Kevin Galim

Deep Learning for Video Depth Estimation from Defocus

Technische Universität München, Department of Informatics

Maxim Maximov

Prof. Dr. Laura Leal-Taixé

Depth Prediction

- Images only supply 2D information
- Depth essential for 3D understanding
- Applications
  - Robotics
  - Refocusing
  - Augmented Reality

Image

Estimated Depth

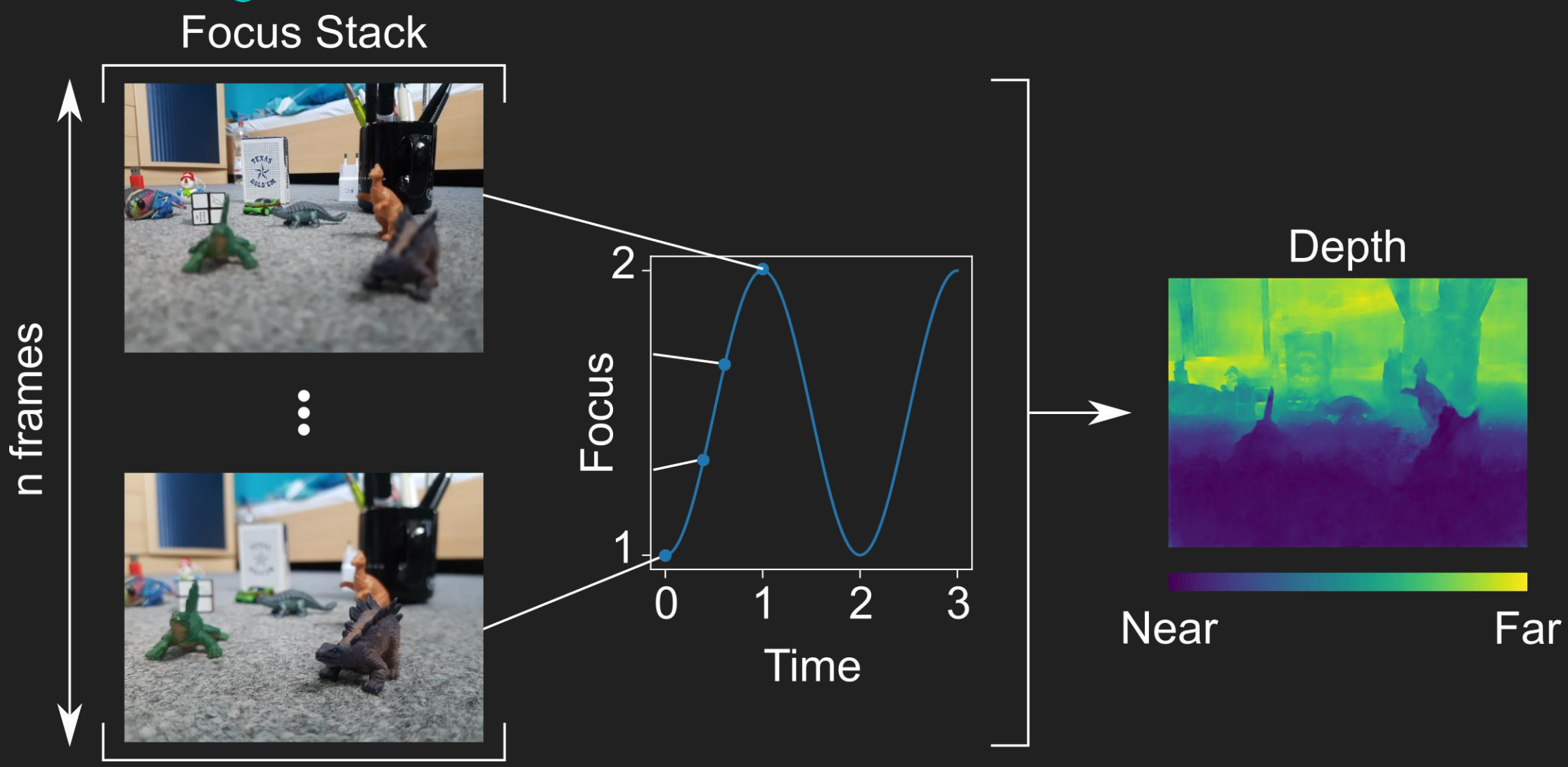
Motivation: Video Depth from Focus

- Video Depth-From-Defocus (Hyeongwoo et al.)
  - Similar aim: video depth from defocus
  - Complex multi-step variational approach
  - Limited regarding the amount of motion which can be compensated
- Deep Depth From Focus (Hazırbaş et al.)
  - Deep learning approach
  - Limited to static scenes
  - Limited generalization

Problem Statement

- Input: RGB sequence
  - Varying focus distance
  - Static or under the influence of motion
- Goal: Predict depth
  - Based on the amount of defocus
  - Using a deep learning approach

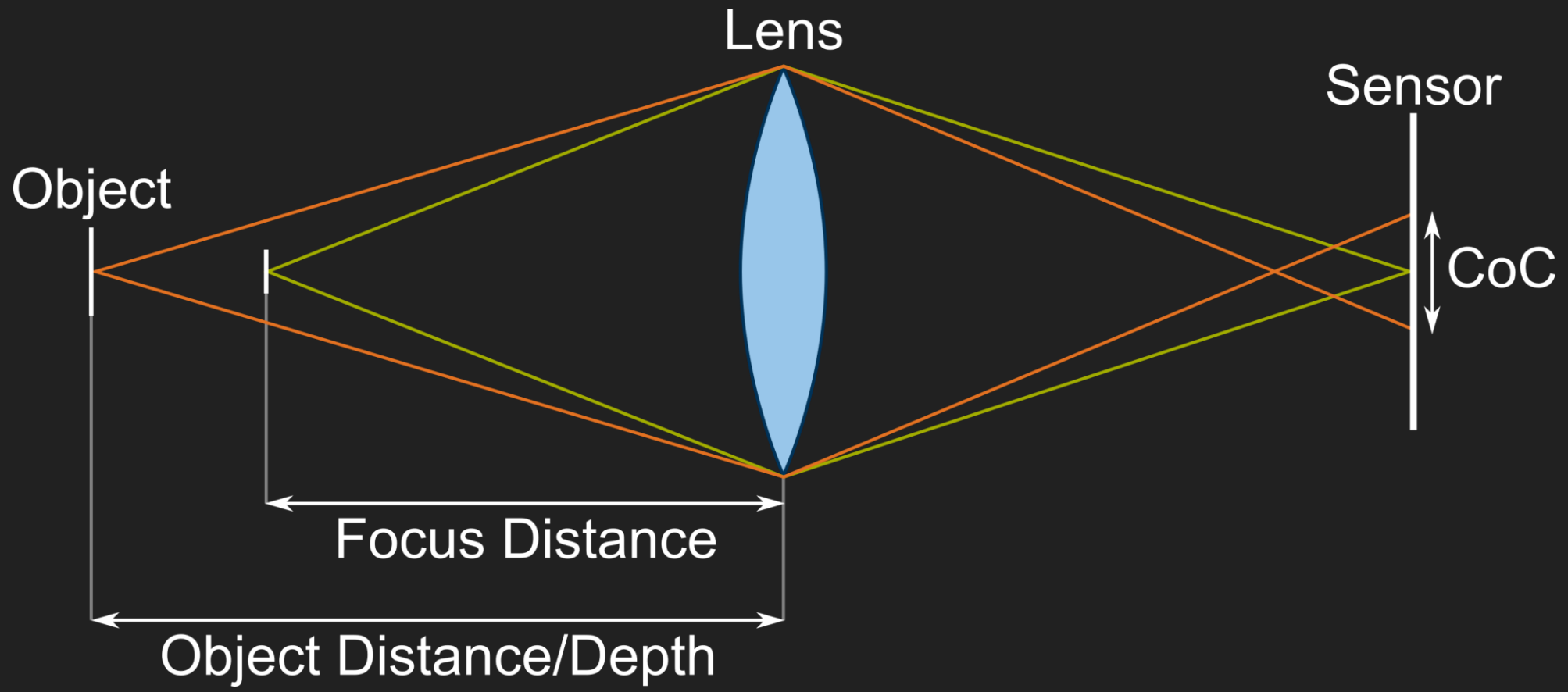
Concept



Key Contributions

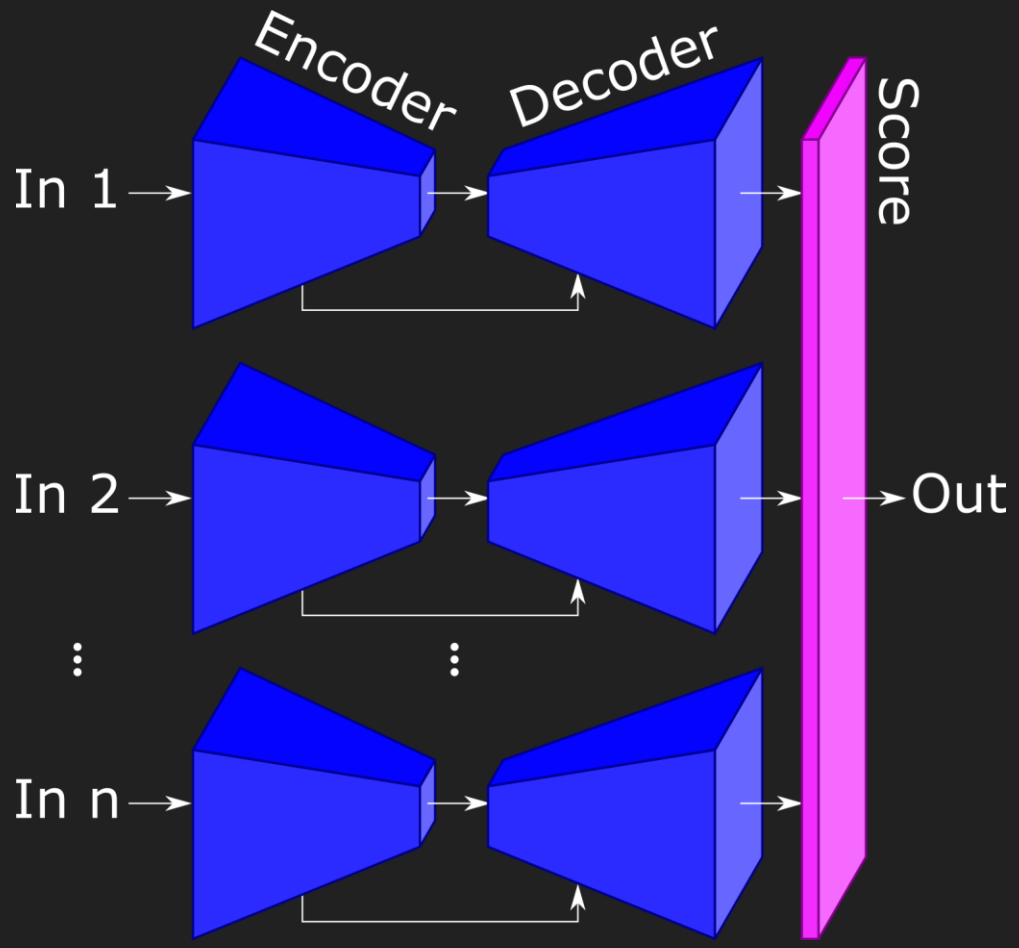
- CNN architectures for depth from focus prediction
  - Generalization for various settings (static and dynamic)
  - Robust to scene movement
- Real-World dataset
  - Android phone and RGB-D sensor
  - Calibration and synchronization
- Synthetic dataset
  - Fully automatic creation of large datasets at little expanse
  - Random scene generation

Circle of Confusion (CoC)



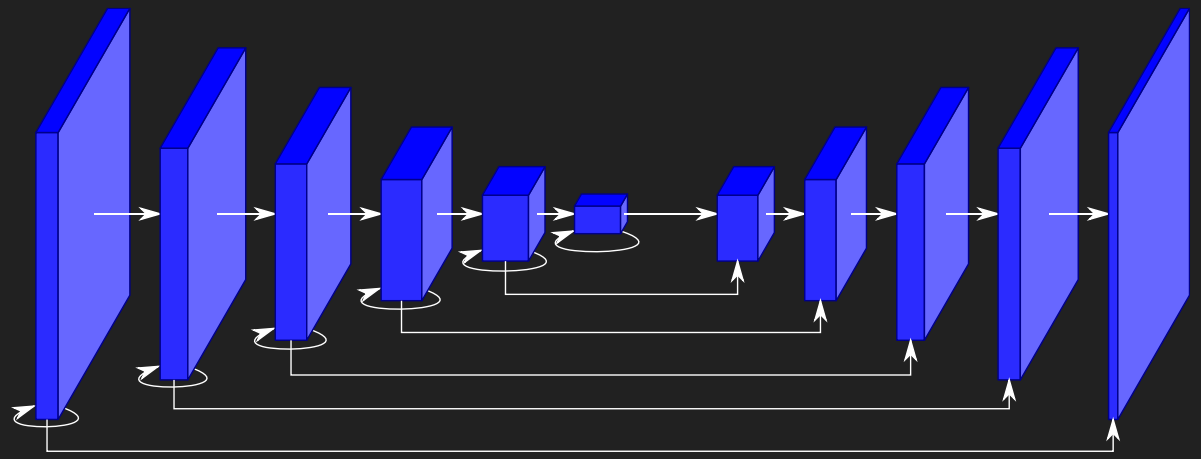
Related Work: Deep Depth From Focus (Hazırbaş et al.)

- DDFFNet
  - VGG16 based encoder-decoder
  - Static focus stacks
  - Frames are processes independently and finally scored
  - Generalization problems



Related Work: Interactive Reconstruction (Chaitanya et al.)

- Recurrent Autoencoder
  - Encoder-Decoder
  - Recurrent blocks having skip connections to themselves
  - Similar to an RNN but is fully convolutional

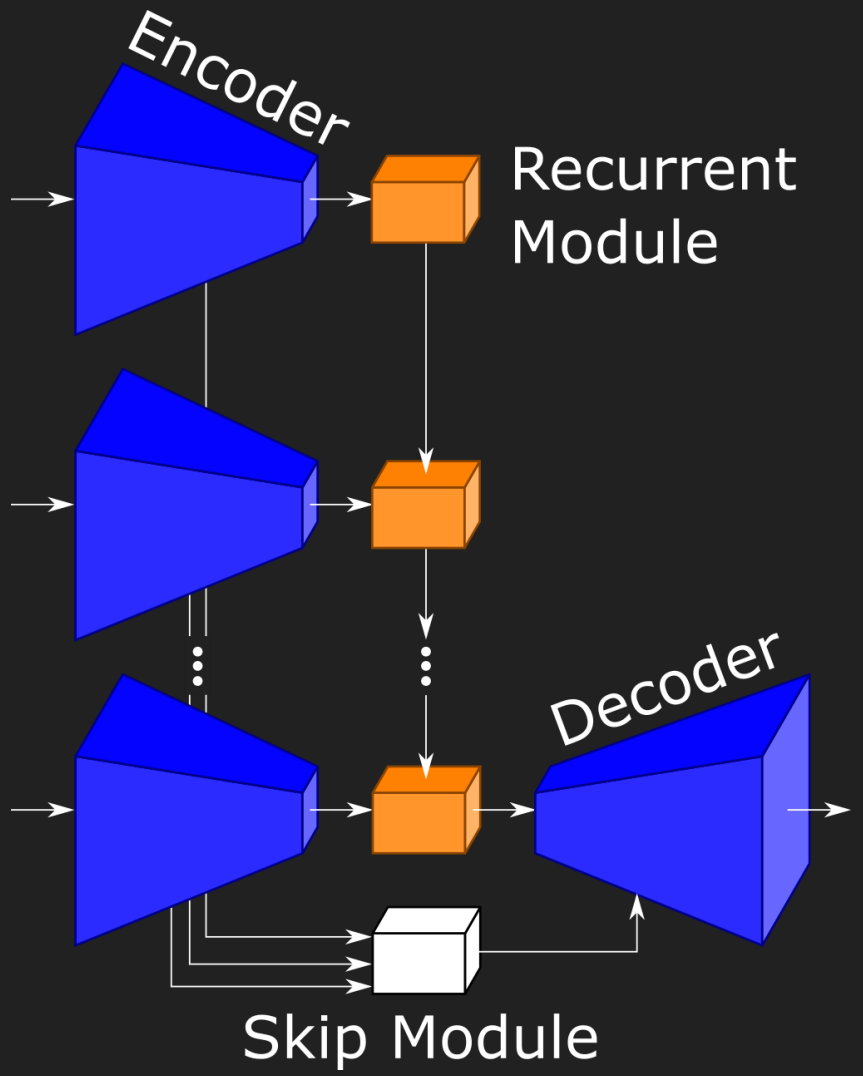


Depth Prediction Approach

- Architectures
  - LSTM-DDFF using the DDFF in joint with an LSTM
  - Modified recurrent autoencoder
    - Batch normalization
    - More feature channels
    - Transposed convolution instead of upsampling
- Extension: Additional decoder for CoC
- Produce one depth map per focus stack

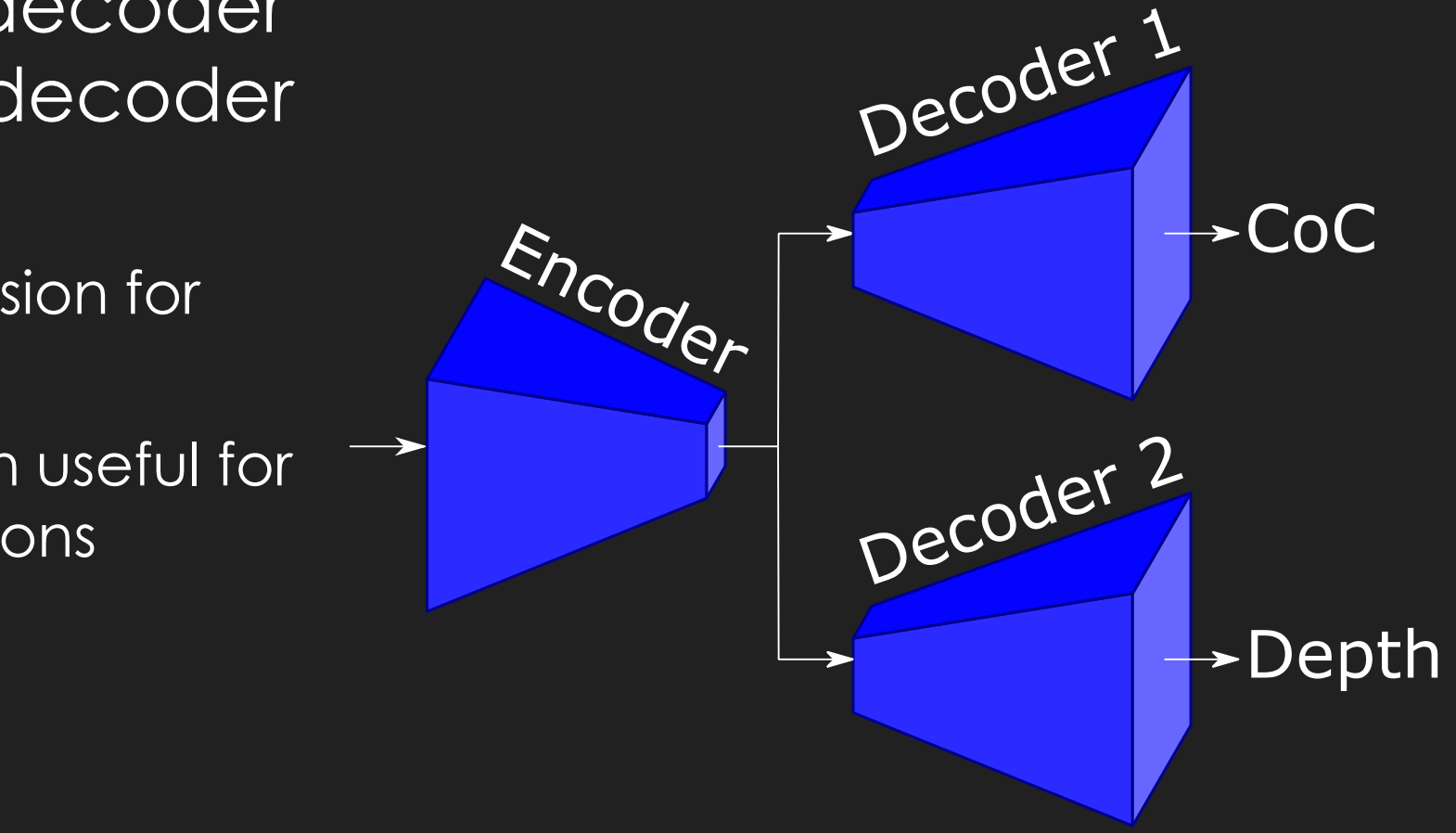
LSTM-DDFF

- DDFF lacks sequential understanding
- Recurrent Module
  - No fully connected layers but long sequence of flattened feature maps
  - Double layered LSTM
- Skip Module: Convolution



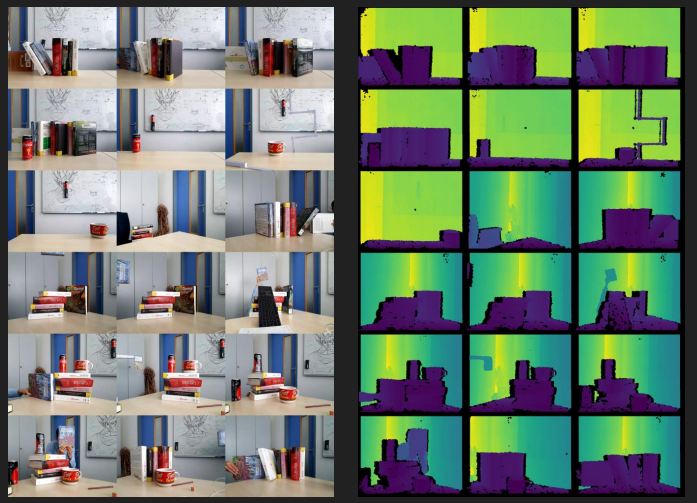
Extension: CoC Supervision

- Add a second decoder to an encoder-decoder system for COC
  - Additional supervision for depth from focus
  - Circle of confusion useful for different applications



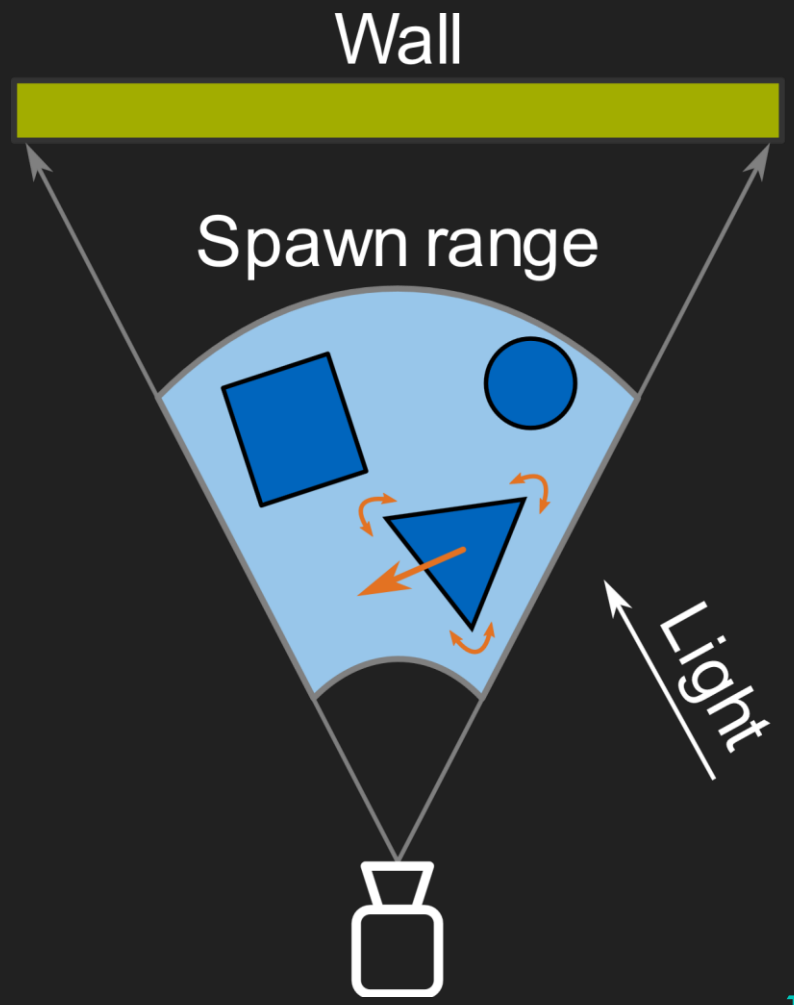
Real-World Dataset

- Samsung Galaxy S7 color recording via an Android application
- Asus Xtion RGB-D sensor for depth ground truth
- Calibration between Android and RGB-D color sensor
- Frame association using common clock
- Limitations: hardware and time



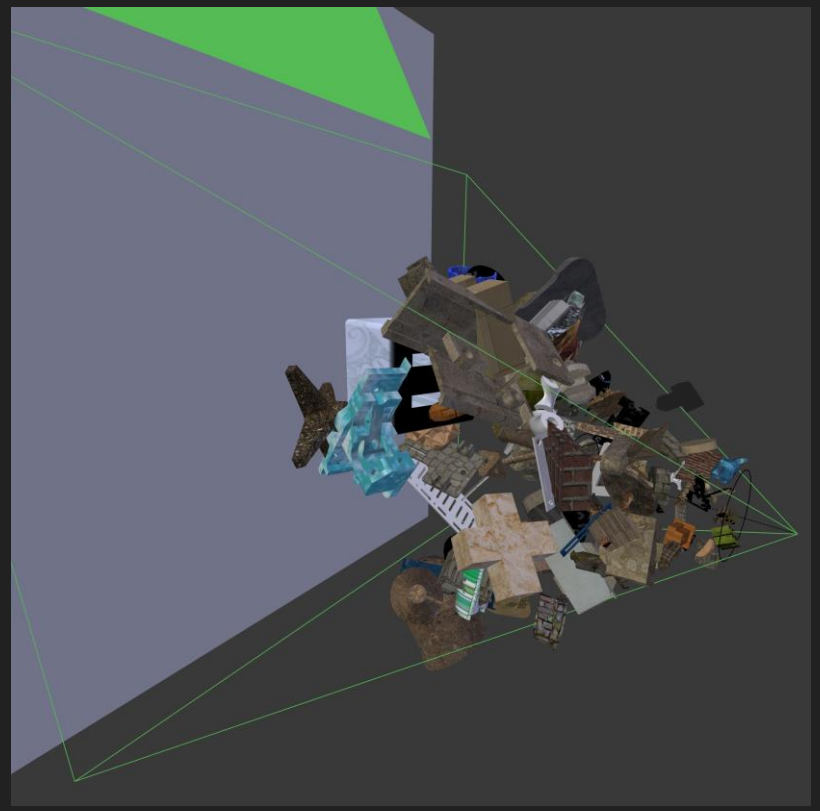
Synthetic Dataset

- CoC domain independent
- Blender for virtual scene rendering
- Physical lens model
- Perfect depth ground truth
- Randomization
  - Objects, textures and light
  - Transformations and animations
- Train-test split object/texture pool

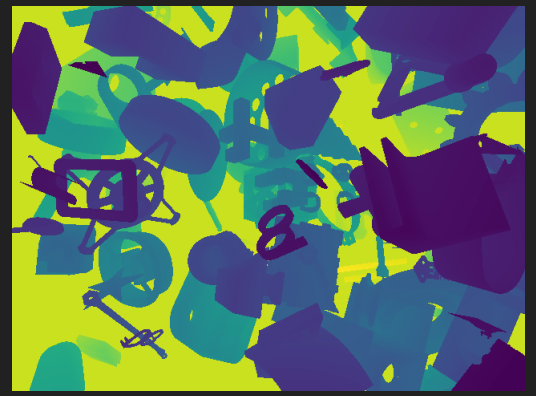


Synthetic Dataset

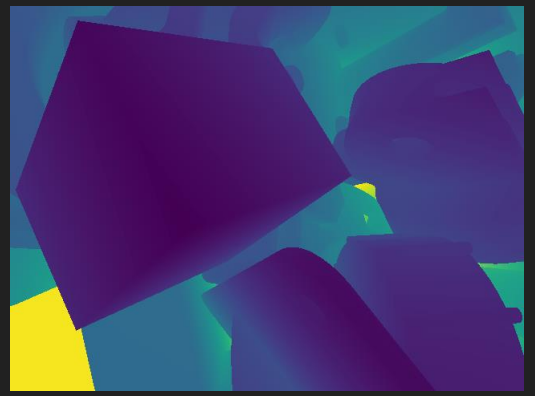
Blender scene



Small obj data



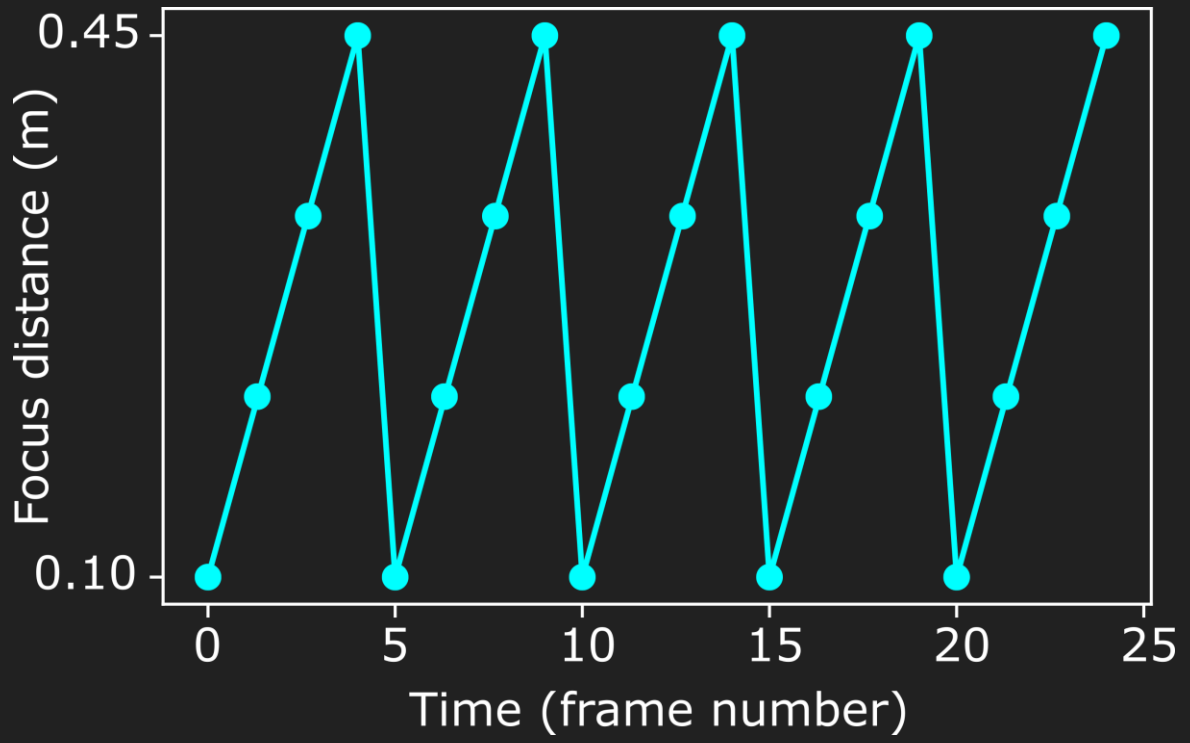
Large obj data



Real-World Dataset: Test Set (Unlabeled)

Carpet dataset

Fast camera motion



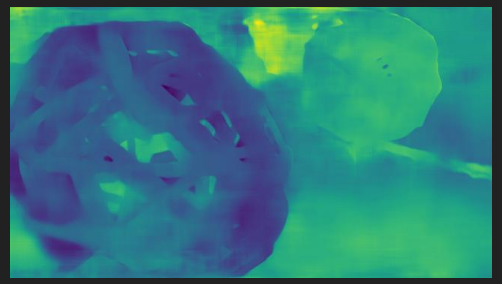
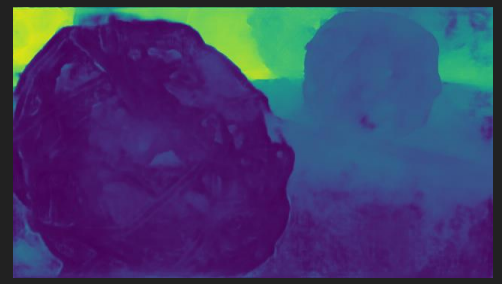
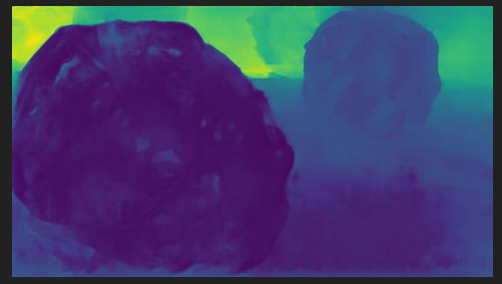
Training

- Using synthetic training set
- Train loss: MSE
- Masked loss to exclude wall
- Input: One focus stack of 4 frames with increasing focus distances
- 256 x 256 random crops

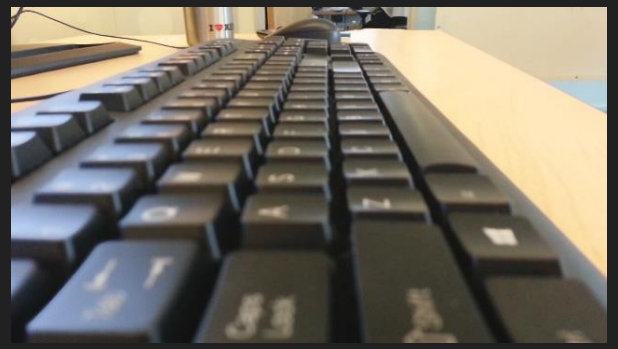
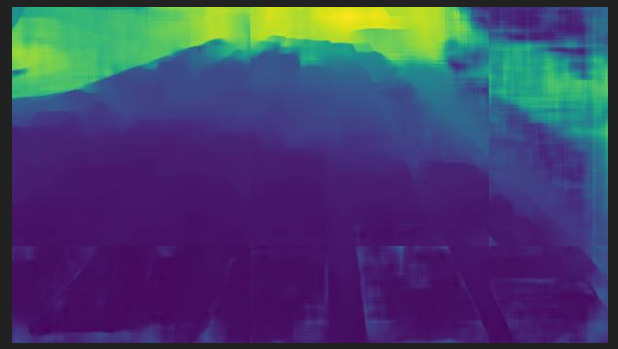
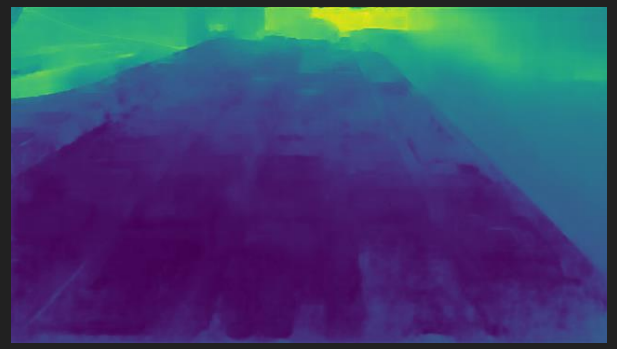
Synthetic Blender Test Set Metrics

Model	MSE	RMS	Accuracy
DDFF Small Obj	1.8×10^{-3}	0.041	86.78
LSTM-DDFF Small Obj	**1.1×10^{-3}**	**0.033**	**94.98**
Recurrent AE Small Obj	3.4×10^{-3}	0.058	69.02
Recurrent AE Small & Large Obj	3.2×10^{-3}	0.056	70.77

Results Static (Suwajanakorn)

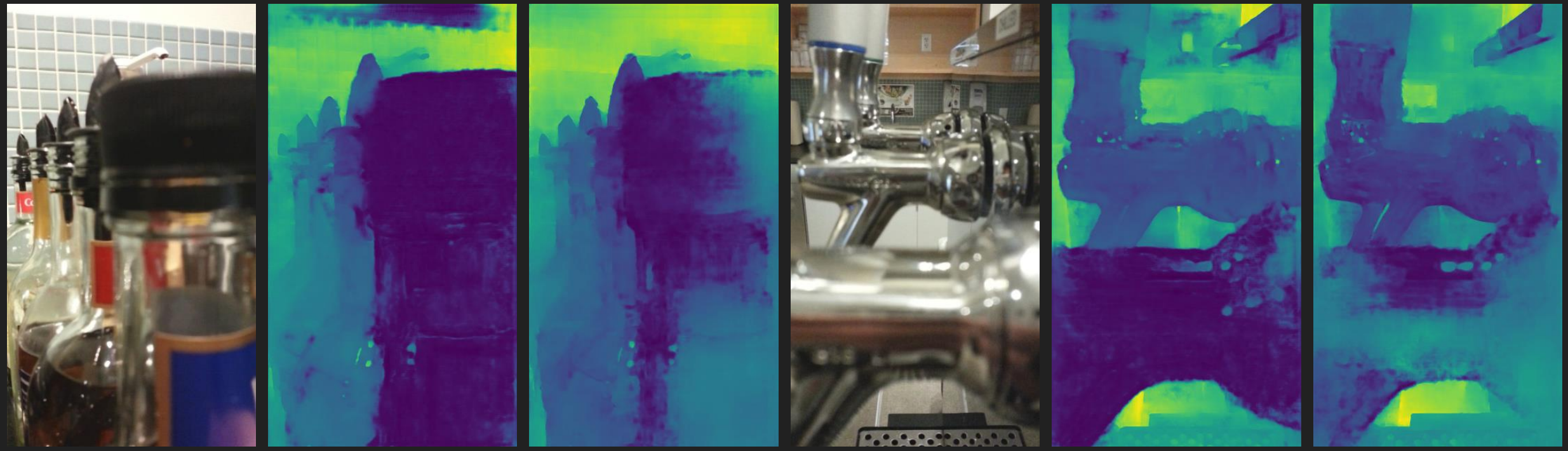
Input	DDFF Small Obj	LSTM-DDFF Small Obj	Recurrent AE Small Obj	Recurrent AE Small & Large
				
				

Results Static (Suwajanakorn)

Input	LSTM-DDFF Small Obj	Recurrent AE Small Obj	Recurrent AE Small & Large Obj
			
			

Results Static (Suwajanakorn): Recurrent Autoencoder

Input	Small Obj	Small & Large Obj	Input	Small Obj	Small & Large Obj



Results Dynamic (Carpet): Recurrent Autoencoder

Input	Recurrent AE Small Obj	Recurrent AE Small & Large Obj
		

Discussion

- Proposed architectures show good real-world generalization
- Synthetic test metrics and real-world performance differs
- Recurrent autoencoder produces best results even in challenging scenarios
- LSTM-DDFF produces seams
- DDFF does not generalize
- Larger data sometimes gives improvement

Limitations

- No best configuration
- Depth from focus issues with
  - Reflections
  - Transparency
  - Uniform coloring
- No confidence for predictions

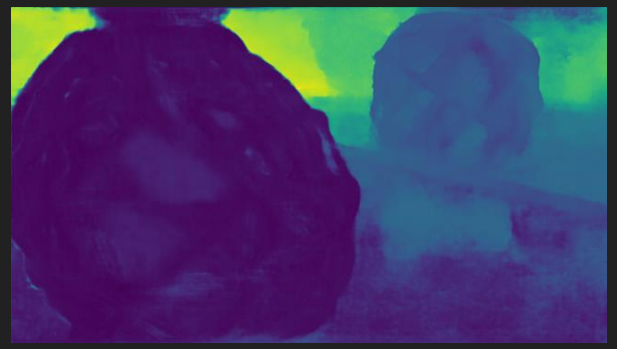
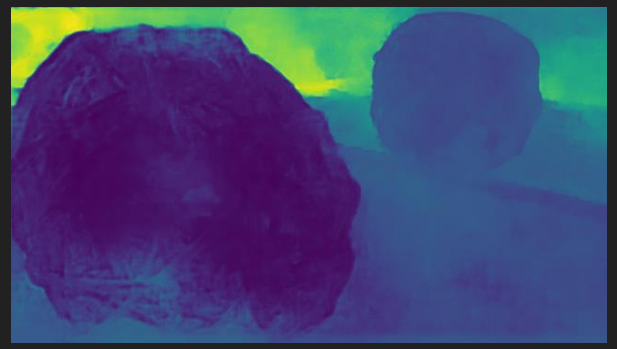
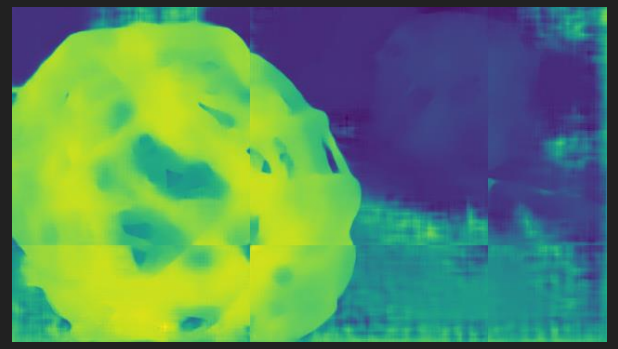
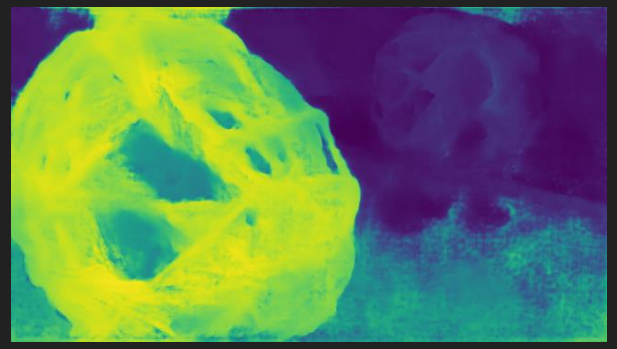
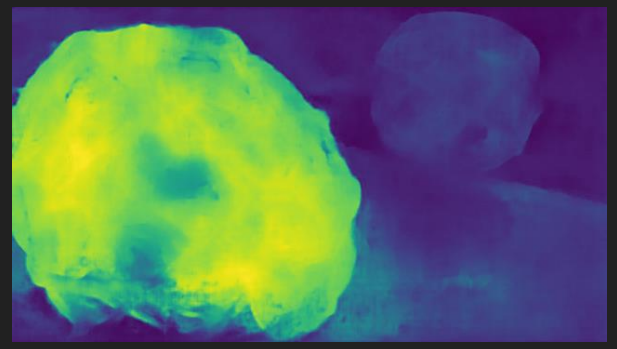
Conclusion

- Modified recurrent AE as a simple and lightweight depth estimator
- Real-world dataset hard to acquire
- Training on synthetic dataset is superior
- Future directions
  - Optimize training data and recurrent autoencoder
  - Recurrent AE interesting for structure from motion

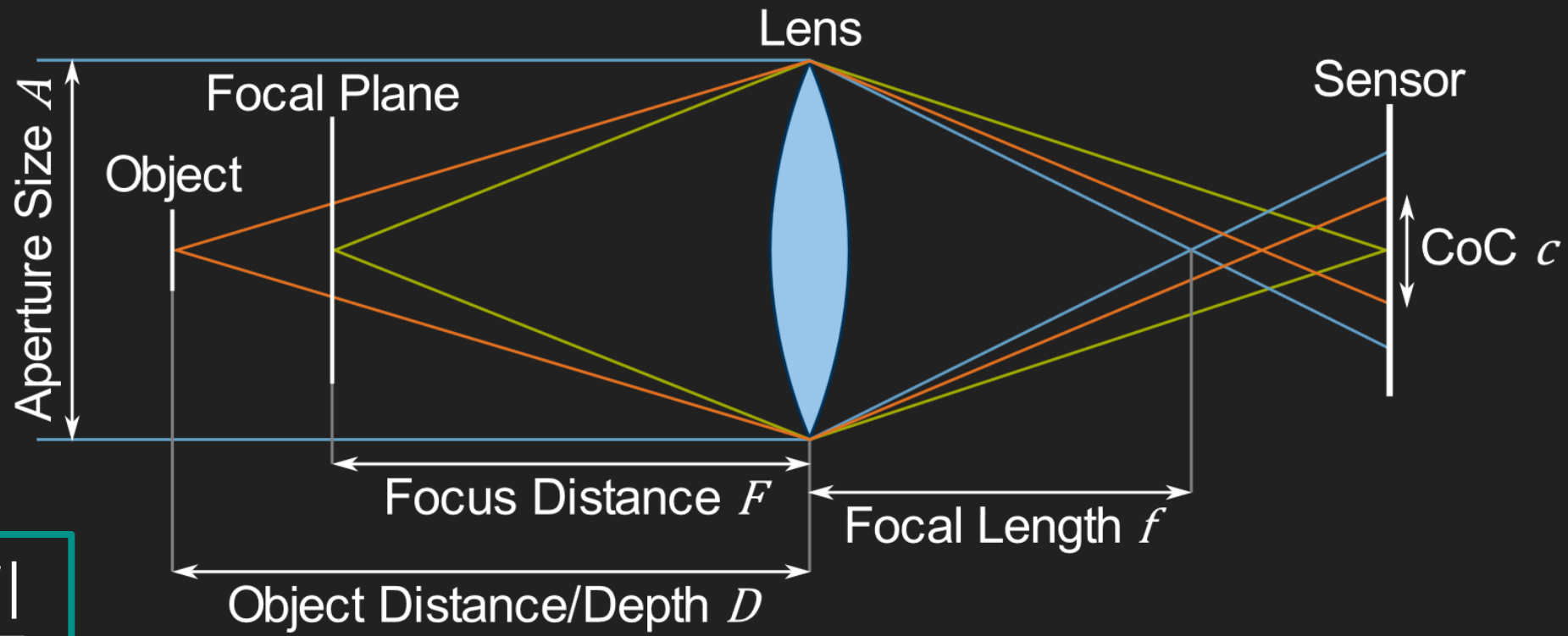
    → Combine with CoC for superior depth prediction

The End

Results Static (Suwajanakorn): CoC Supervision

Input	LSTM-DDFF Small Obj	Recurrent AE Small Obj	Recurrent AE Small & Large Obj
			
			

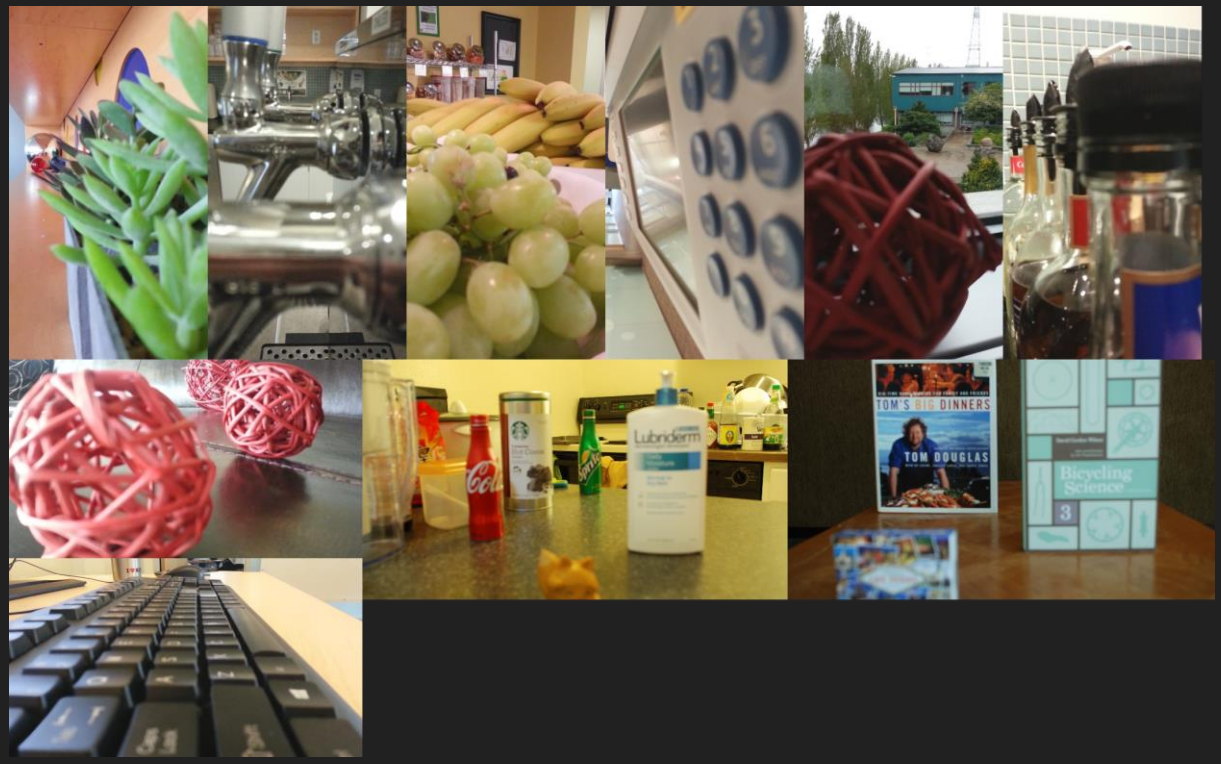
Circle of Confusion (CoC)



$$c = \frac{Af|D - F|}{D(F - f)}$$

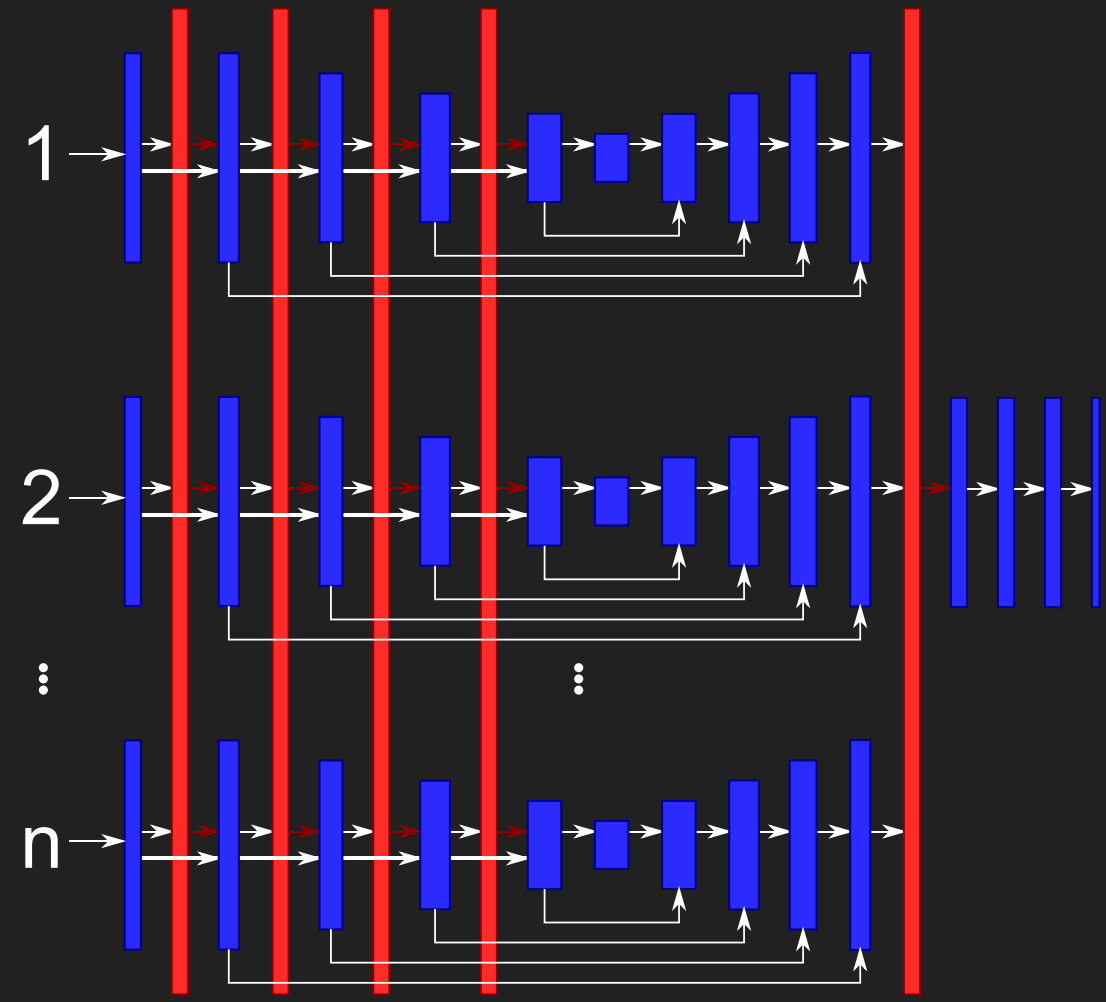
Related Work: Depth from Focus (Suwajanakorn et. al)

- Variational approach for static focus stacks
- Perfectly static dataset

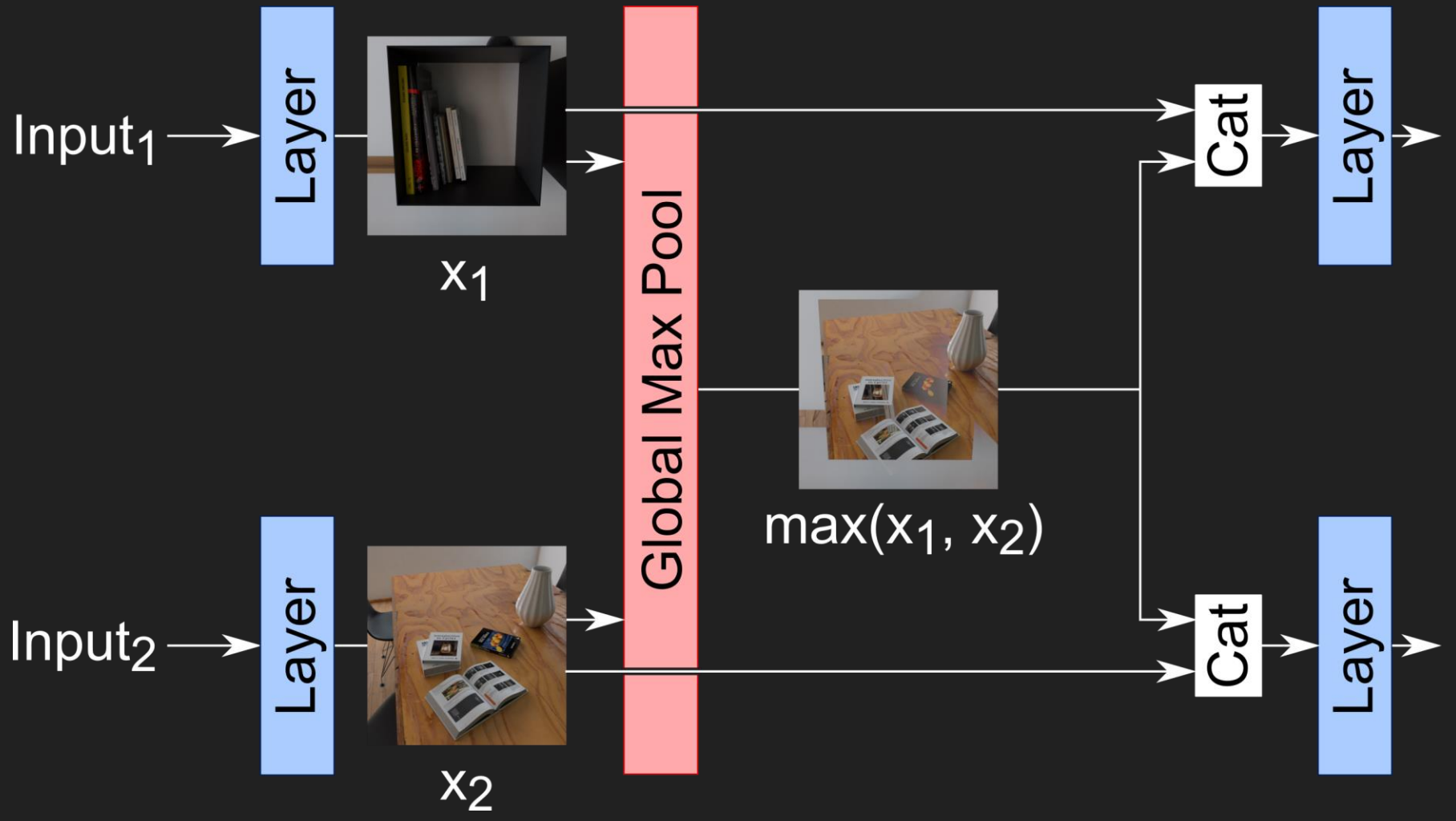


Related Work: Burst Image Deblurring (Aittala et. al)

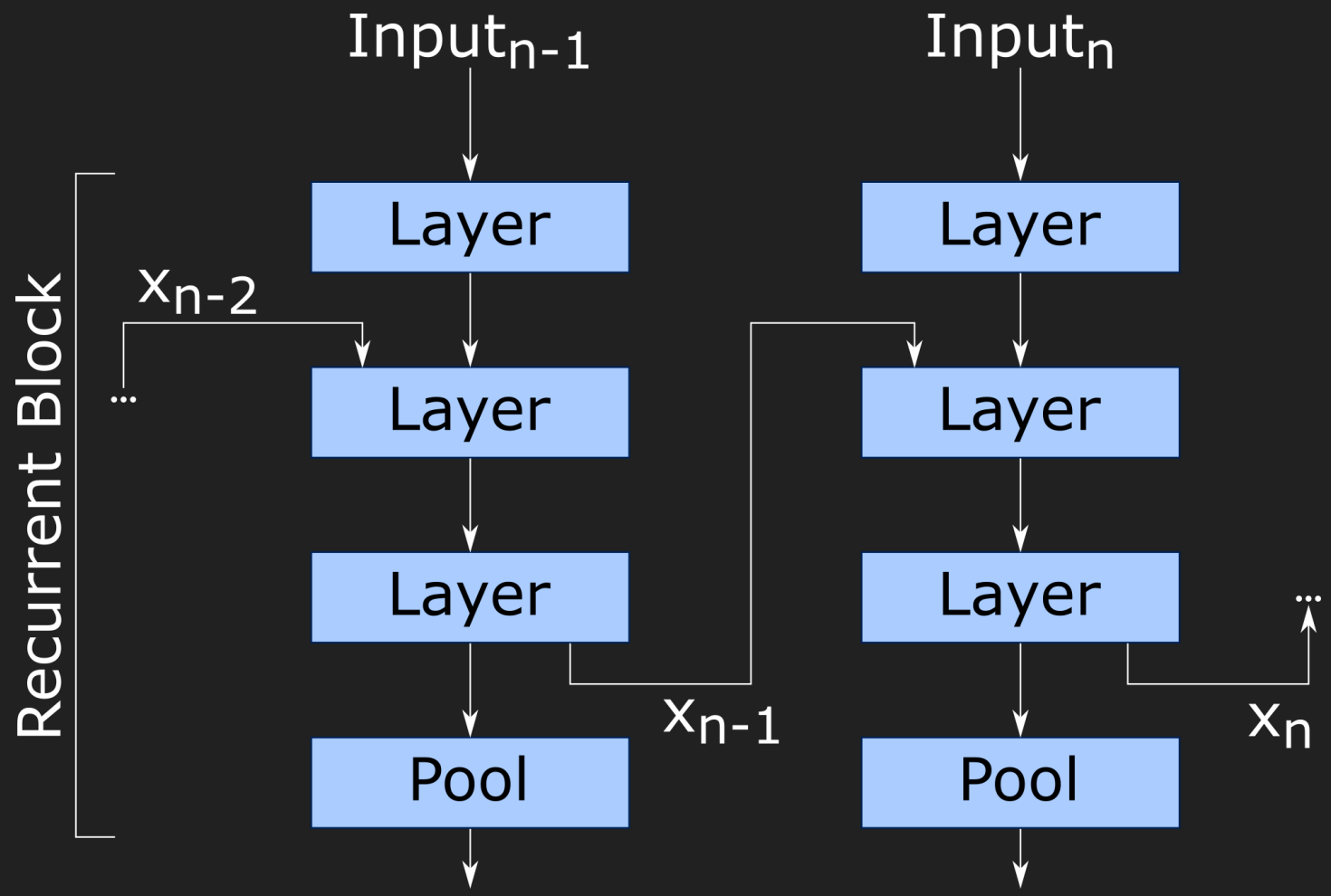
- Permutation Invariant CNN
  - Encoder-Decoder
  - Sequence input
  - Order invariant
  - Global max-pool layers for information exchange



Related Work: Burst Image Deblurring (Aittala et. al) - Global Max-Pooling

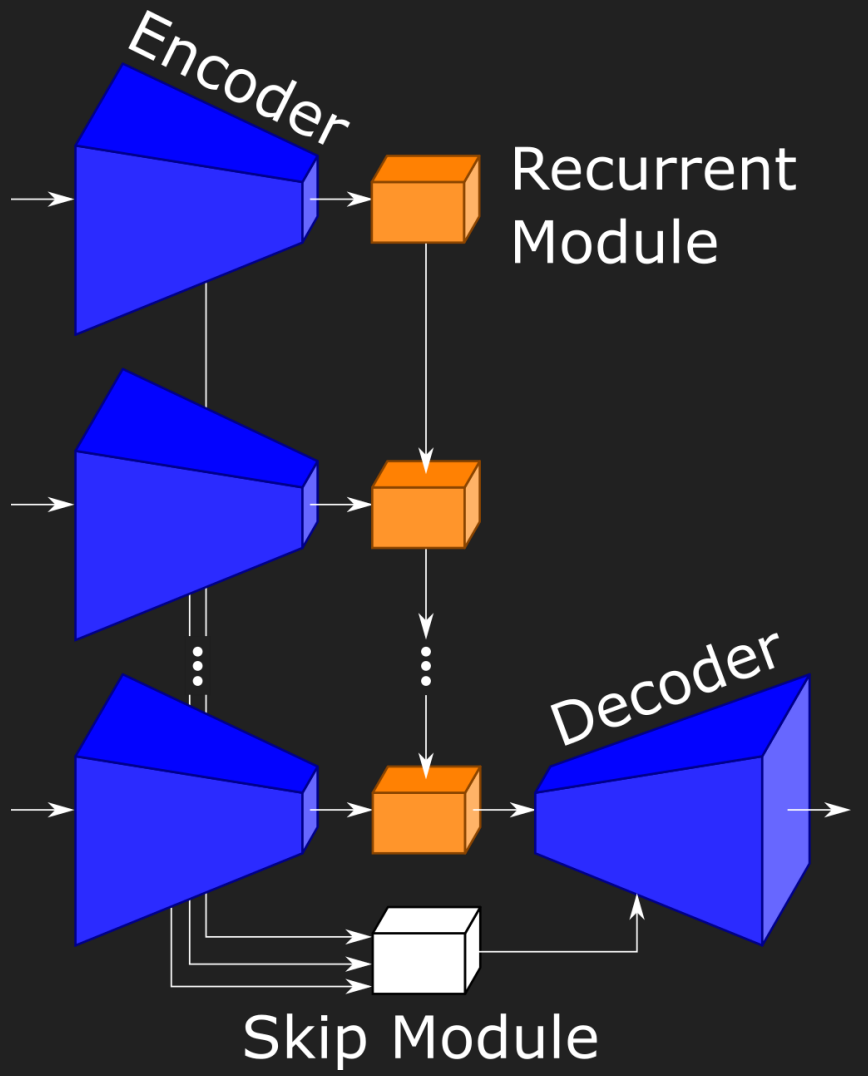


Related Work: Interactive Reconstruction (Chaitanya et al.) - Recurrent Block



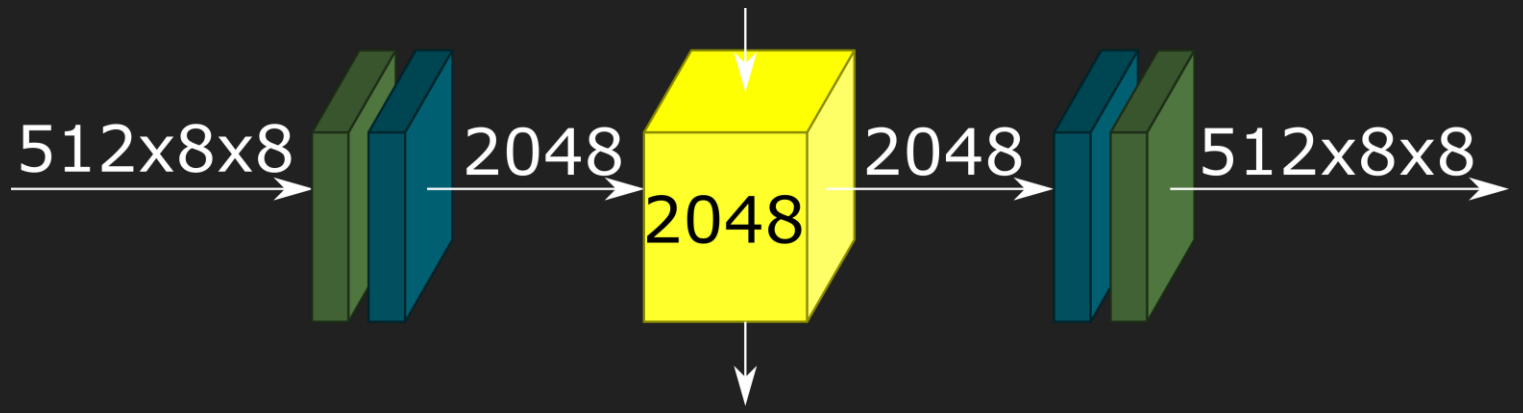
LSTM-DDFF

- DDFF lacks sequential understanding
  - → Use LSTM
- Recurrent Module
  - Flattening and dimension reduction
  - LSTM
- Skip connection aggregation
  - Last encoder only
  - Convolution

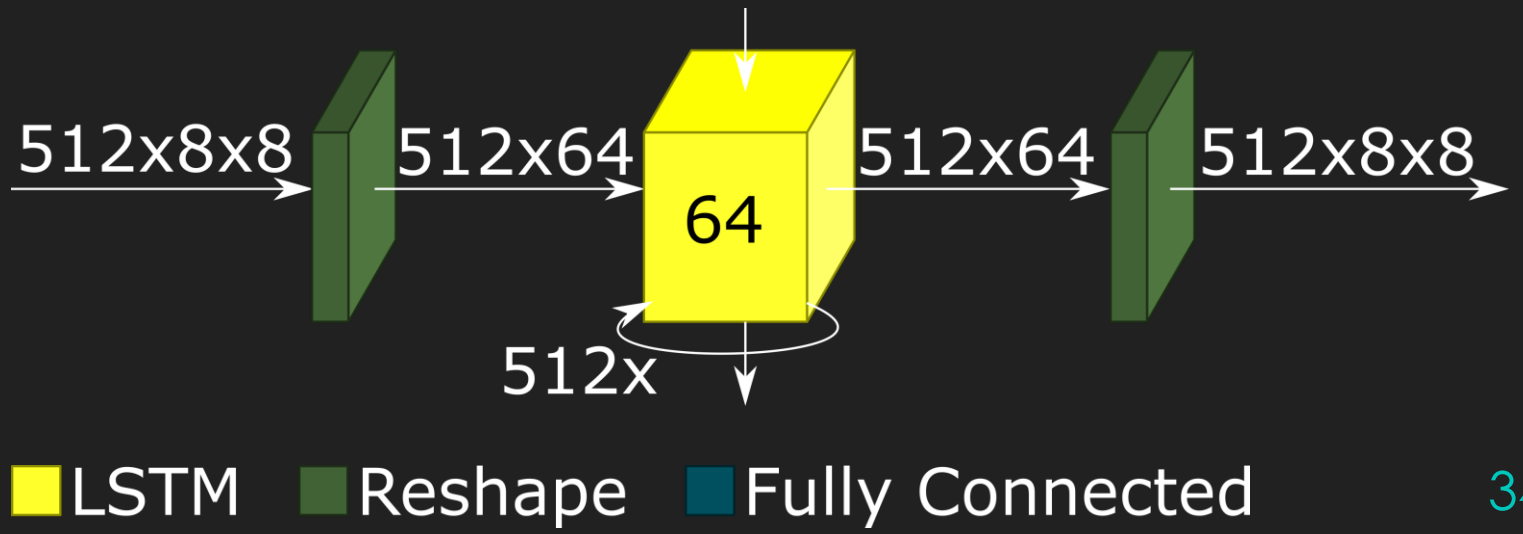


LSTM-DDFF: Recurrent Module Types

- Fully connected
  - Flatten C x H x W
  - Large LSTM dimension



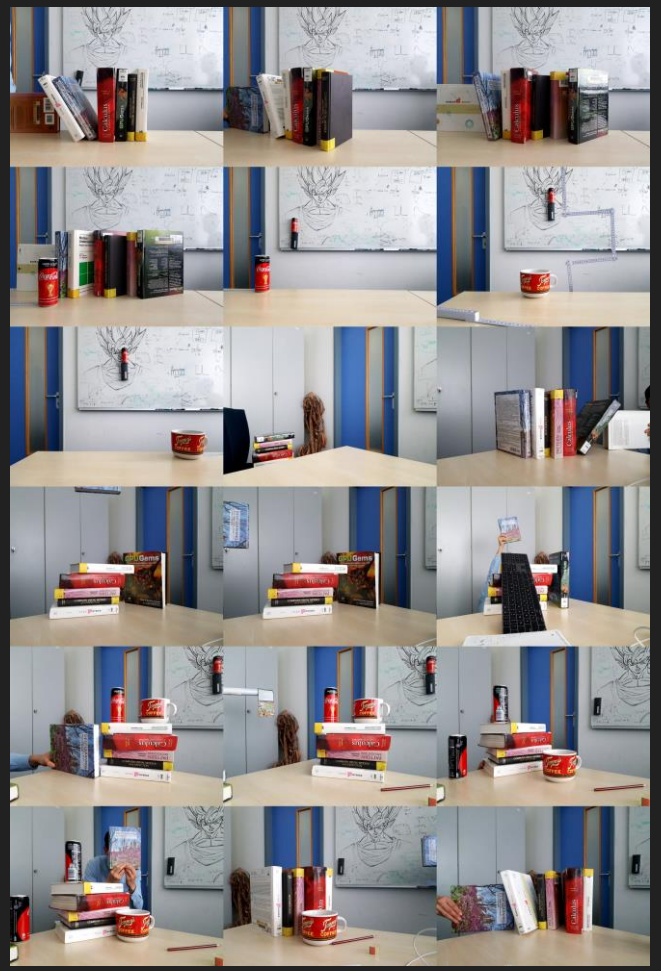
- Stacking
  - Flatten H x W
  - Long sequence
  - Double layered LSTM

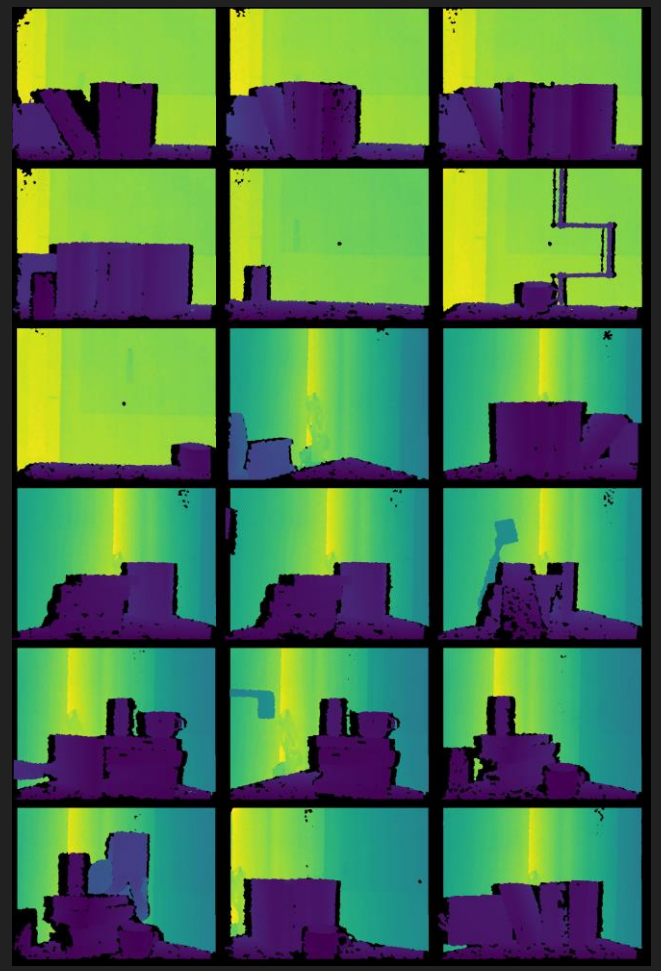


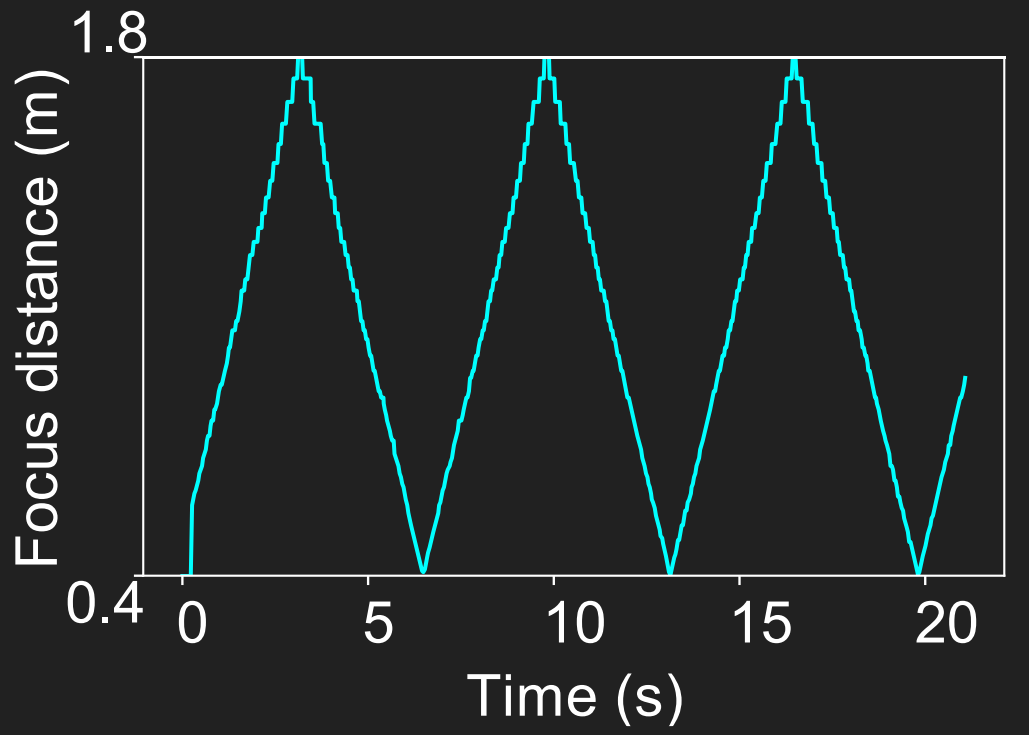
PoolNet

- Modification of the Permutation Invariant CNN
  - Batch Normalization
  - Less feature channels
  - No global pooling in the decoder
- Variations
  - Giving focus distance as channel for each frame
  - Two PoolNets in sequence
    - Predict CoC maps first for all inputs (leave out last pooling)
    - Then predict depth using predicted CoC

Real-World Dataset: Train Set





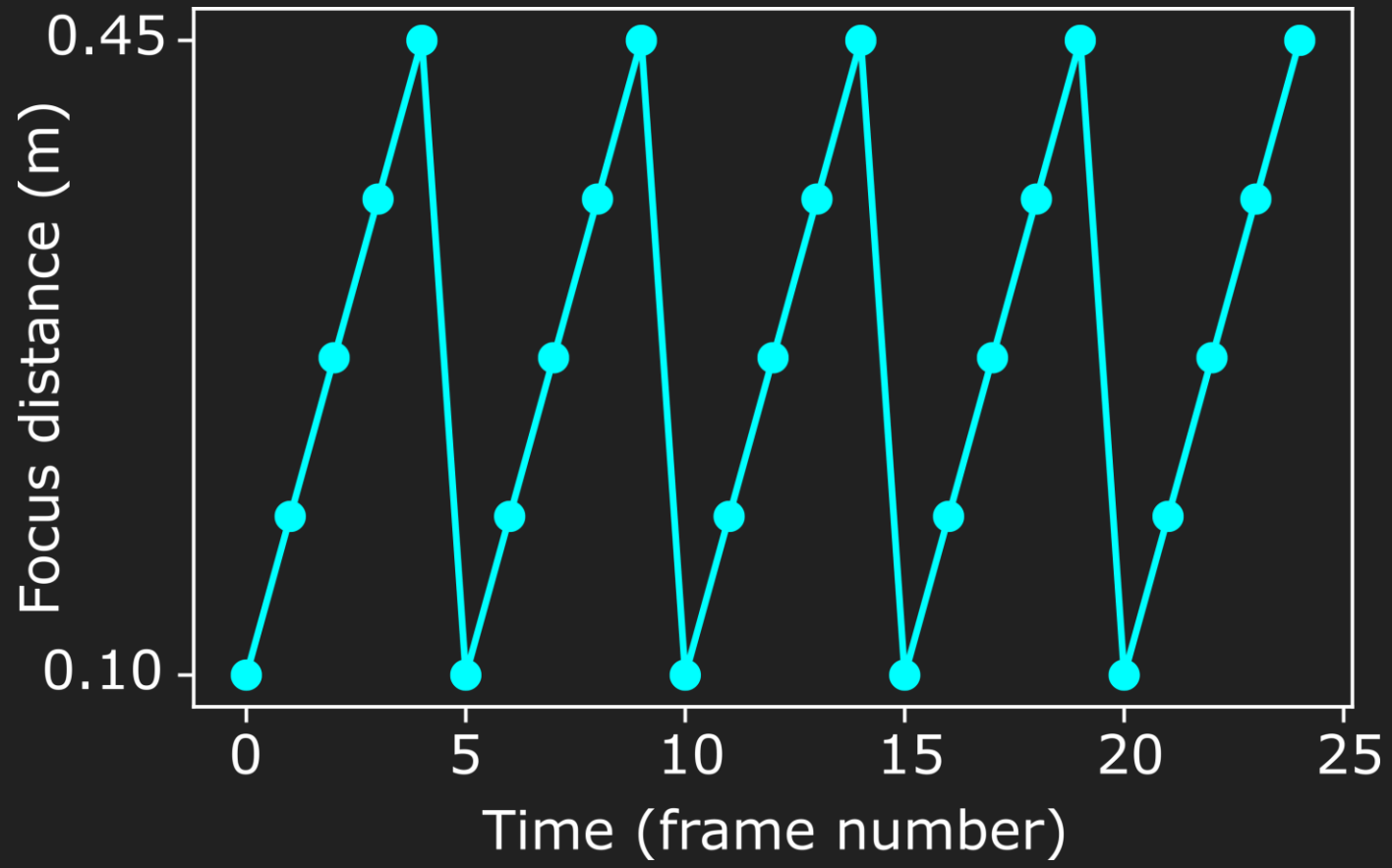


Real-World Dataset: Train Set Limitations

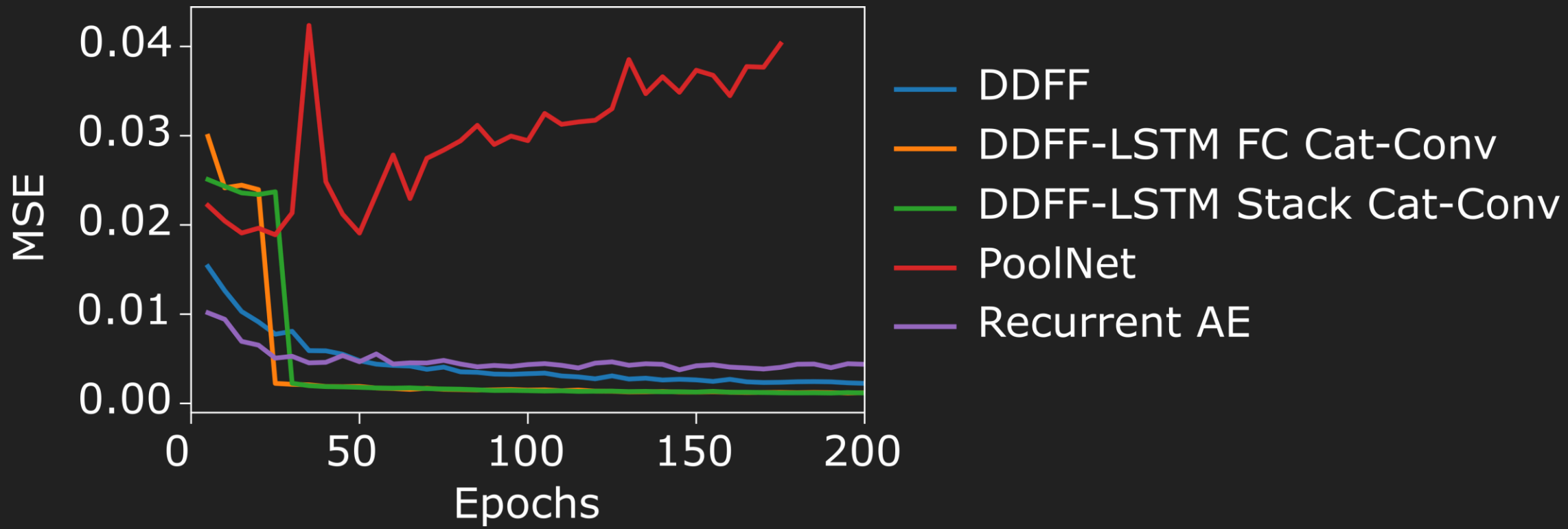
- CoC hardly noticeable
  - RGB-D sensor works only for larger distances
  - Mobile lenses are small and have a large DoF for larger distances
- Invalid depth values
  - Reflections
  - Occlusions
- Time consuming

Synthetic Dataset

- 5 ramps consisting of 5 frames
- No ramp from max to min
- Android phone also capable to jump to any focus distance without delay



Training: Validation Loss



Results Dynamic (Carpet): Recurrent Autoencoder

Input	Recurrent AE Small Obj	Recurrent AE Small & Large Obj