GPU TECHNOLOGY CONFERENCE

April 4-7, 2016 | Silicon Valley

Neural Attention for Object Tracking

Brian Cheung

bcheung@berkeley.edu

Redwood Center for Theoretical Neuroscience, UC Berkeley

Visual Computing Research, NVIDIA

PRESENTED BY NVIDIA

REDWOOD CENTER for Theoretical Neuroscience

Source: Wikipedia “School Bus”

Motivation

Solving complex vision problems

- Question Answering
- Search
- Navigation

Two core components:

- Attention
- Memory

Emergent Properties from Attention

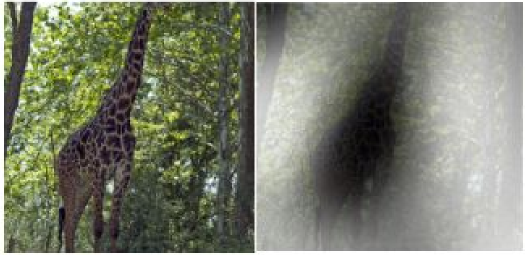
A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

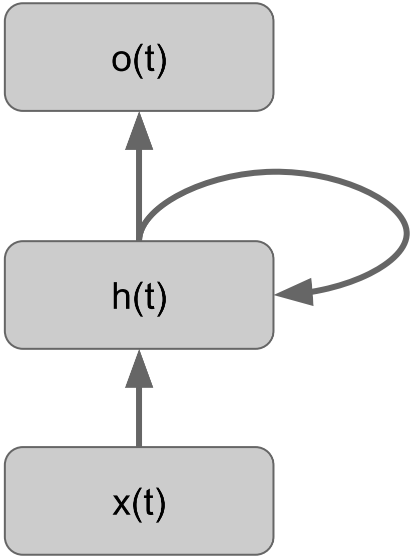
A group of people sitting on a boat in the water.



A giraffe standing in a forest with trees in the background.

Xu et. al. 2015

Recurrent Networks



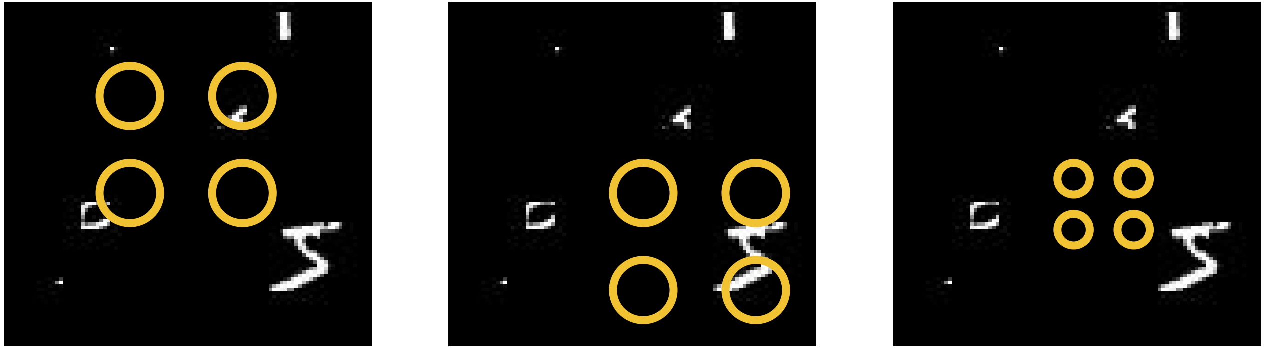
$$h(t) = \sigma(x(t)W_{xh} + h(t-1)W_{hh})$$
$$o(t) = \sigma(h(t)W_{ho})$$
$$h(t+1) = \sigma(x(t+1)W_{xh} + h(t)W_{hh})$$

Formulating a Glimpse

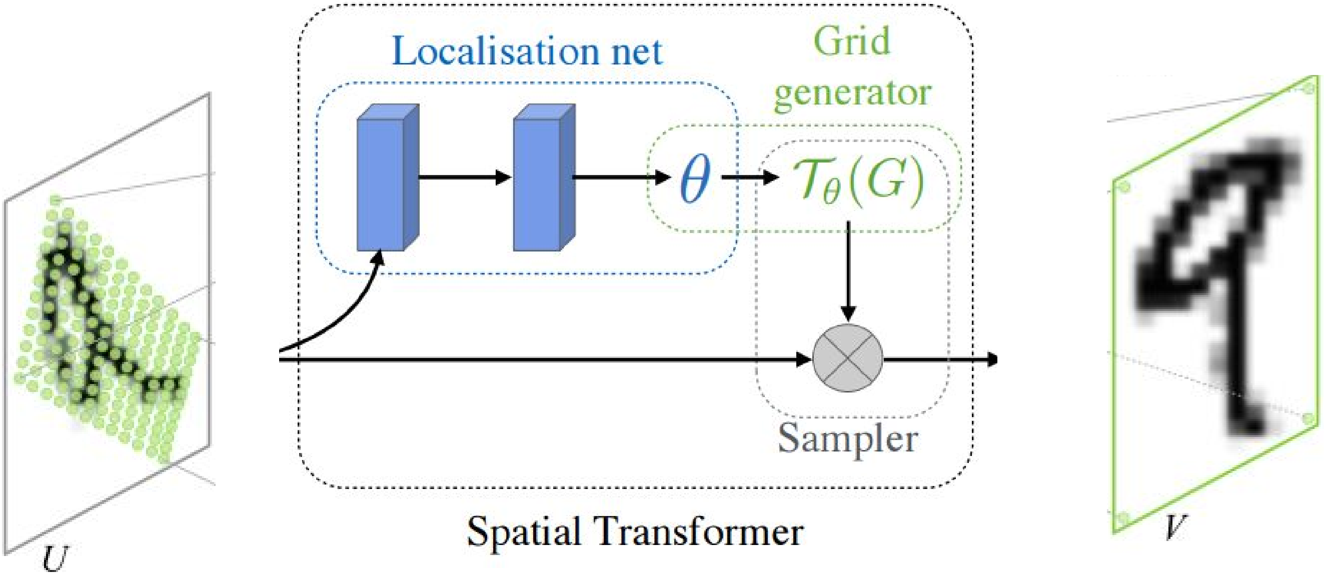
$$V_i = \sum_{n}^{H} \sum_{m}^{W} U(n,m) k_i(m,n)$$
$$\forall i \in [1, ..., H'W']$$

Parameters in the kernel control the layout of the attention window over the original image.

Translation **Scale**

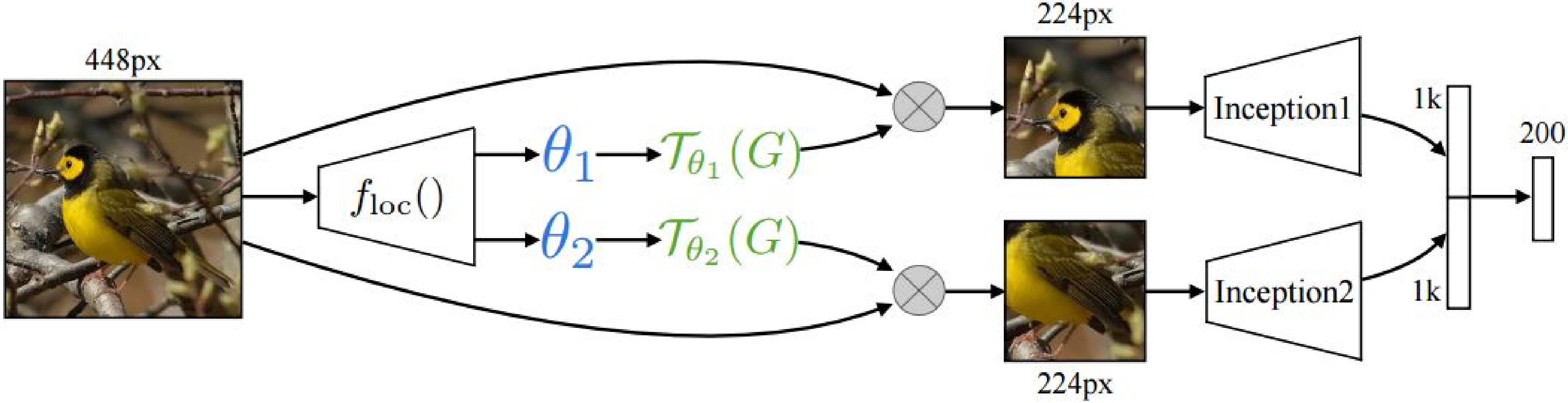


Spatial Transformer



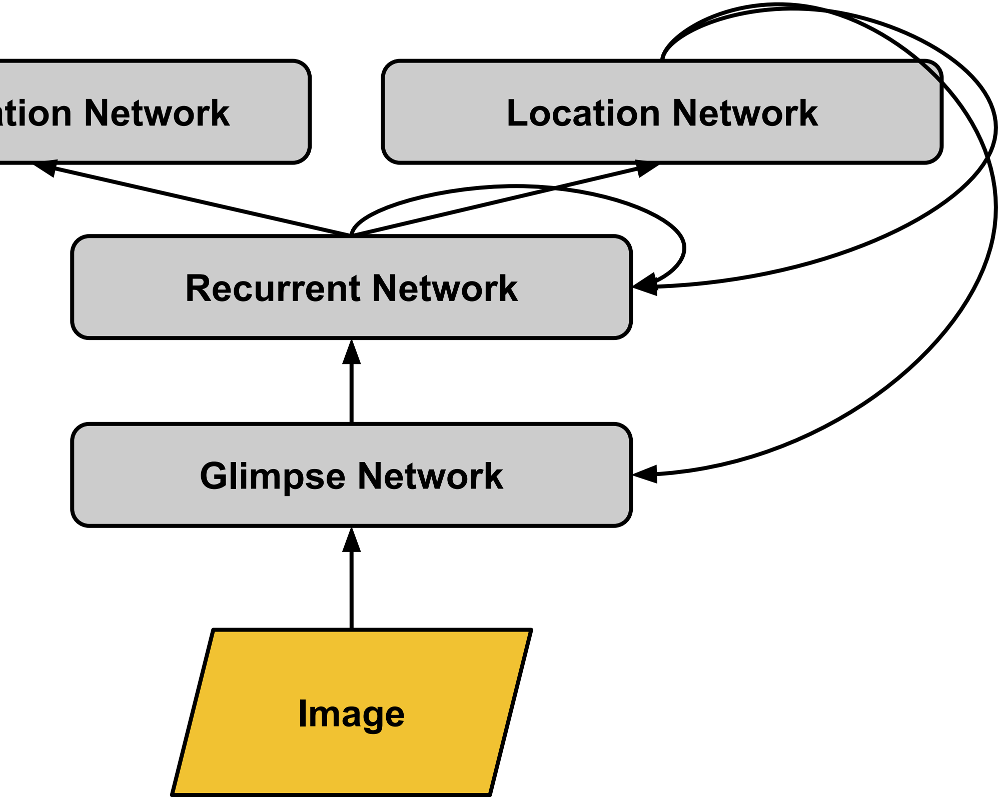
Jaderberg et. al. 2015

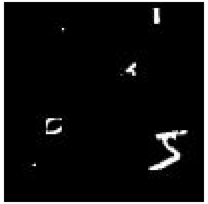
Spatial Transformer Network



Jaderberg et. al. 2015

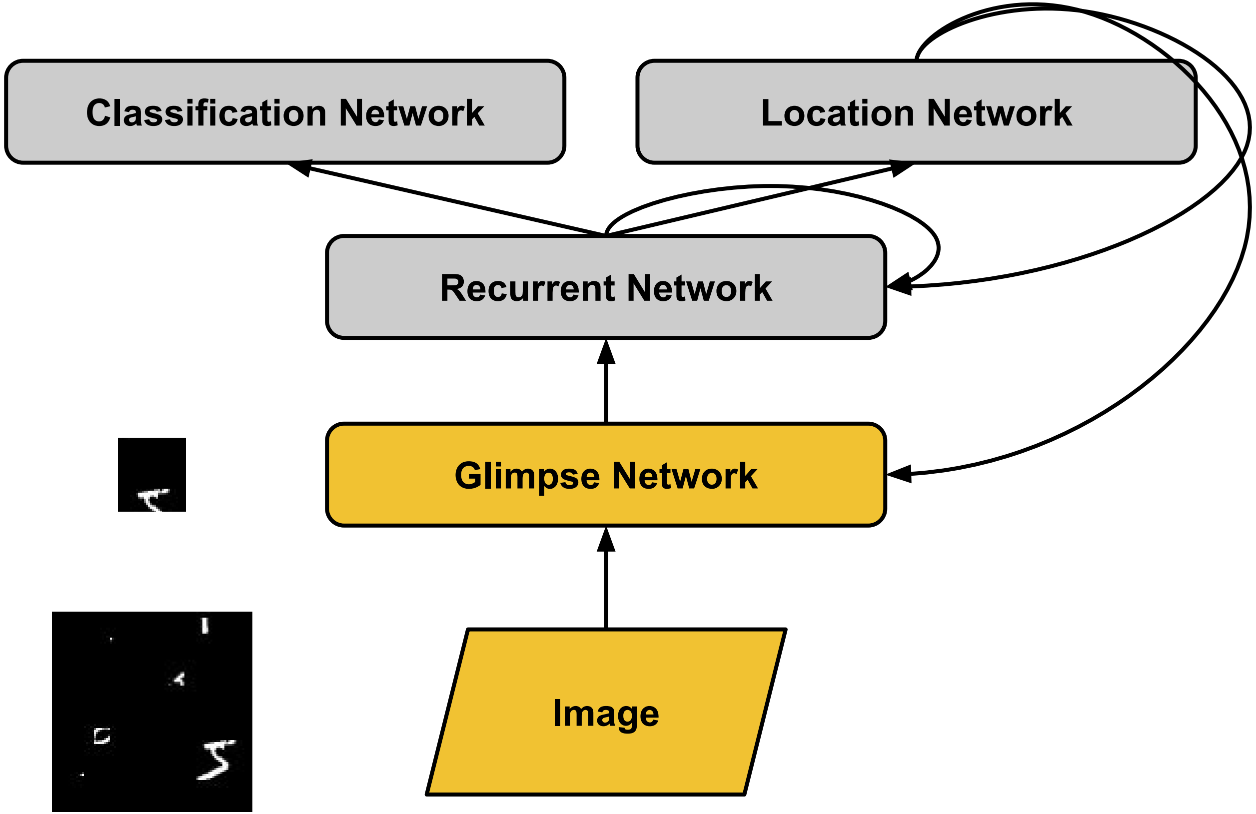
Foveal Attention Network





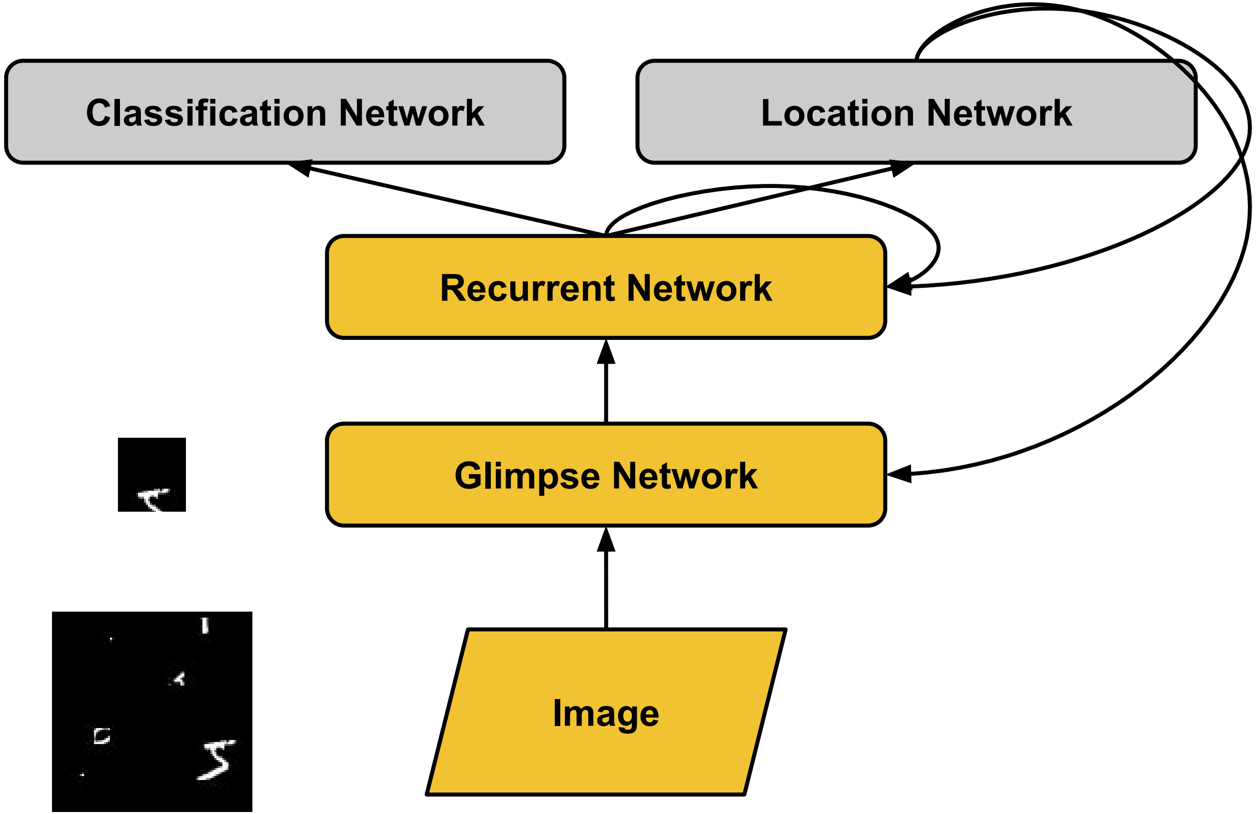
Cheung et. al. 2015

Foveal Attention Network



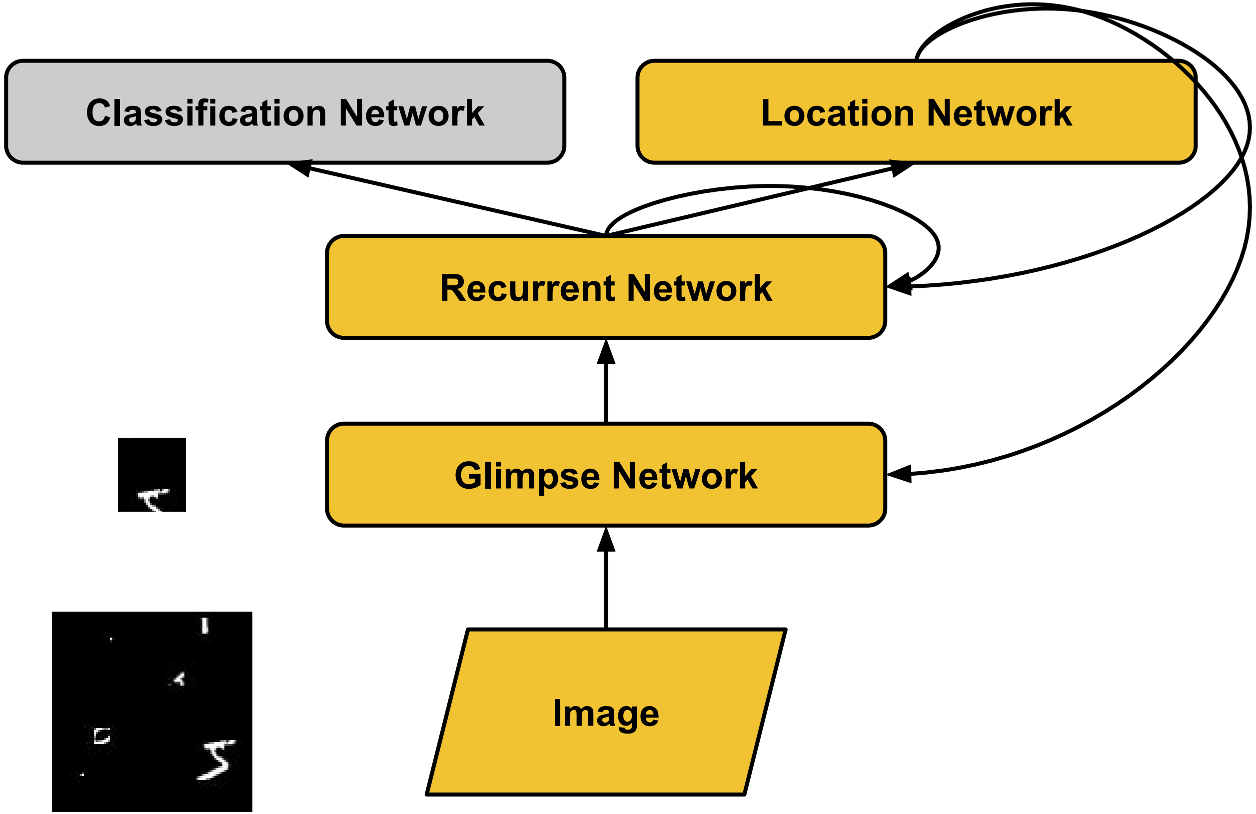
Cheung et. al. 2015

Foveal Attention Network



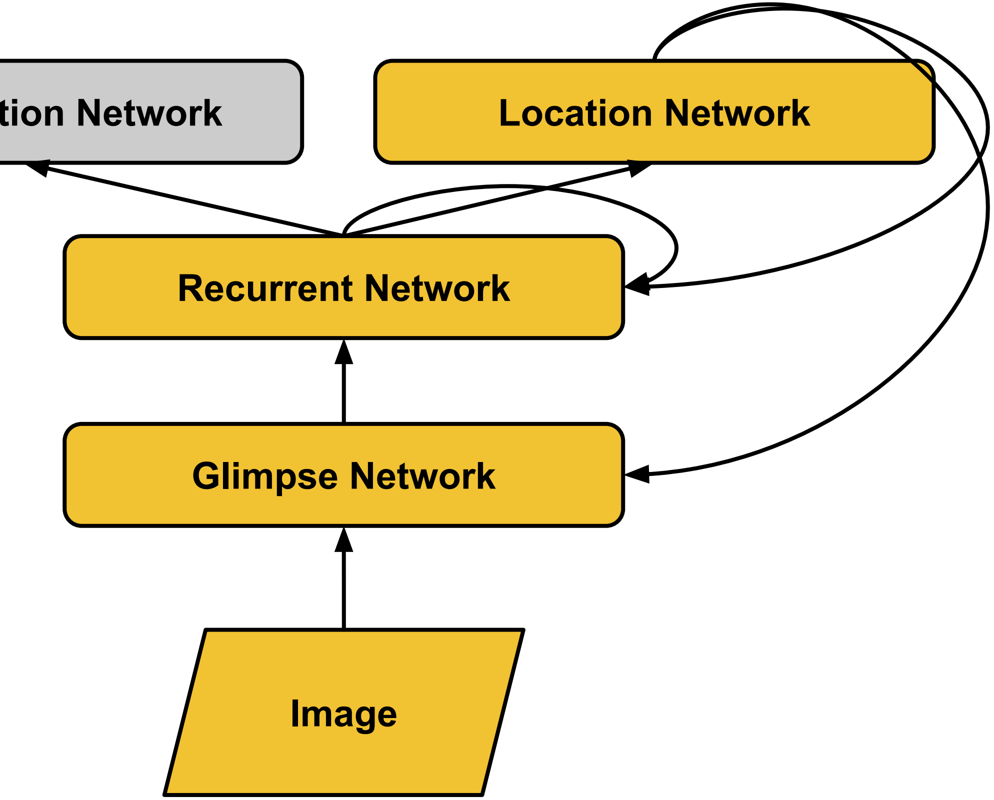
Cheung et. al. 2015

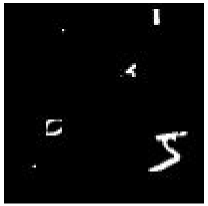
Foveal Attention Network



Cheung et. al. 2015

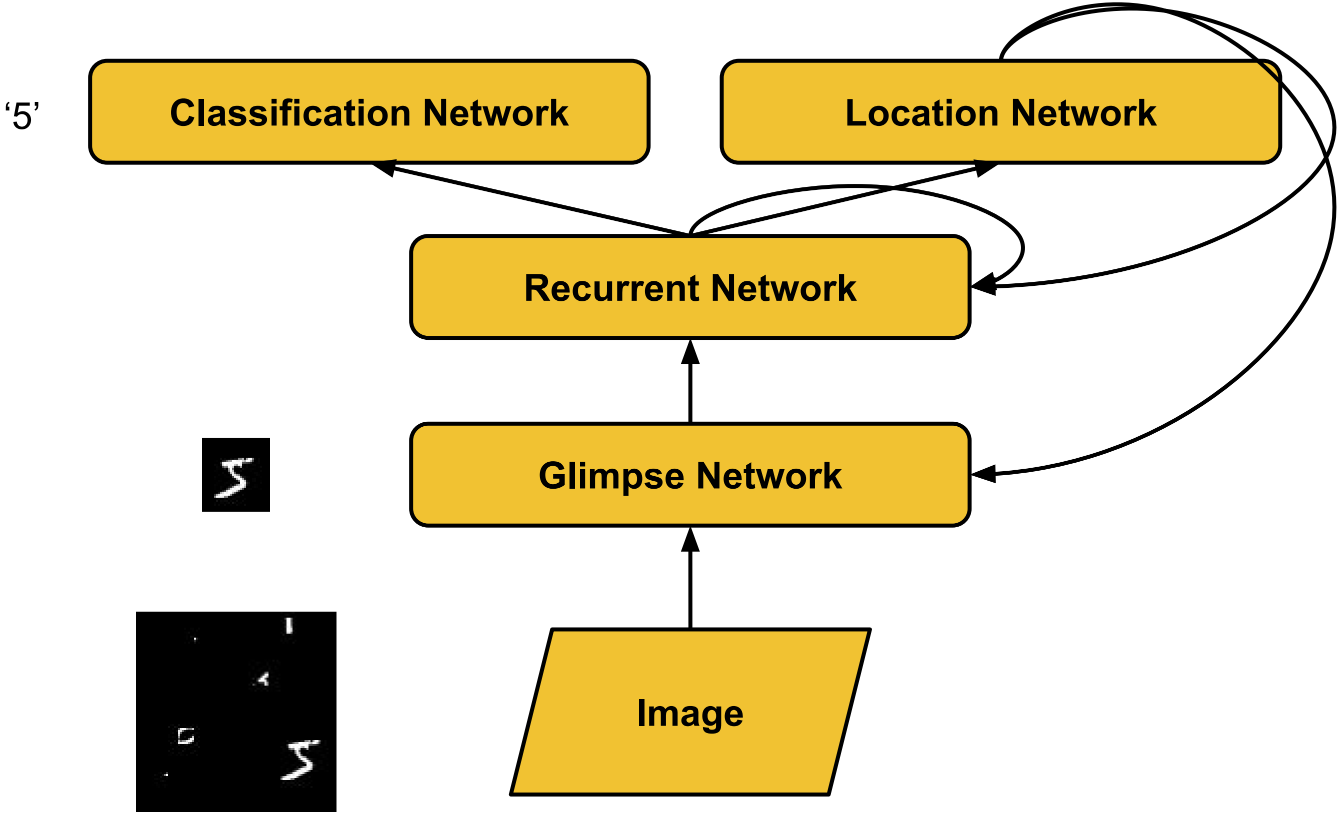
Foveal Attention Network





Cheung et. al. 2015

Foveal Attention Network



Cheung et. al. 2015

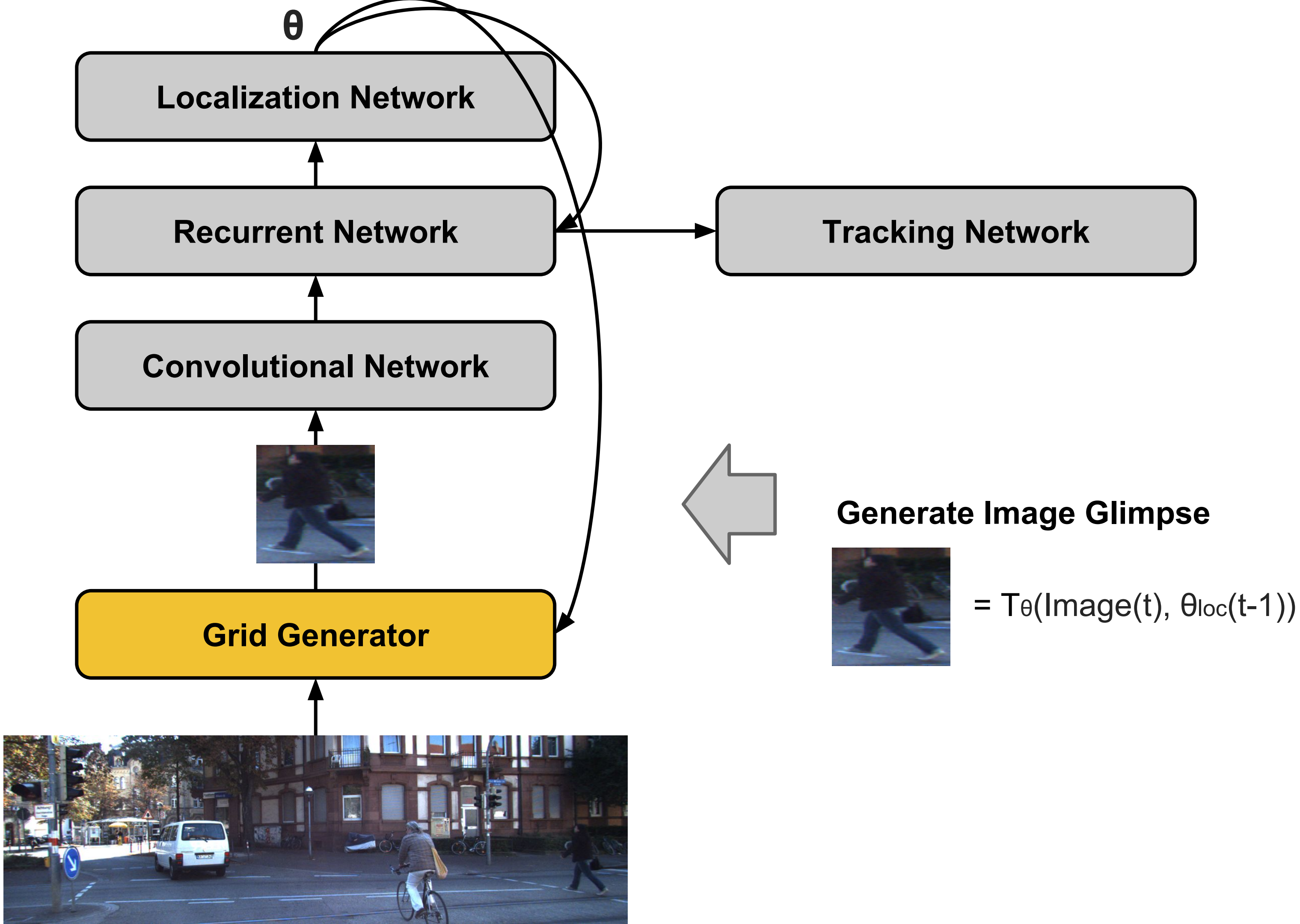
Benefits of Attention

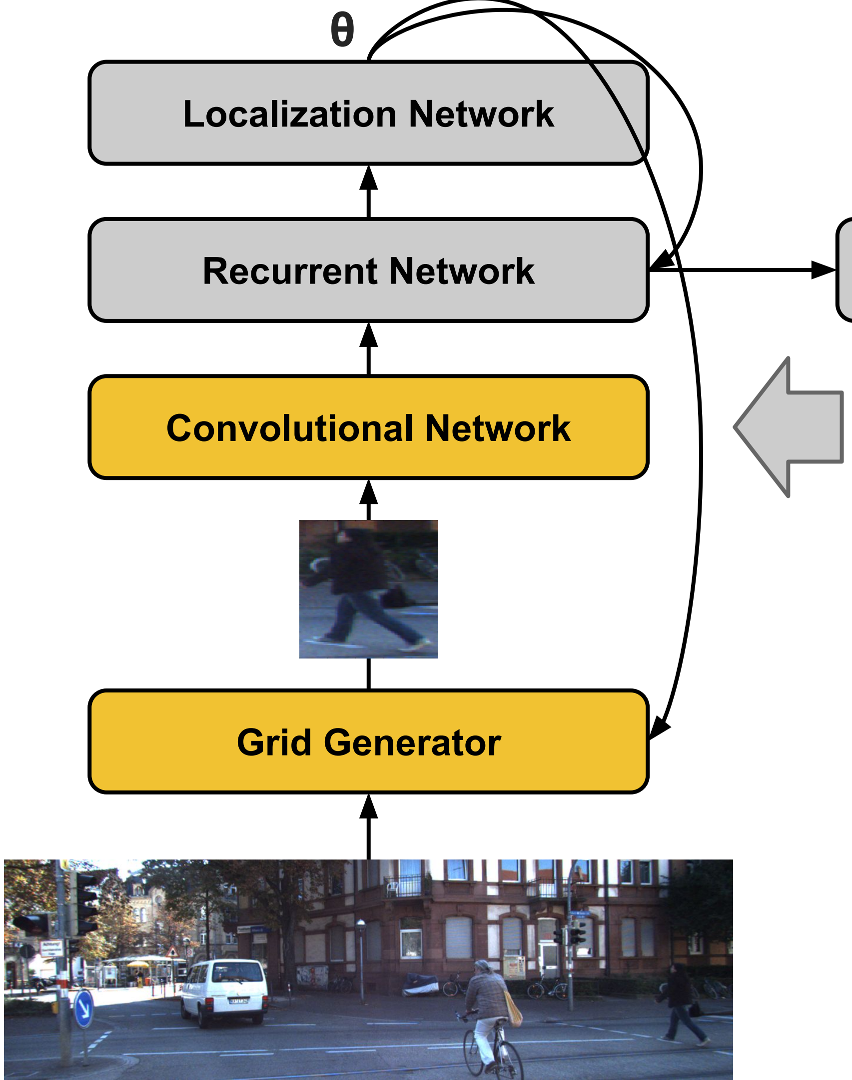
- Less parameters/Less Computation
  - Smaller Convolutional Network
- Better Performance
  - Significant performance over ConvNet over entire image
  - Breaks down complex problems into a sequence of simpler problems
  - Filters out noise and distractors
- Localization information is free

KITTI Tracking Dataset

Geiger et. al. 2012

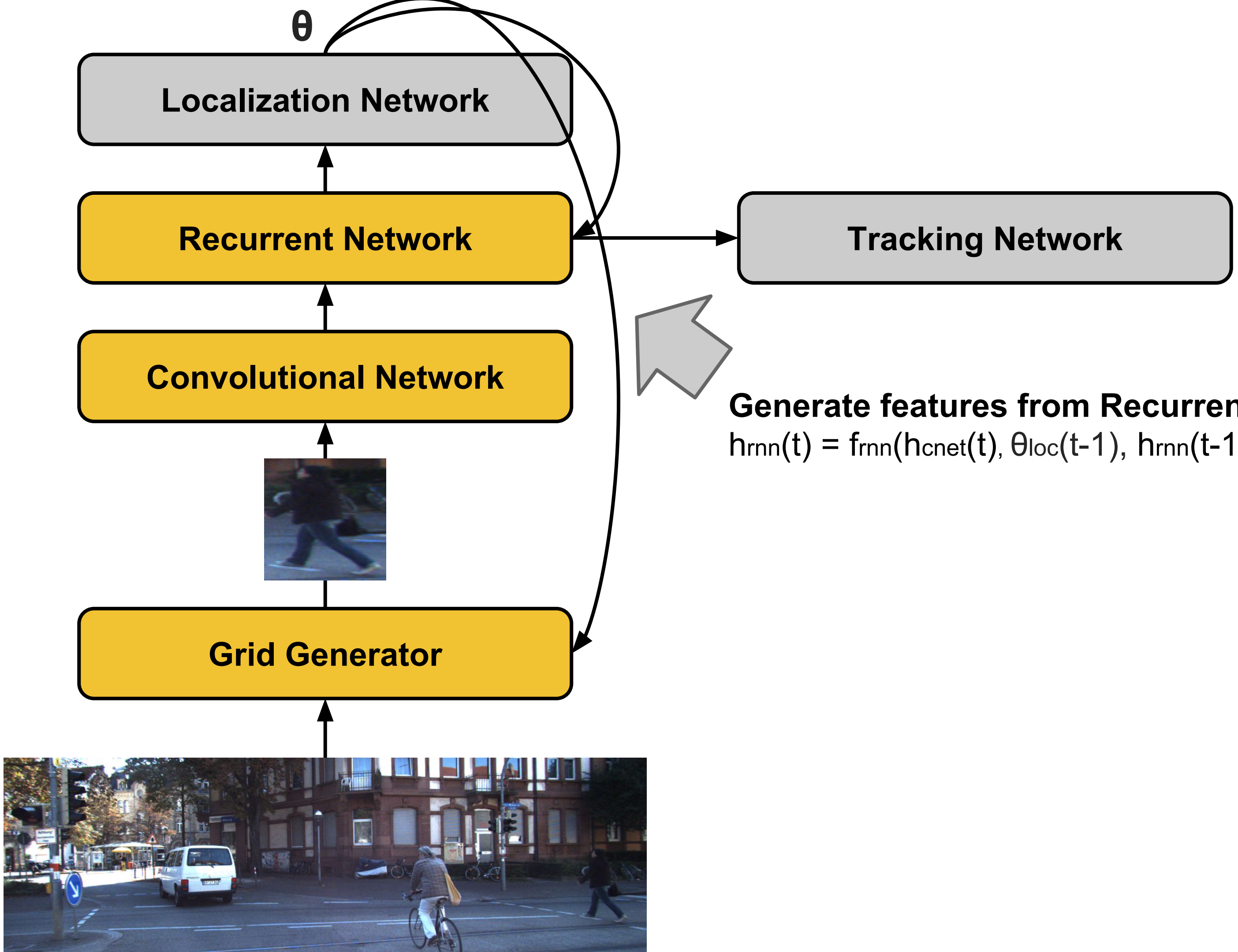
- 375x1240 video
- Bounding boxes over time of cars, pedestrians, etc.


θ
Localization Network
Recurrent Network
Tracking Network
Convolutional Network
Generate Image Glimpse
= Tθ(Image(t), θloc(t-1))
Grid Generator



Generate features from ConvNet

$h_{cnet}(t) = f_{cnet}($  $)$

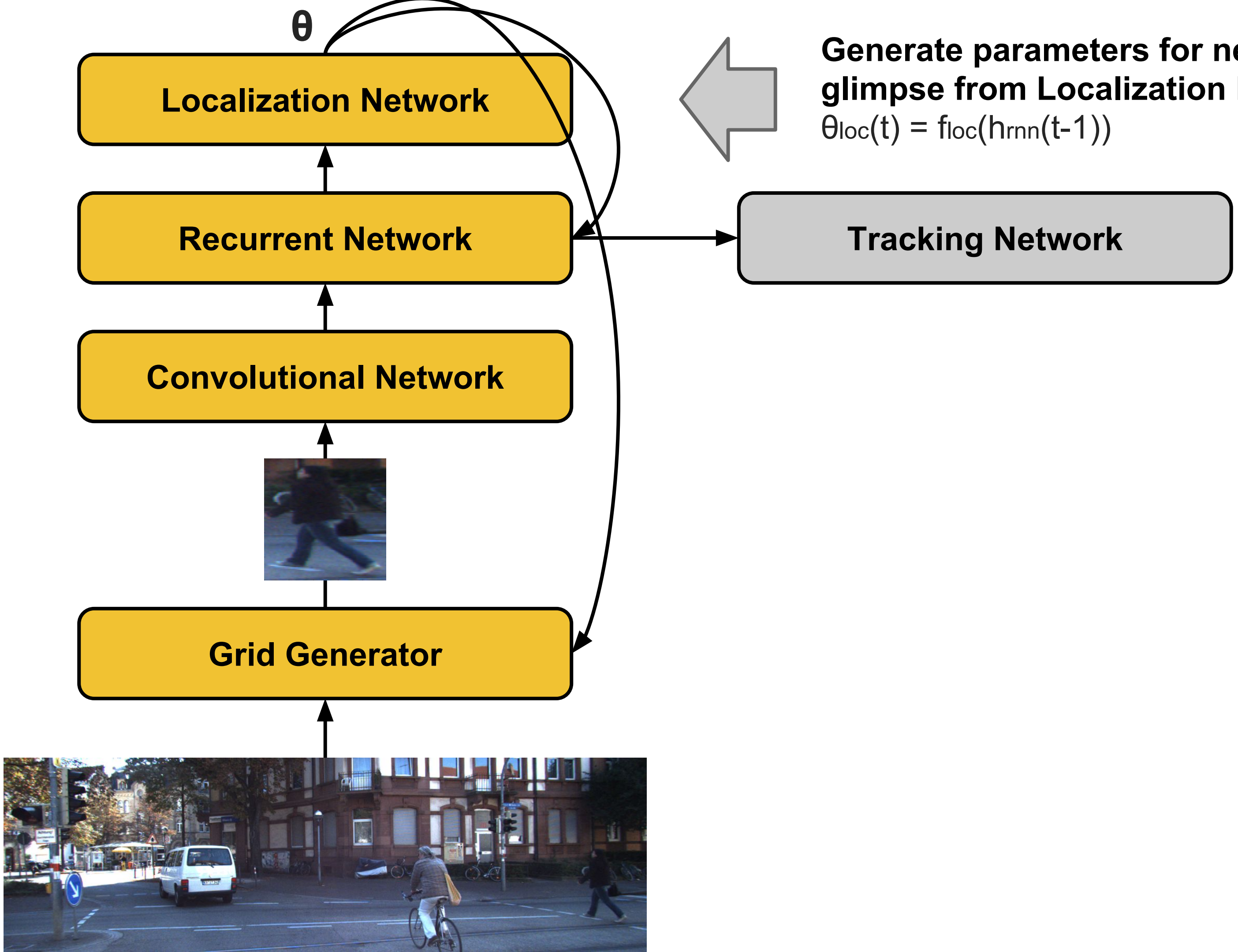


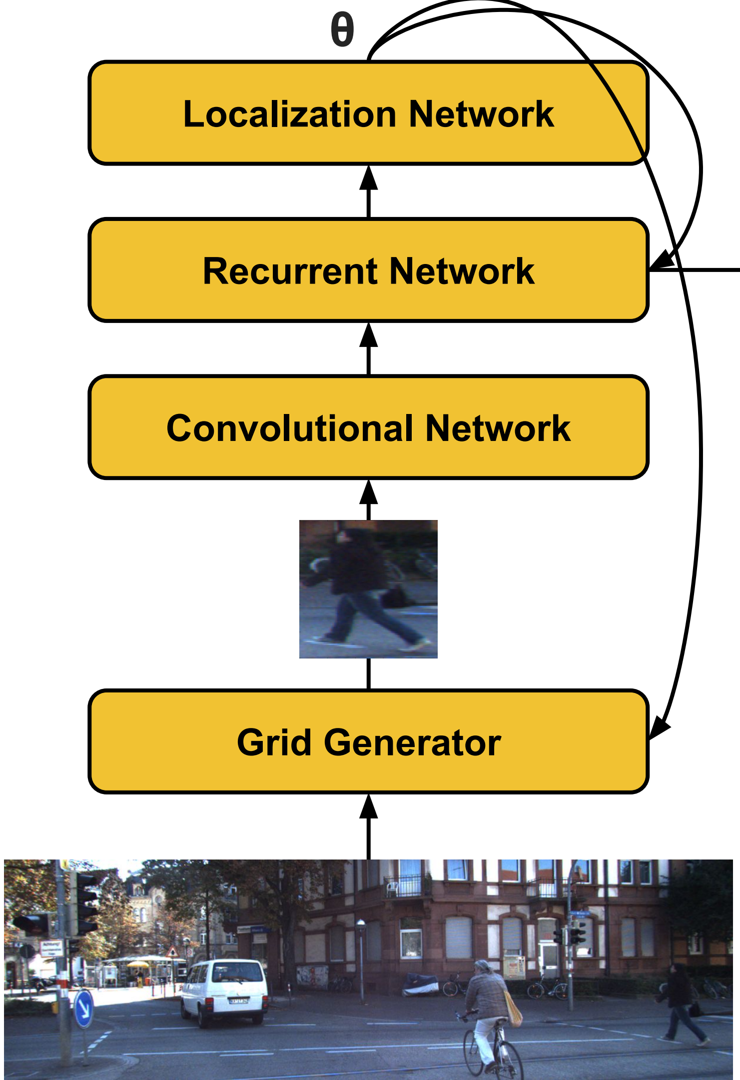
Generate features from Recurrent Network

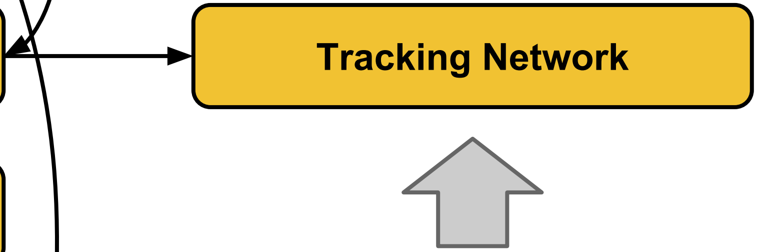
$h_{rnn}(t) = f_{rnn}(h_{cnet}(t), \theta_{loc}(t-1), h_{rnn}(t-1))$

Generate parameters for next glimpse from Localization Network

$\theta_{loc}(t) = f_{loc}(h_{rnn}(t-1))$



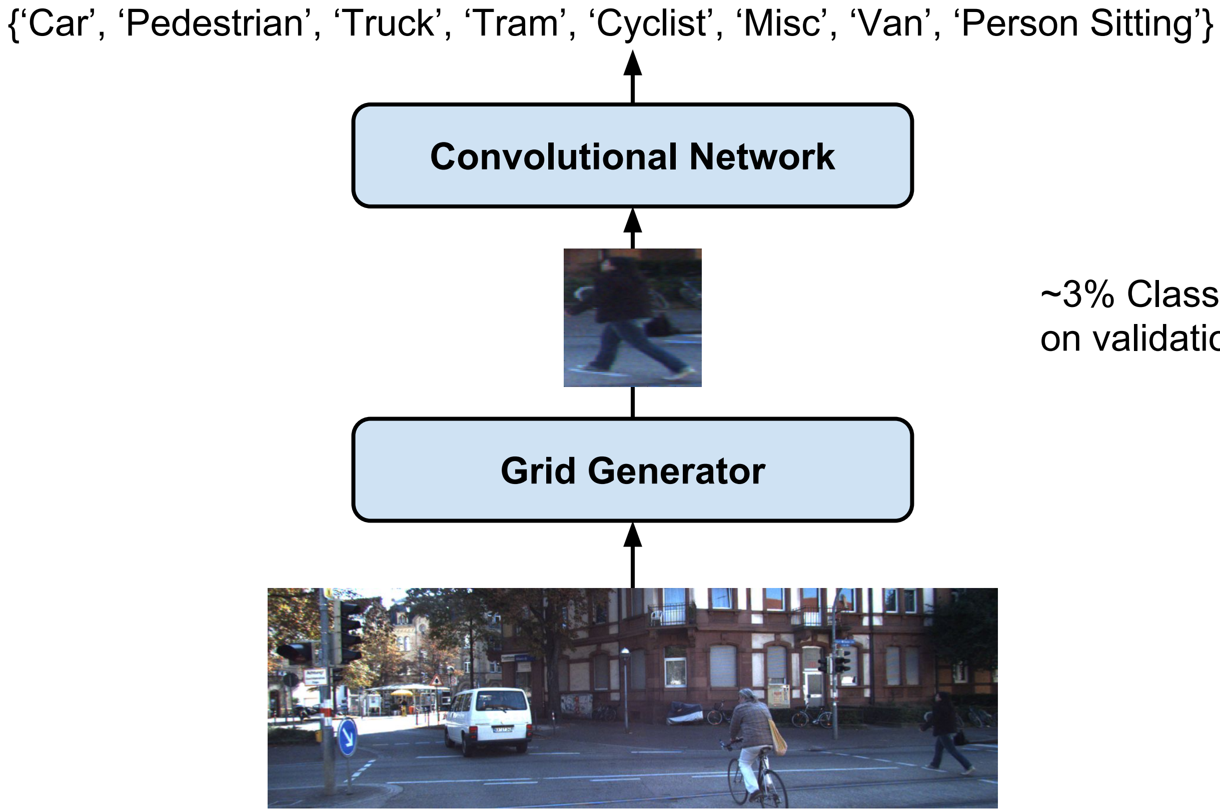




Generate tracking prediction from Tracking Network

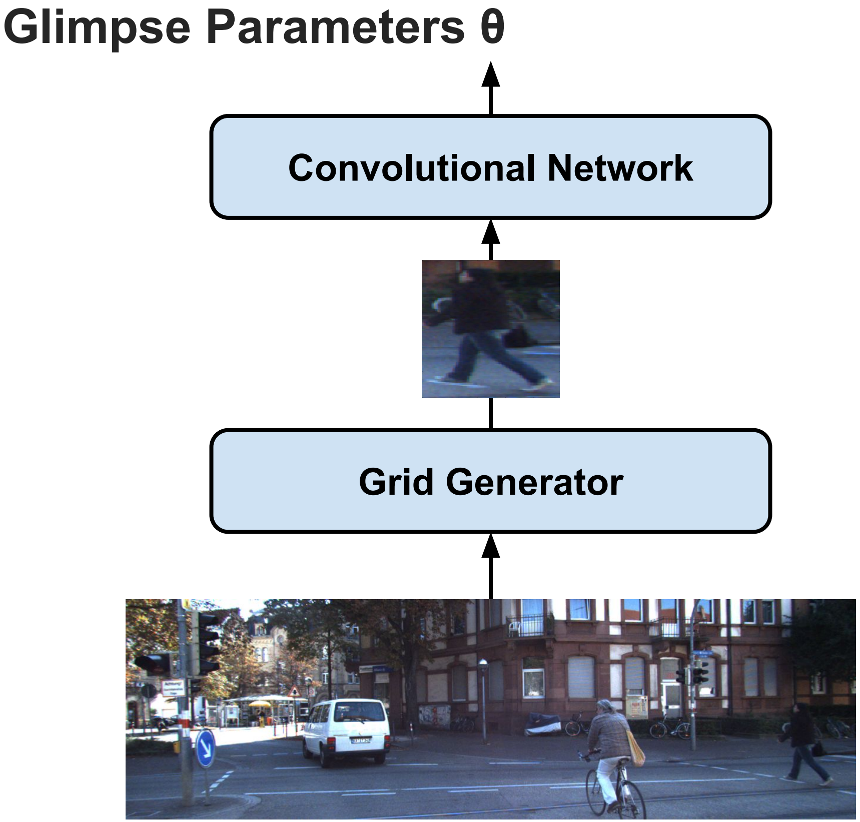
$\theta_{pred}(t),\ y_{pres}(t) = f_{tracking}(h_{rnn}(t-1))$

Pretraining on Classification Task



~3% Classification Error
on validation set

Pretraining on the Registration Task



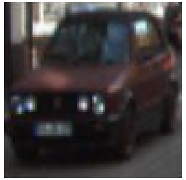
Pretraining on the Registration Task

- Simpler task similar to tracking: Fix a bad glimpse
- Useful signal for Localization Network

Input Glimpse

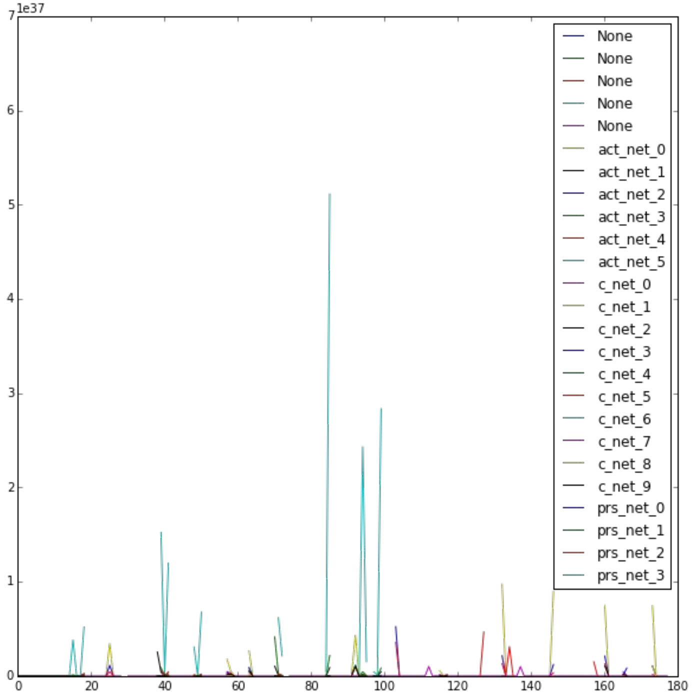
Predicted Correction

Actual Correction

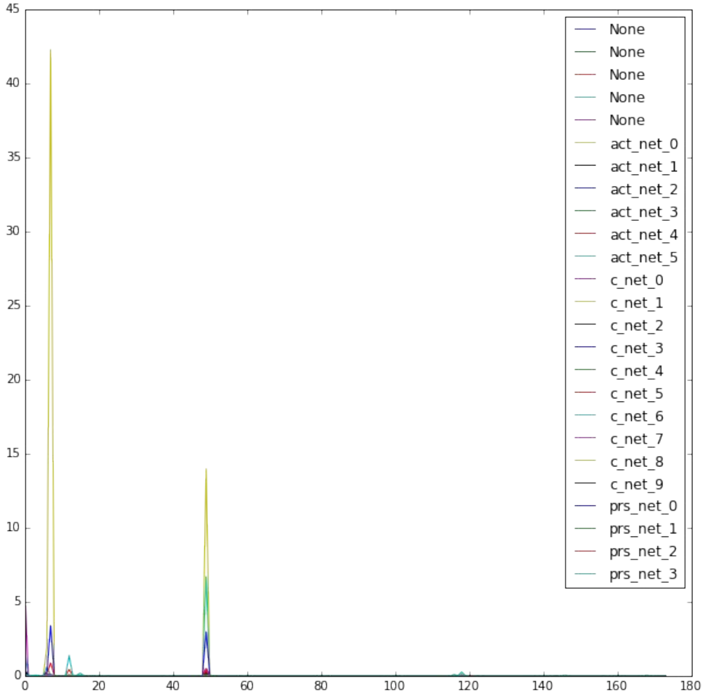


Comparing Training Gradients

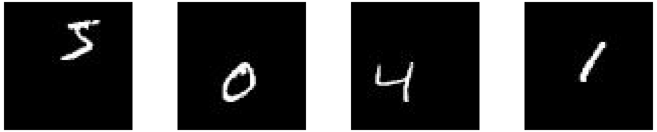
Without pretraining (Random Initialization)



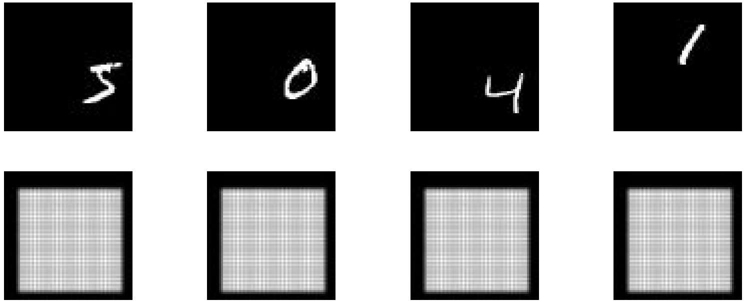
With ConvNet Pretraining



Bouncing MNIST

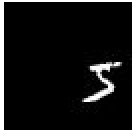


Bouncing MNIST



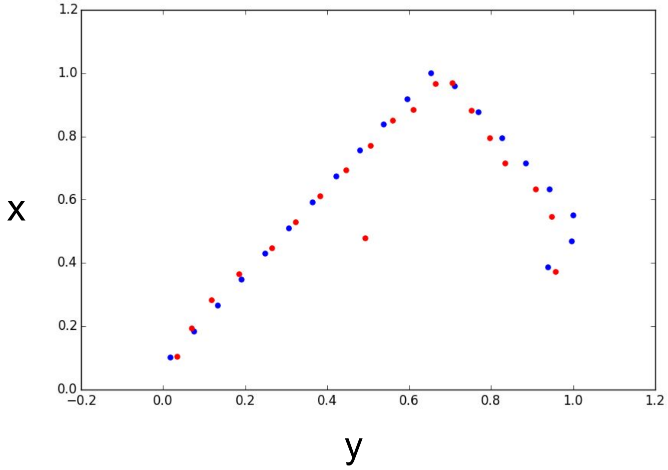
Bouncing MNIST

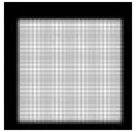
Output:



MNIST Position

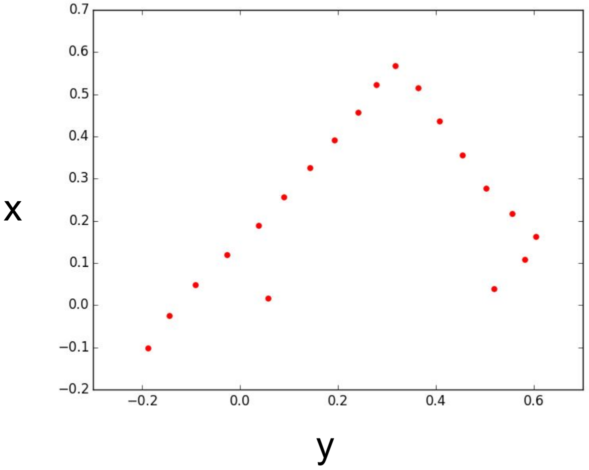
Tracking Network





Attention Position

Localization Network



Ground Truth

Prediction

Conclusions

- End-to-End visual attention works for simple tasks
- Robust to encoding of attention parameters

Conclusions

- Difficult to train on more complex tasks
  - **First Step toward Model-Free, Anonymous Object Tracking with Recurrent Neural Networks (Gan et. al. 2015)**
  - **RATM: Recurrent Attentive Tracking Model (Kahou et. al. 2015)**
- Scaling computational costs

Future Work

- Integrate more tailored components
  - Spatial Memory (Weiss et. al. 2015)
- Train compact ImageNet models for initialization
- Exploration/Unsupervised strategies to recover from mistakes
  - Error Based Attention (Rezende et. al. 2016)

Acknowledgements

Special thanks to:

Shalini Gupta

Jan Kautz

Pavlo Molchanov

Stephen Tyree

Eric Weiss

NVIDIA.