBED	NIKOLAI	UCLA
DENMARK	JOSEPHUS	NES	WHITISH	TOSSED	OSKAR	VASSAR
GERMANY	MOSES	NINTENDO	BLACKISH	SQUEEZED	JOSEF	CLAREMONT
PORTUGAL	SIN	GAMECUBE	SILVERY	BLASTED	ANDREI	BYU
SWEDEN	HEAVEN	PSP	GREYISH	TANGLED	GIUSEPPE	USC
AUSTRIA	SALVATION	AMIGA	PALER	SLASHED	PIETRO	LSU

Learning Codes for Natural Language Processing

[Collobert & Weston ICML 2008, ACL 2008]

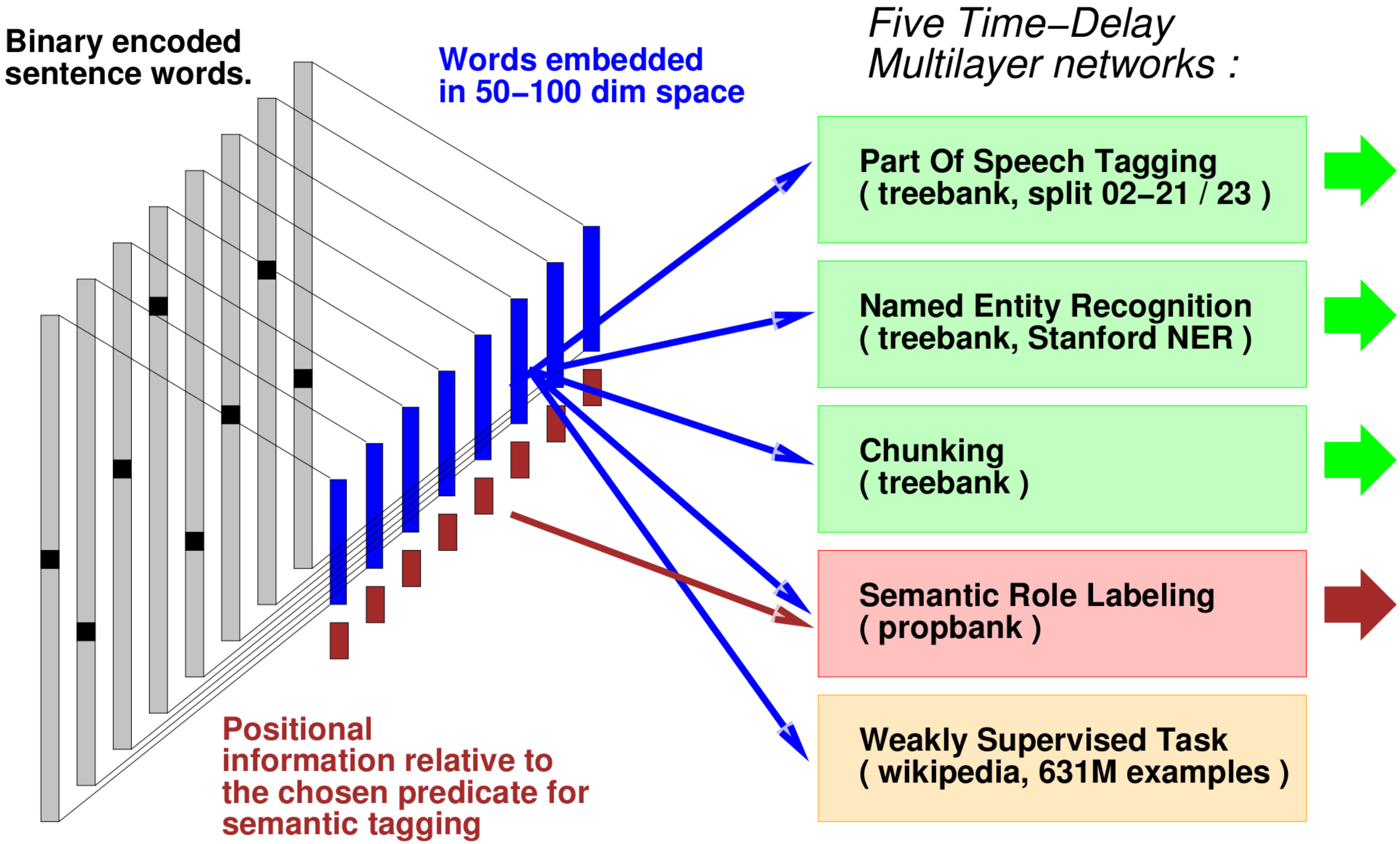
- **Convnet on word window**



Learning Codes for Natural Language Processing

[Collobert & Weston ICML 2008, ACL 2008]

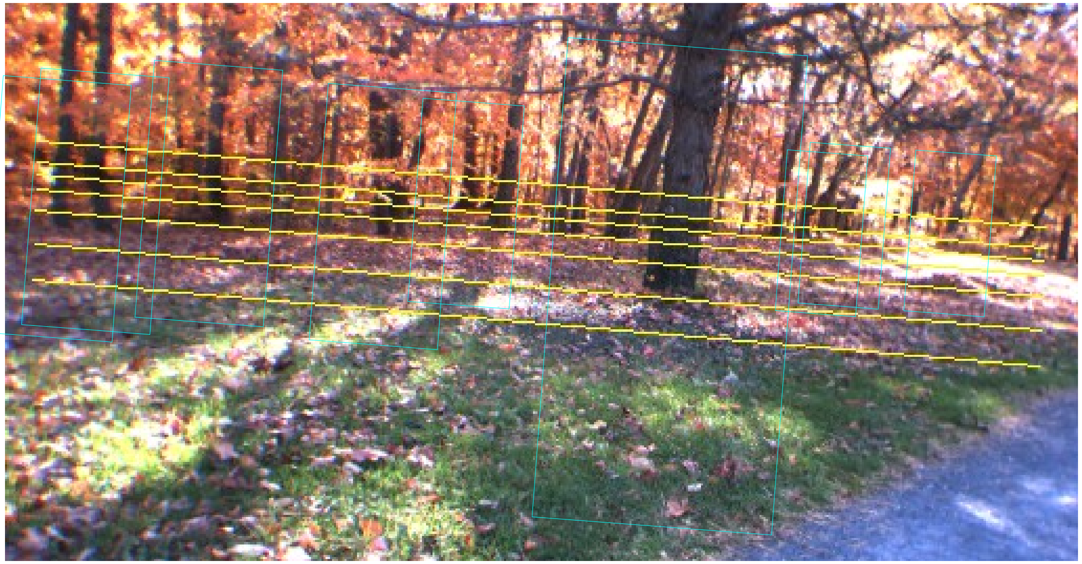
Convnet on word window



DARPA/LAGR: Learning Applied to Ground Robotics

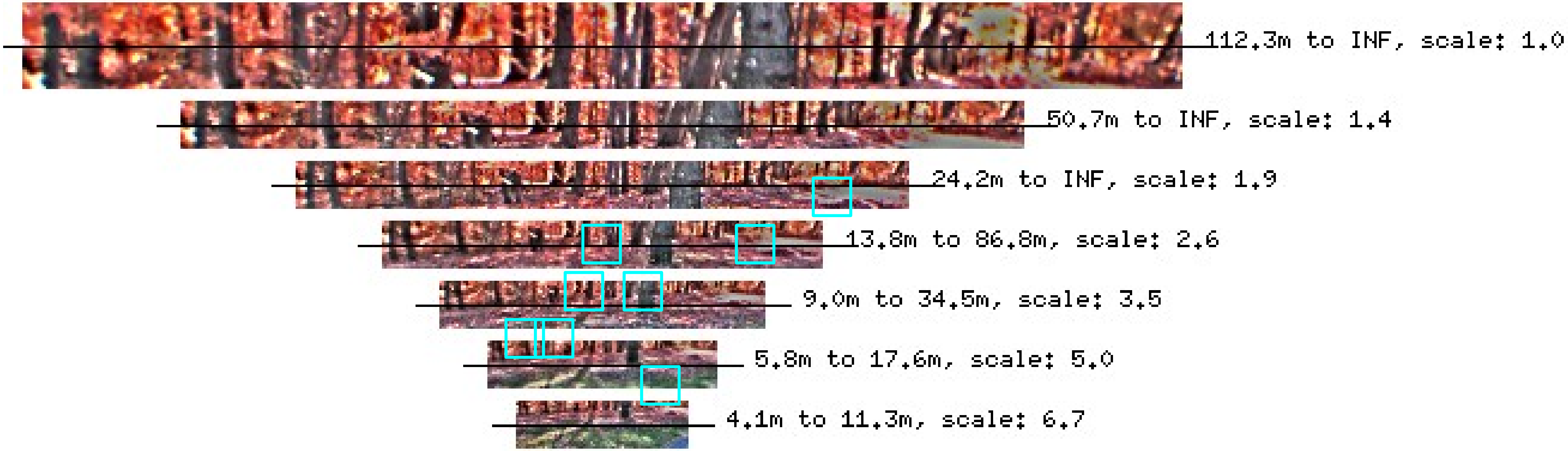
- Getting a robot to drive autonomously in unknown terrain solely from vision (camera input).
- Our team (NYU/Net-Scale Technologies Inc.) was one of 8 participants funded by DARPA
- All teams received identical robots and can only modify the software (not the hardware)
- The robot is given the GPS coordinates of a goal, and must drive to the goal as fast as possible. The terrain is unknown in advance. The robot is run 3 times through the same course.
- Long-Range Obstacle Detection with on-line, self-trained ConvNet
- Uses temporal consistency!

Long Range Vision: Distance Normalization



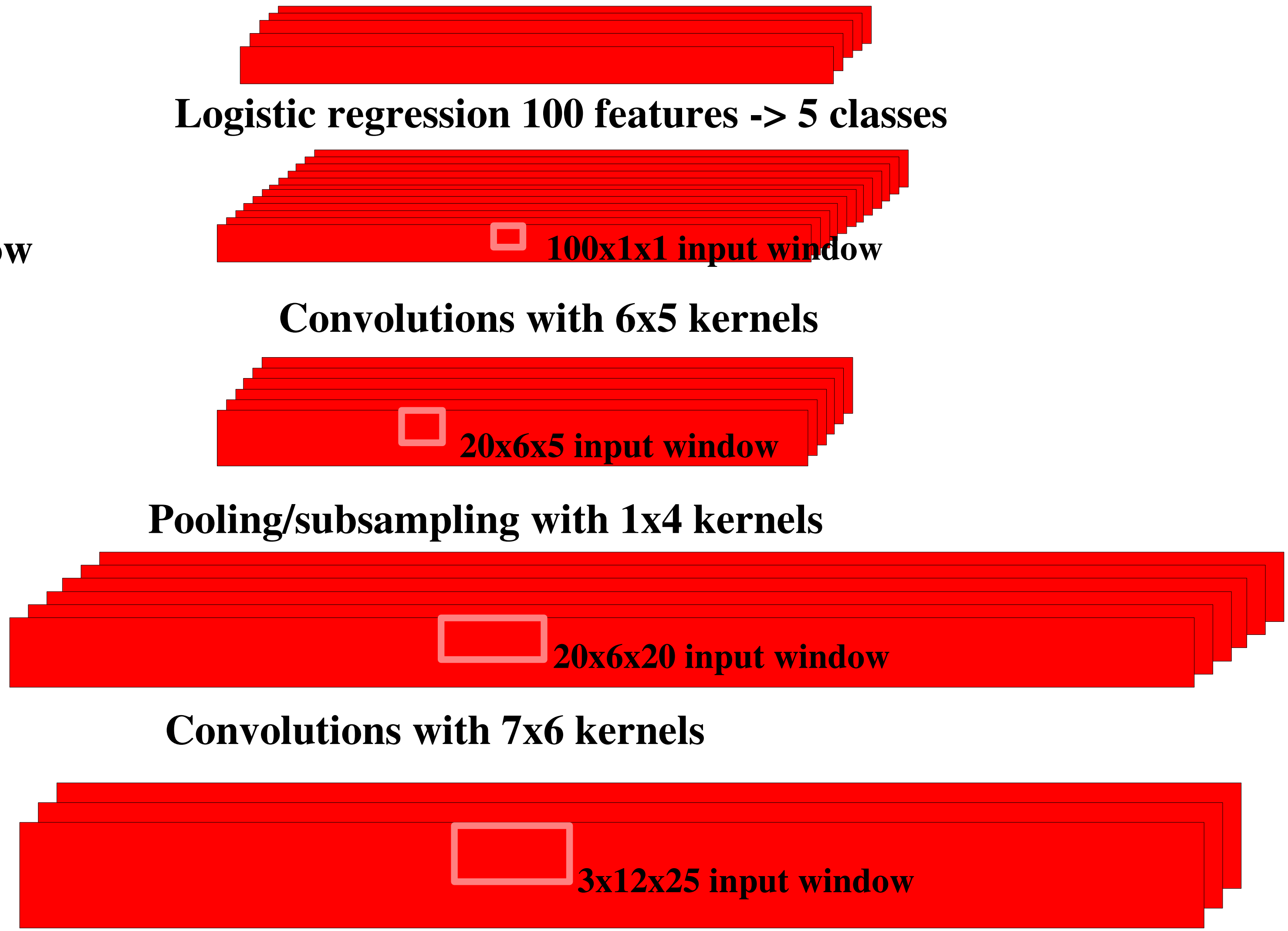
Pre-processing (125 ms)

- Ground plane estimation
- Horizon leveling
- Conversion to YUV + local contrast normalization
- Scale invariant pyramid of distance-normalized image “bands”



Convolutional Net Architecture

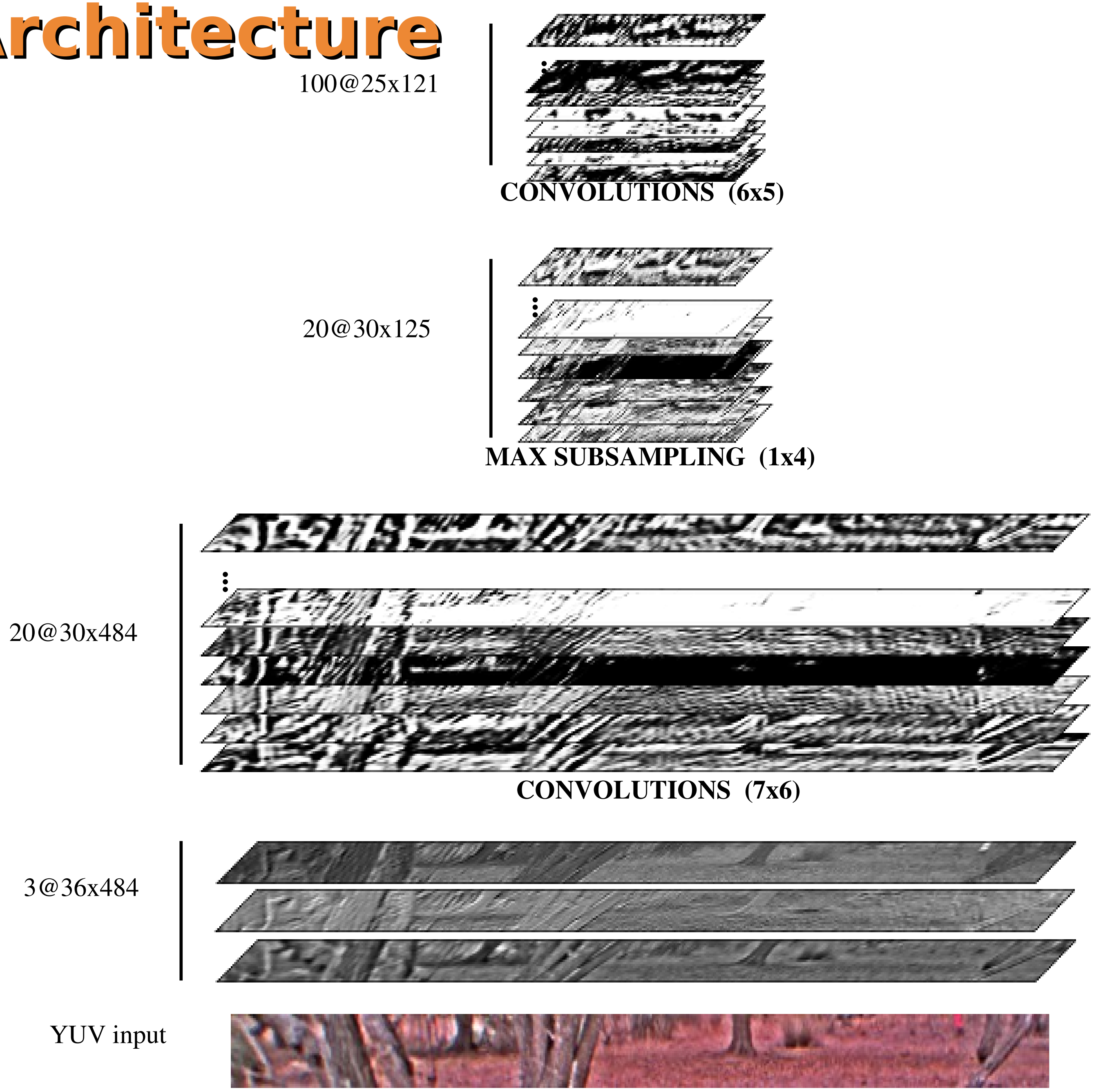
- Operates on 12x25 YUV windows from the pyramid



100 features per
3x12x25 input window

YUV image band
20-36 pixels tall,
36-500 pixels wide

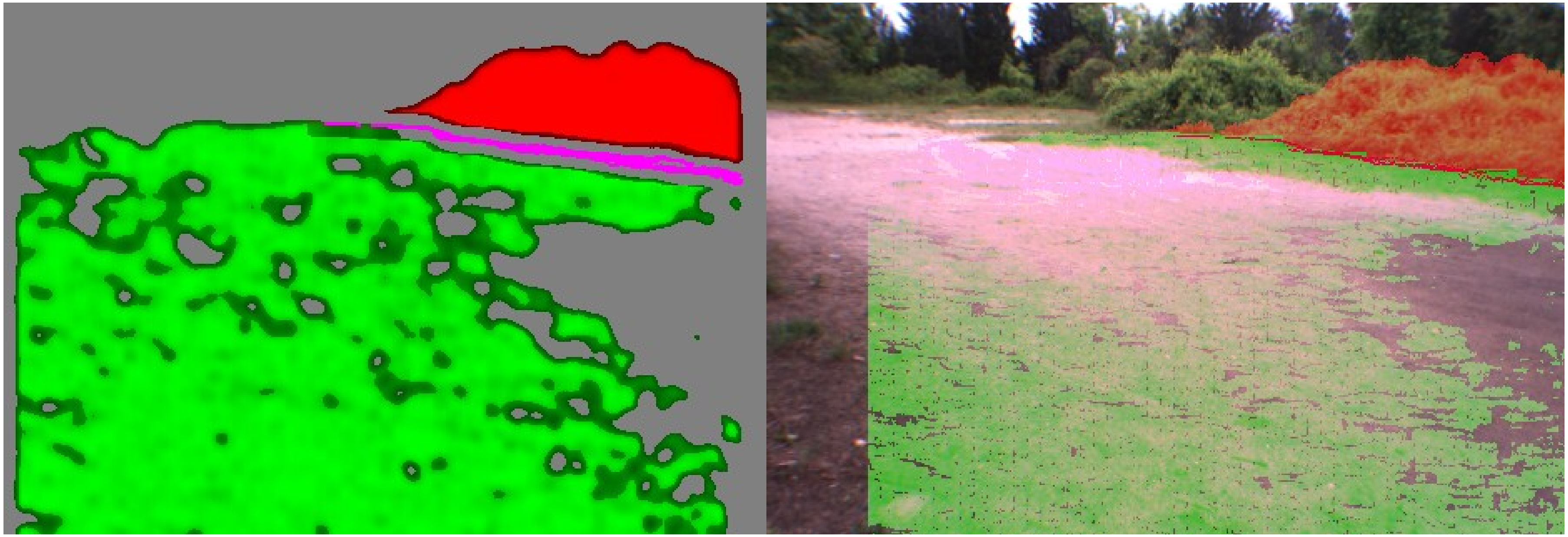
Convolutional Net Architecture



Long Range Vision: 5 categories

Online Learning (52 ms)

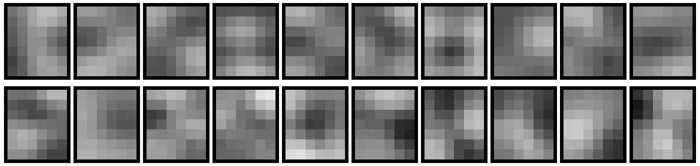
- Label windows using stereo information – 5 classes

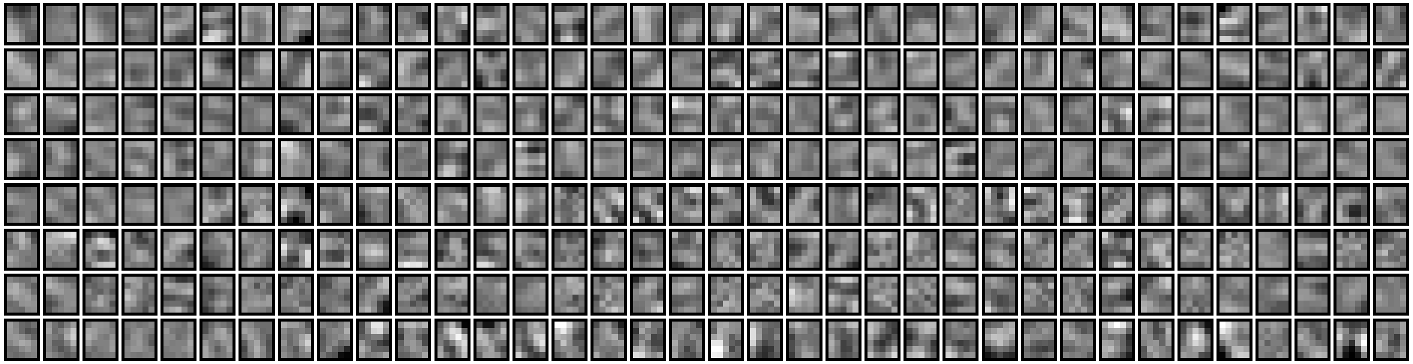


super-ground ground footline obstacle super-obstacle

Trainable Feature Extraction

- "Deep belief net" approach to unsupervised feature learning
- Two stages are trained in sequence
  - each stage has a layer of convolutional filters and a layer of horizontal feature pooling.
  - Naturally shift invariant in the horizontal direction
- Filters of the convolutional net are trained so that the input can be reconstructed from the features
  - 20 filters at the first stage (layers 1 and 2)
  - 300 filters at the second stage (layers 3 and 4)
- Scale invariance comes from pyramid.
  - for near-to-far generalization

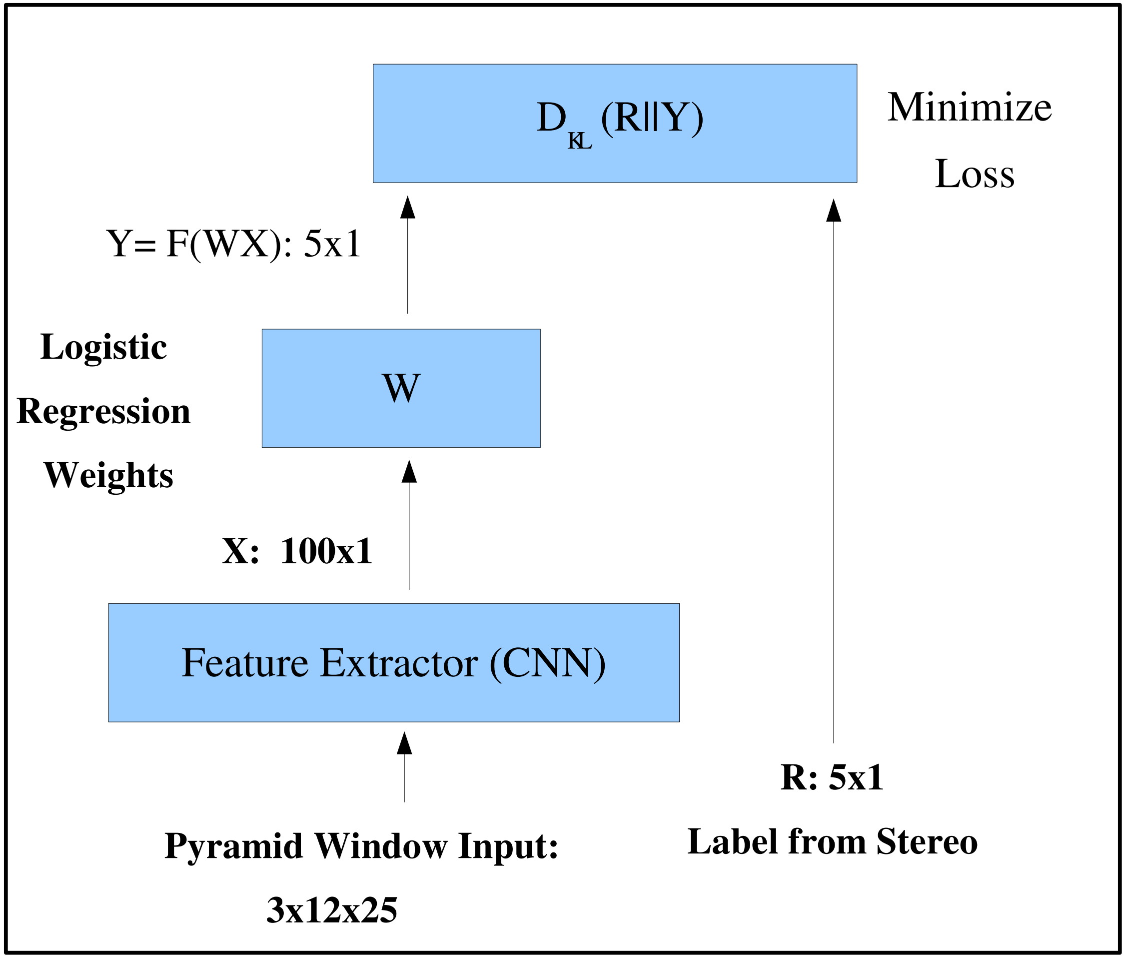




Long Range Vision: the Classifier

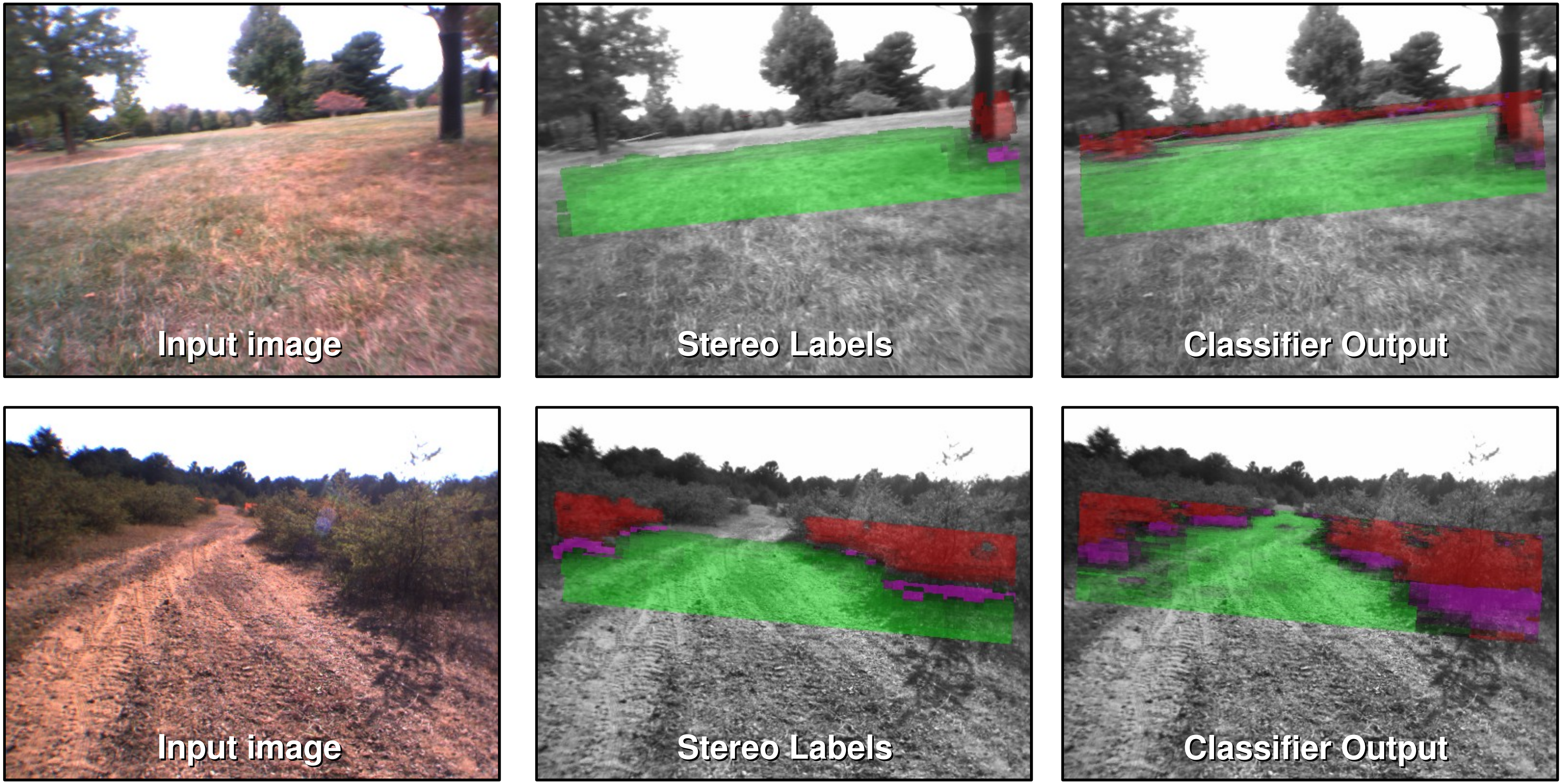
Online Learning (52 ms)

- Train a logistic regression on every frame, with cross entropy loss function

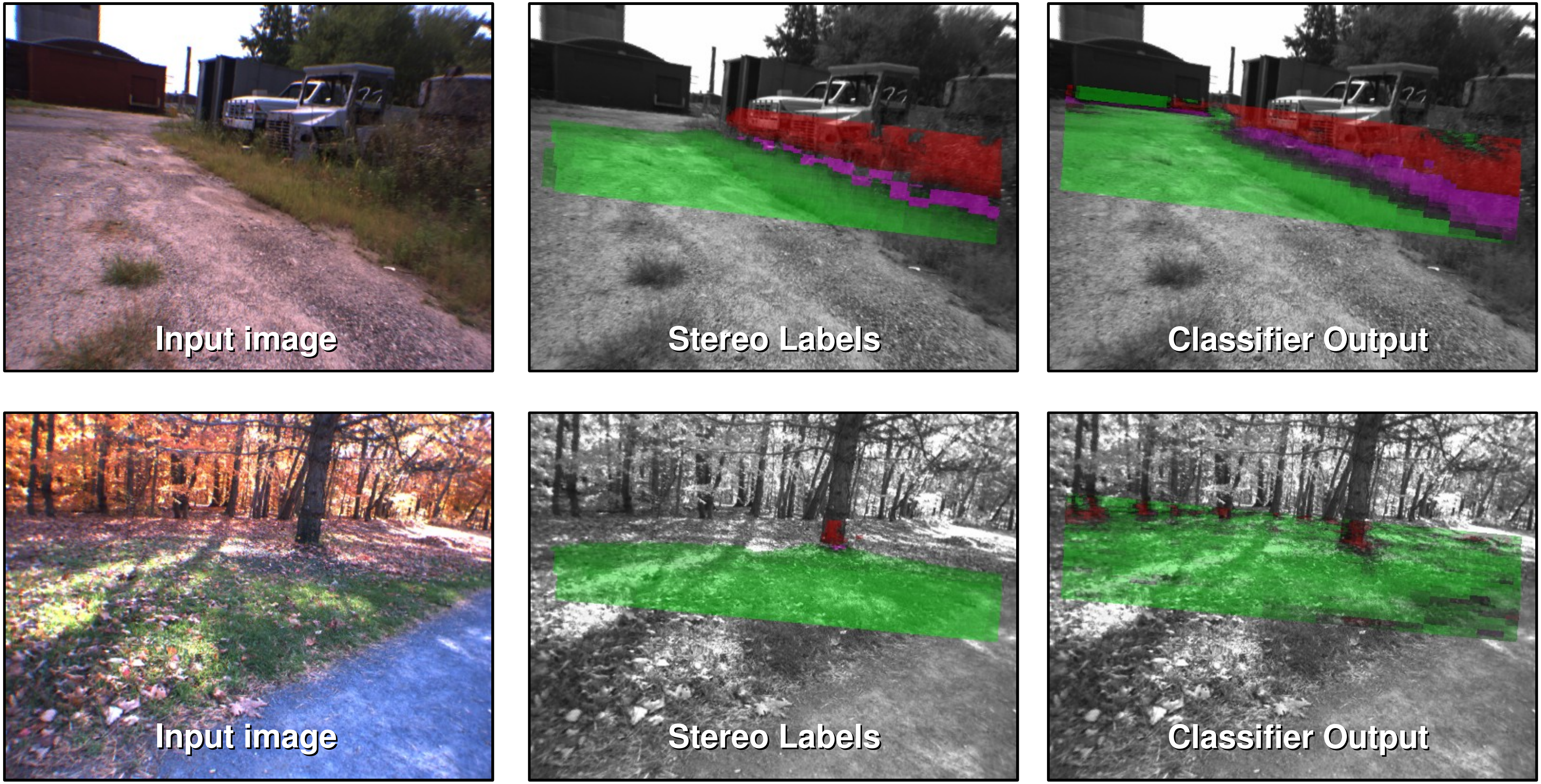


- 5 categories are learned
- 750 samples of each class are kept in a ring buffer: short term memory.
- Learning “snaps” to new environment in about 10 frames
- Weights are trained with stochastic gradient descent
- Regularization by decay to default weights

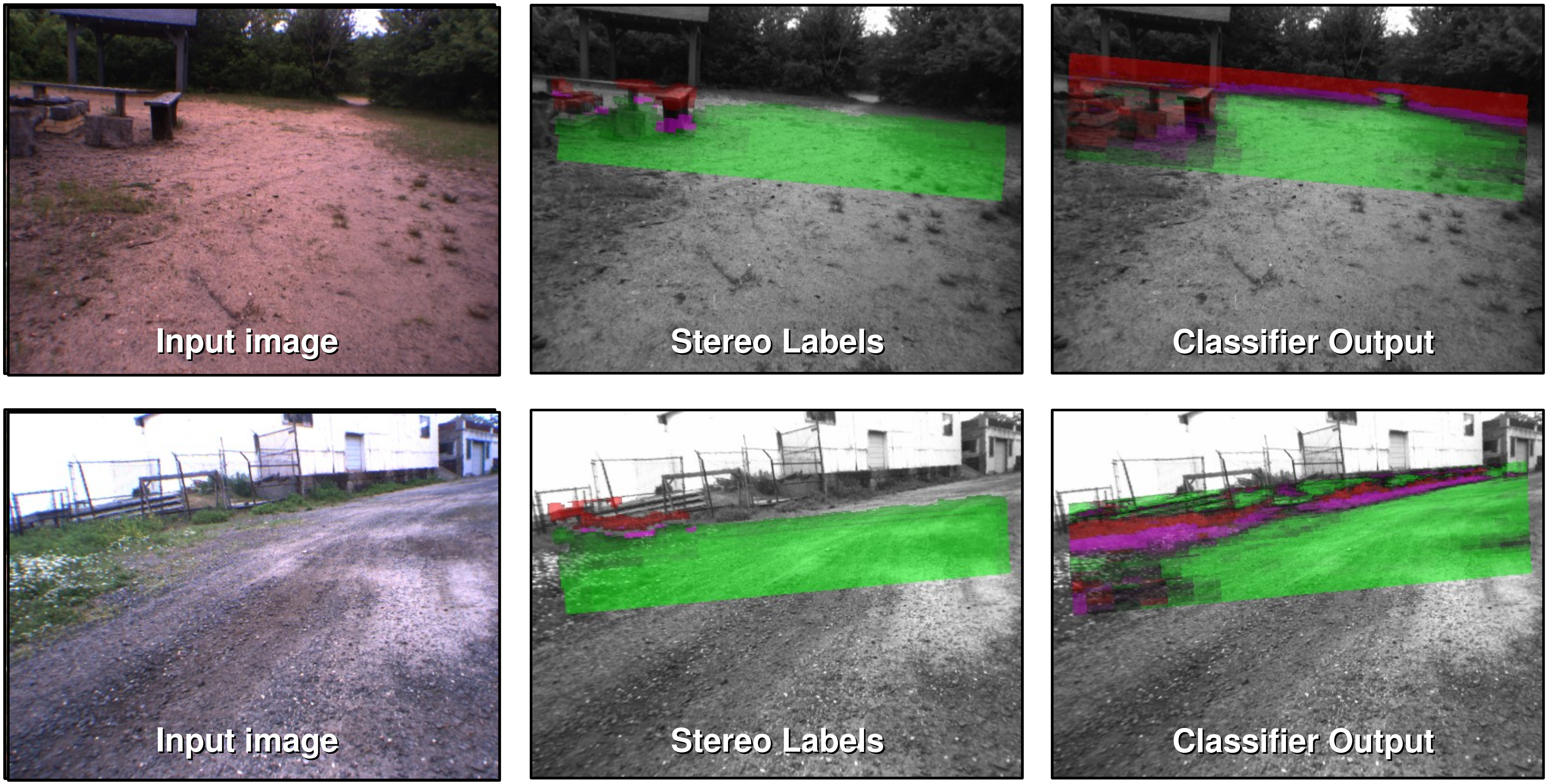
Long Range Vision Results

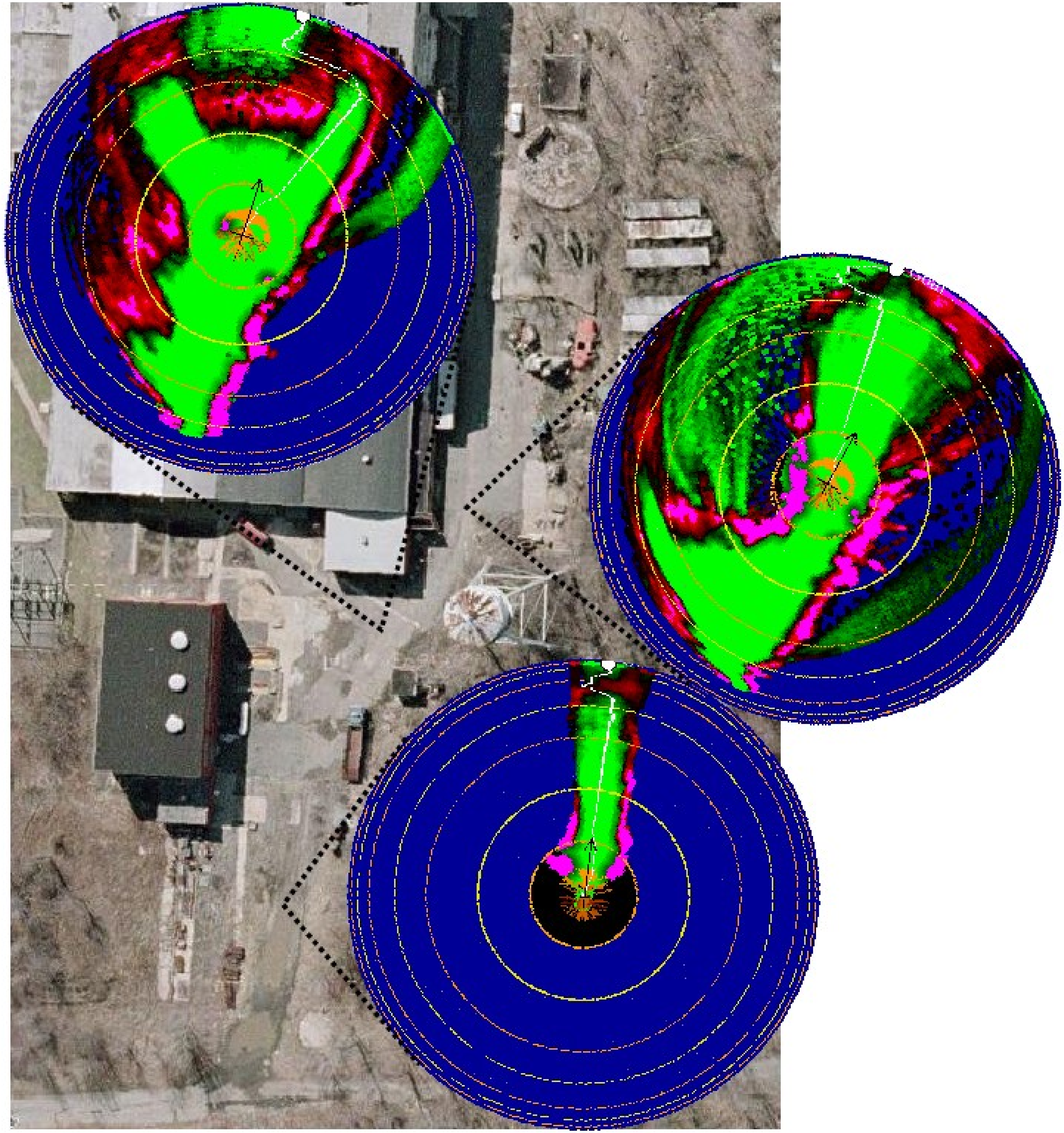


Long Range Vision Results



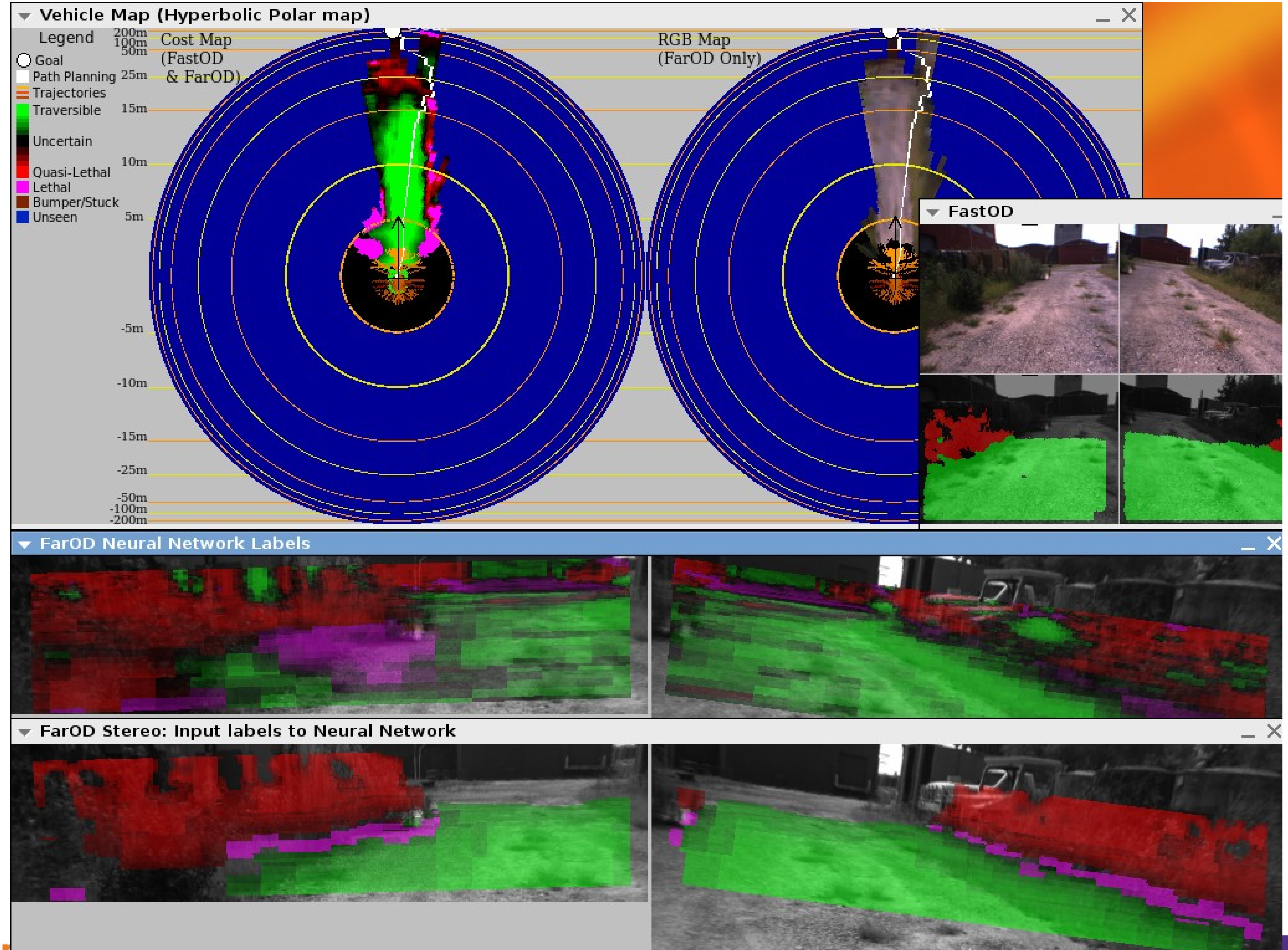
Long Range Vision Results



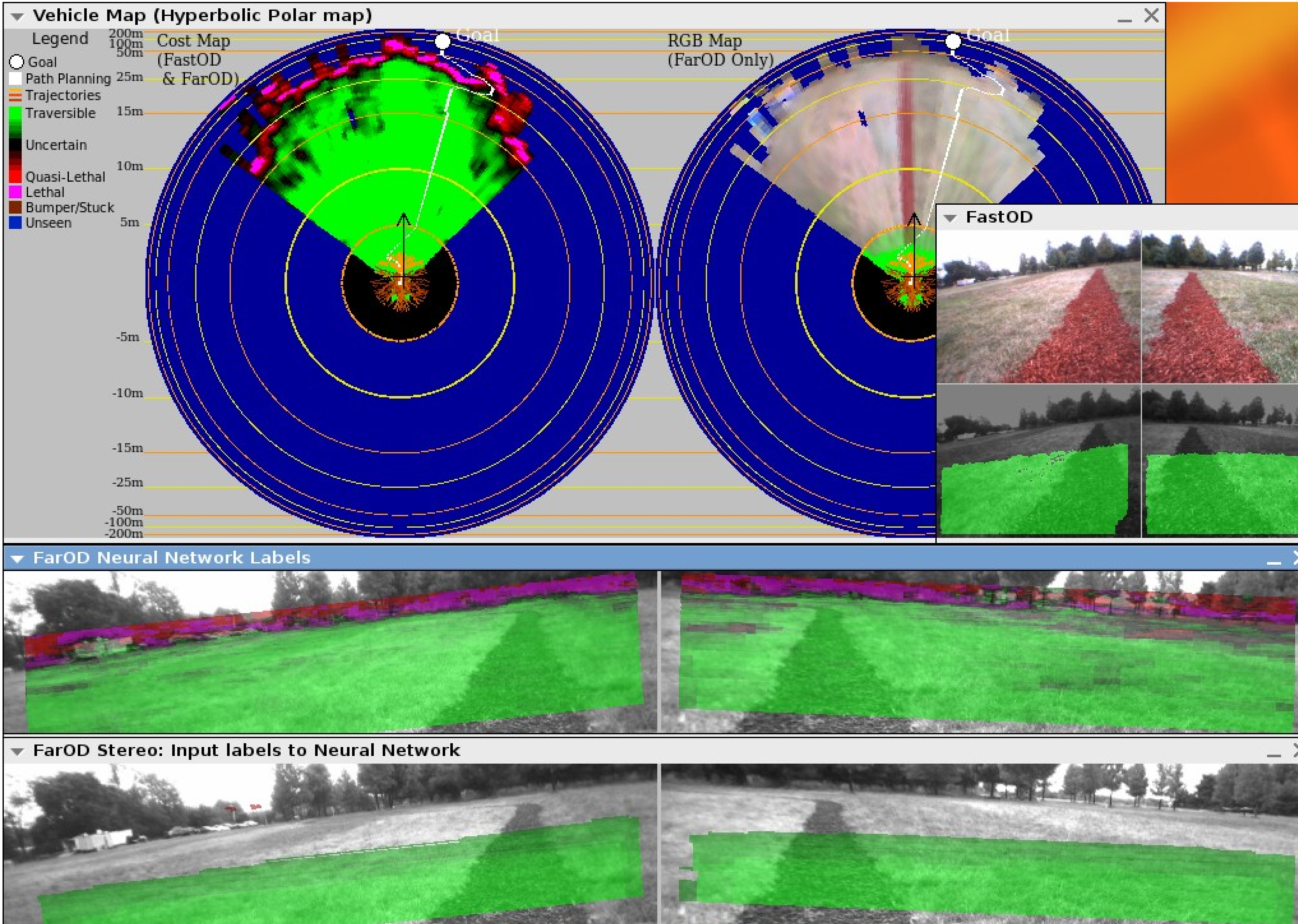


net SCALE Technologies, Inc.

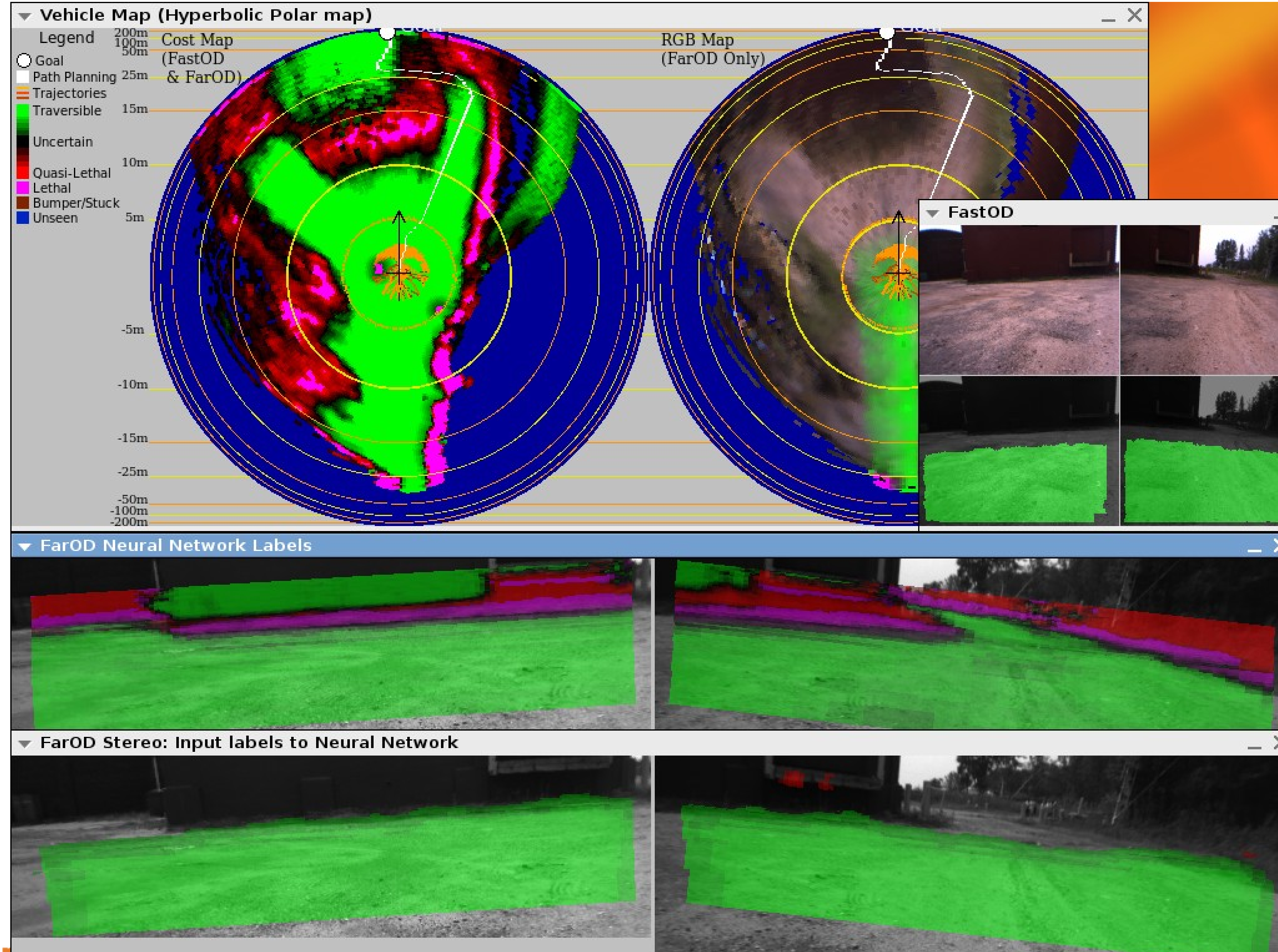
RSITY

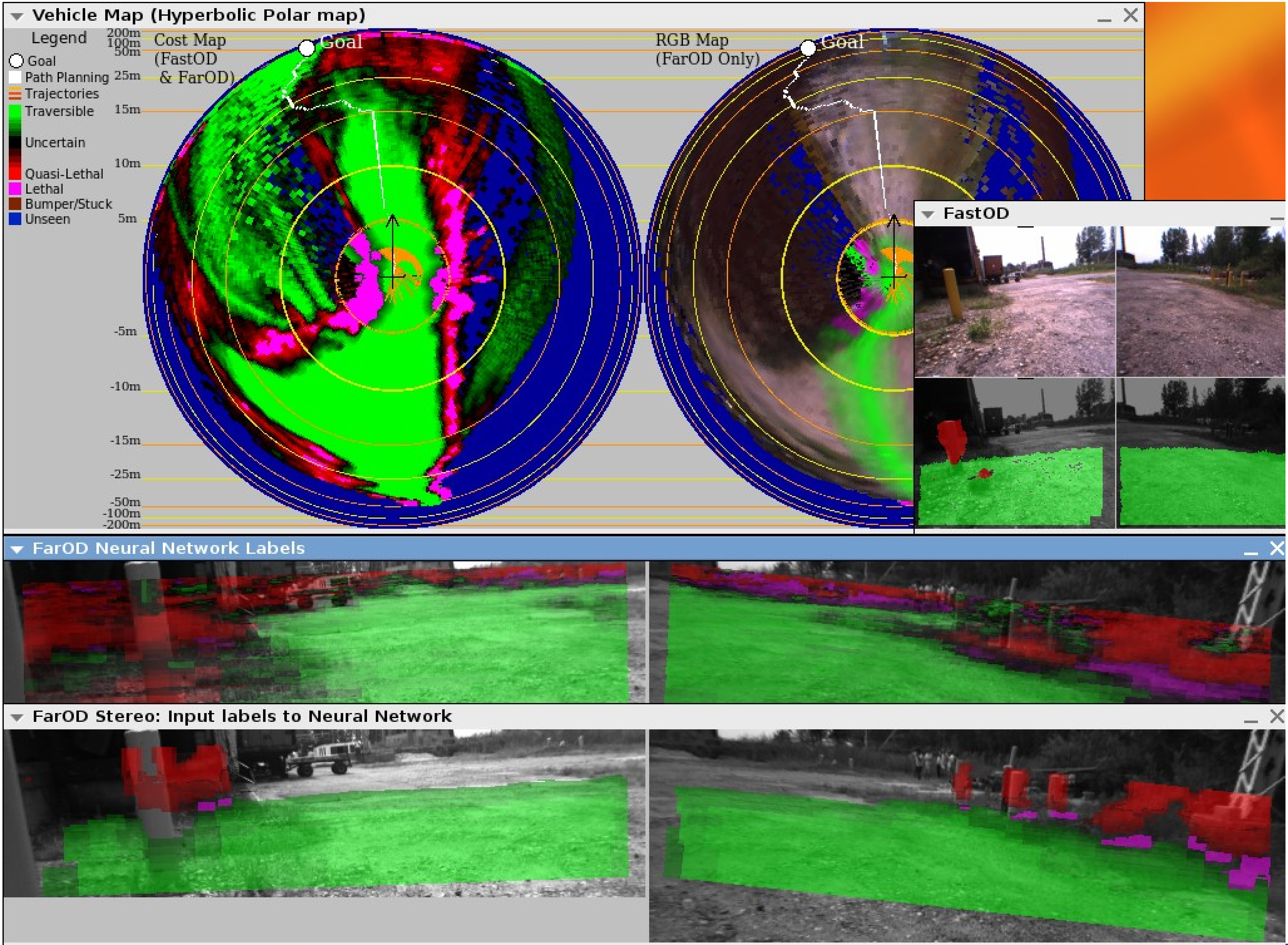


Vehicle Map (Hyperbolic Polar map)
Legend
Goal
Path Planning
Trajectories
Traversible
Uncertain
Quasi-Lethal
Lethal
Bumper/Stuck
Unseen
200m
100m
50m
25m
15m
10m
5m
-5m
-10m
-15m
-25m
-50m
-100m
-200m
Cost Map
(FastOD
& FarOD)
RGB Map
(FarOD Only)
FastOD
FarOD Neural Network Labels
FarOD Stereo: Input labels to Neural Network



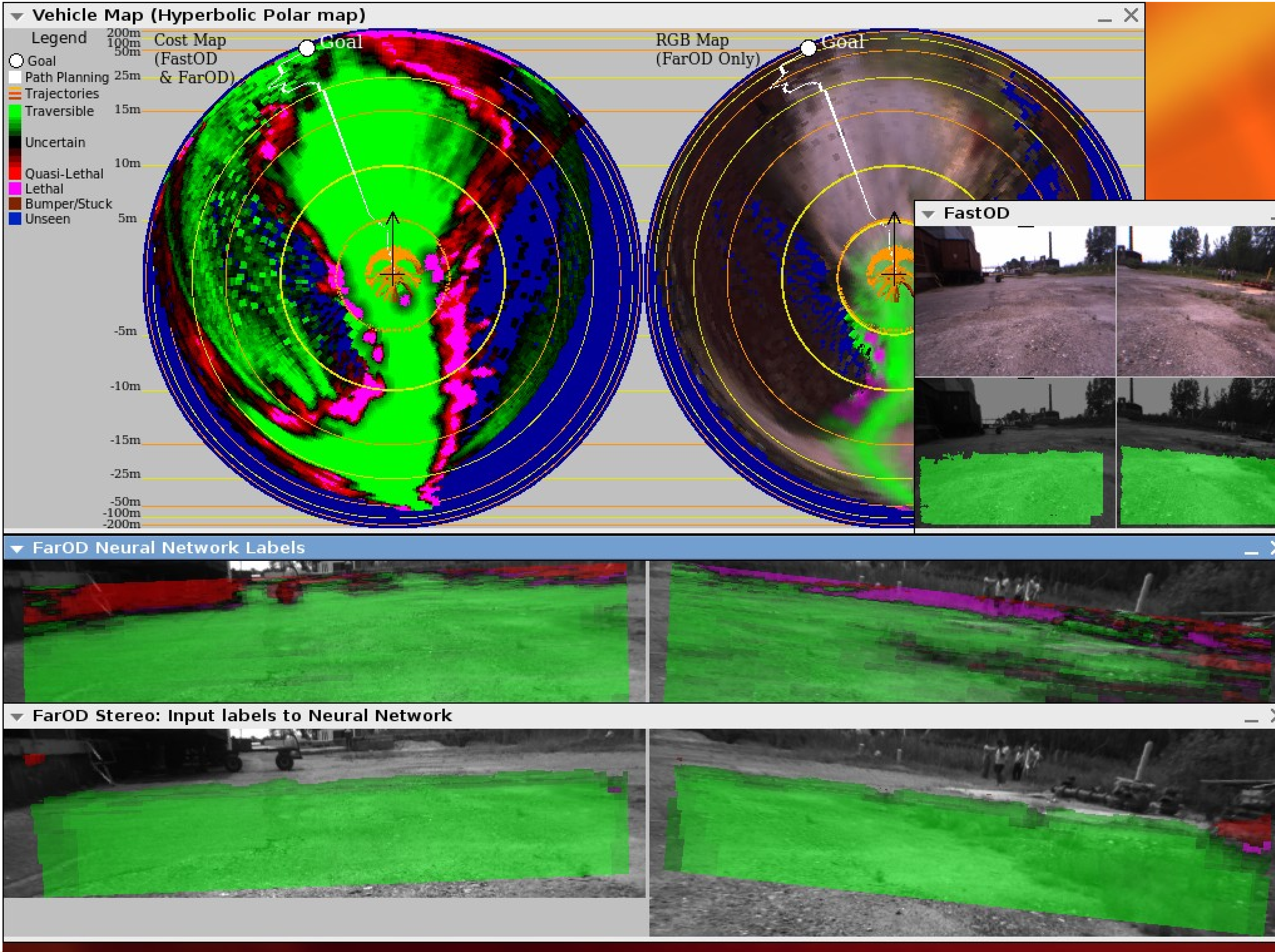
Vehicle Map (Hyperbolic Polar map)
Legend
Goal
Path Planning
Trajectories
Traversible
Uncertain
Quasi-Lethal
Lethal
Bumper/Stuck
Unseen
200m
100m
50m
25m
15m
10m
5m
-5m
-10m
-15m
-25m
-50m
-100m
-200m
Cost Map
(FastOD
& FarOD)
Goal
RGB Map
(FarOD Only)
Goal
FastOD
FarOD Neural Network Labels
FarOD Stereo: Input labels to Neural Network


Vehicle Map (Hyperbolic Polar map)
Legend
Goal
Path Planning
Trajectories
Traversible
Uncertain
Quasi-Lethal
Lethal
Bumper/Stuck
Unseen
200m
100m
50m
25m
15m
10m
5m
-5m
-10m
-15m
-25m
-50m
-100m
-200m
Cost Map (FastOD & FarOD)
RGB Map (FarOD Only)
FastOD
FarOD Neural Network Labels
FarOD Stereo: Input labels to Neural Network



Technologies, Inc.

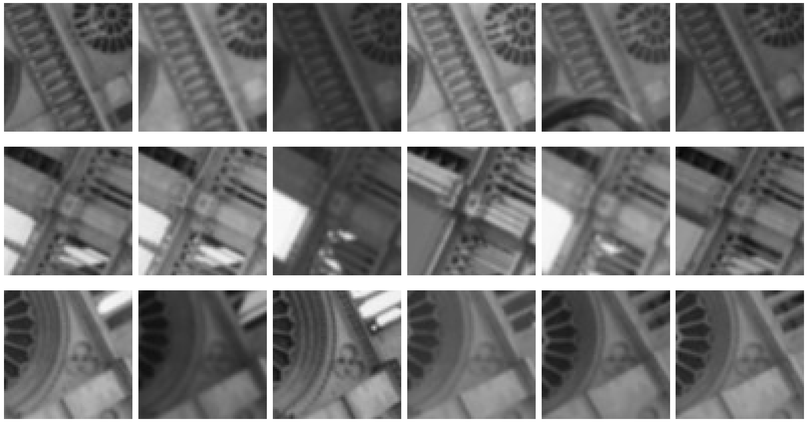
NEW YORK UNIVERSITY


Vehicle Map (Hyperbolic Polar map)
Legend
Goal
Path Planning
Trajectories
Traversible
Uncertain
Quasi-Lethal
Lethal
Bumper/Stuck
Unseen
200m
100m
50m
25m
15m
10m
5m
-5m
-10m
-15m
-25m
-50m
-100m
-200m
Cost Map
(FastOD
& FarOD)
Goal
RGB Map
(FarOD Only)
Goal
FastOD
FarOD Neural Network Labels
FarOD Stereo: Input labels to Neural Network

Learning Deep Invariant Features with DrLIM

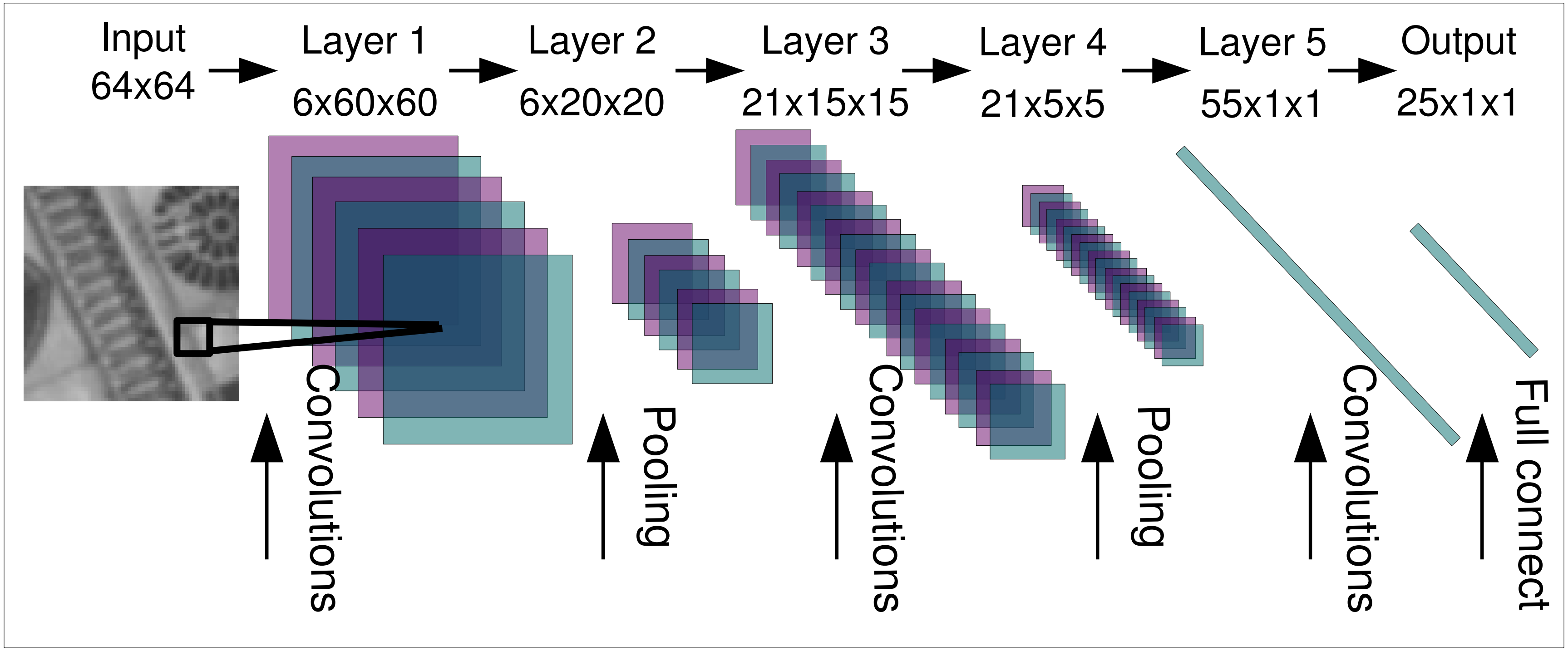
Co-location patch data

- multiple tourist photos
- 3d reconstruction
- groundtruth matches



Uses temporal consistency

- Pull together outputs for same patch
- Push away outputs for different patches



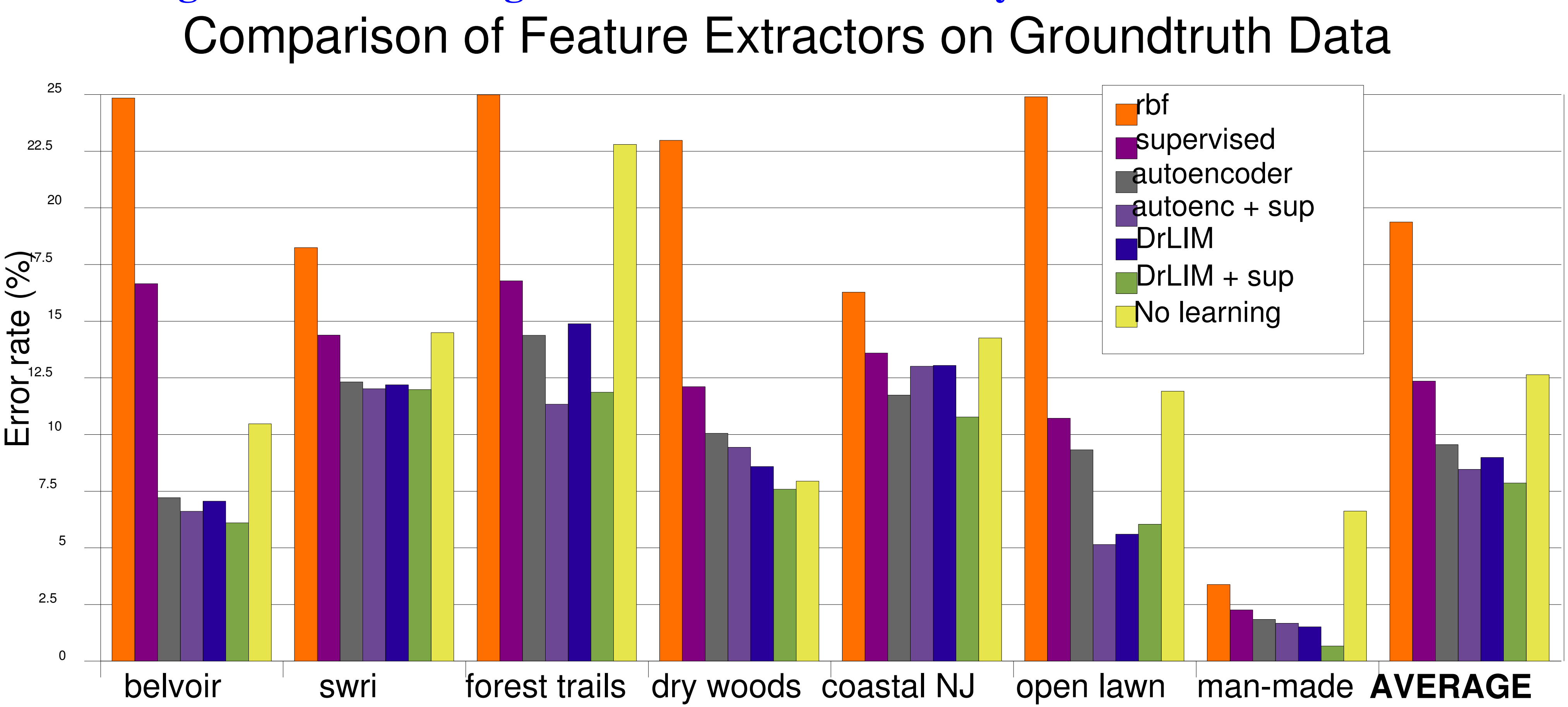
data from: Winder and Brown, CVPR 07

Feature Learning for traversability prediction (LAGR)

Comparing

- purely supervised
- stacked, invariant auto-encoders
- DrLIM invariant learning

Testing on hand-labeled groundtruth frames – binary labels



The End