An intriguing failing of convolutional neural networks and the CoordConv solution

Rosanne Liu[1] rosanne@uber.com

Joel Lehman[1] joel.lehman@uber.com

Piero Molino[1] piero@uber.com

Felipe Petroski Such[1] felipe.such@uber.com

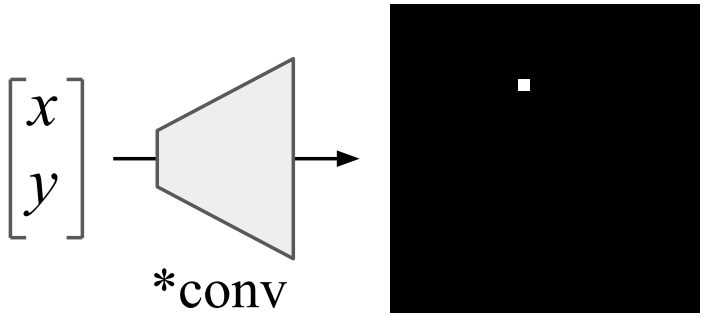
Eric Frank[1] mysterefrank@uber.com

Alex Sergeev[2] asergeev@uber.com

Jason Yosinski[1] yosinski@uber.com

To be presented at NIPS 2018

Toy Tasks

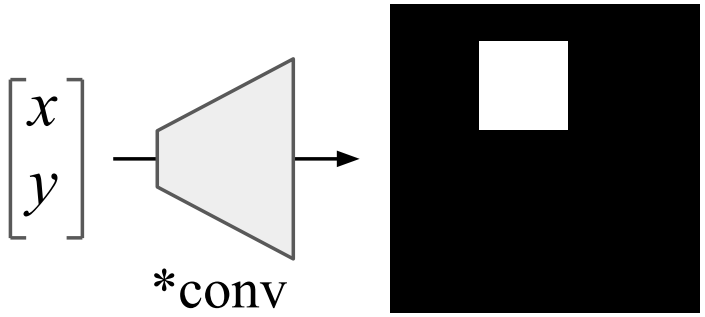


Supervised Coordinate Classification

Output: softmax Loss: cross entropy

(With reverse arrow) **Supervised Coordinate Regression**

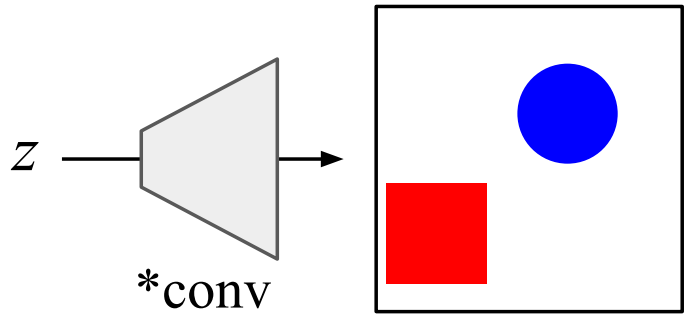
Output: linear Loss: MSE



Supervised Rendering

Output: per-pixel sigmoid

Loss: cross entropy

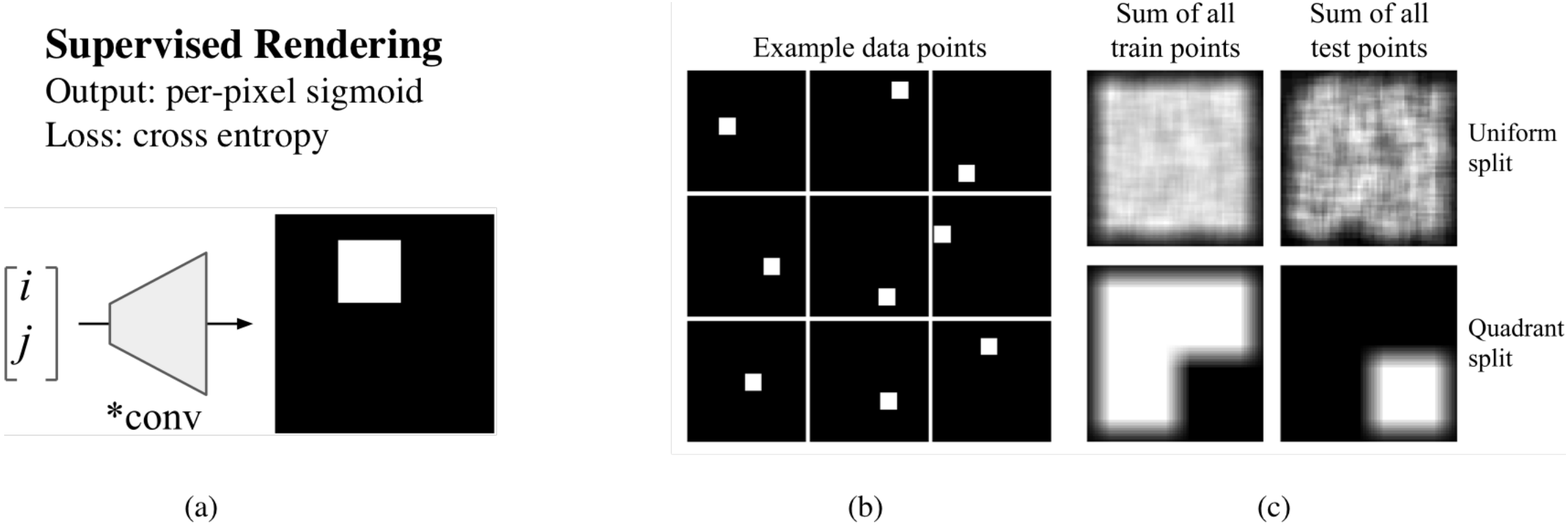


Unsupervised Density Learning (GAN)

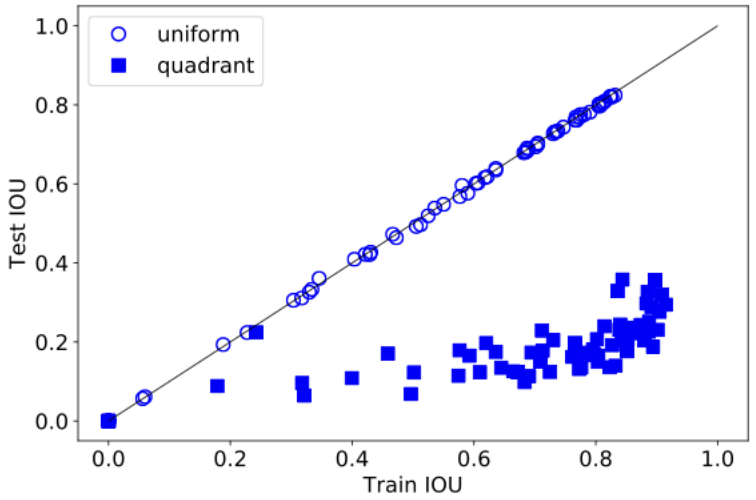
Output: per-pixel, per-channel sigmoid

Loss: learned GAN discriminator

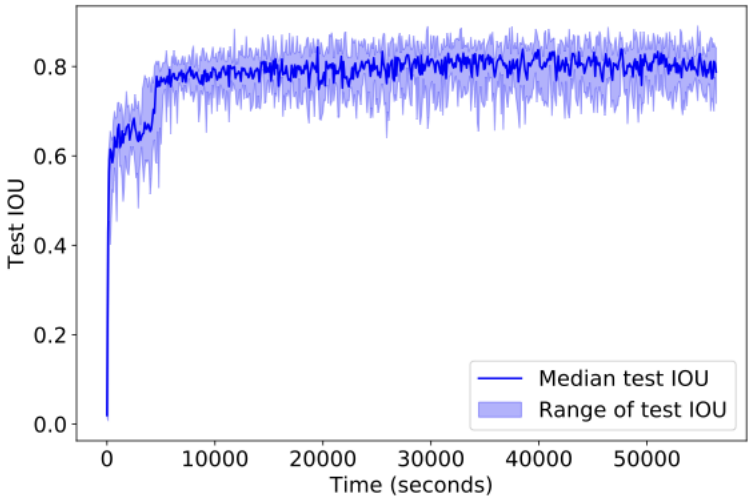
Supervised Rendering



Surprisingly Difficult



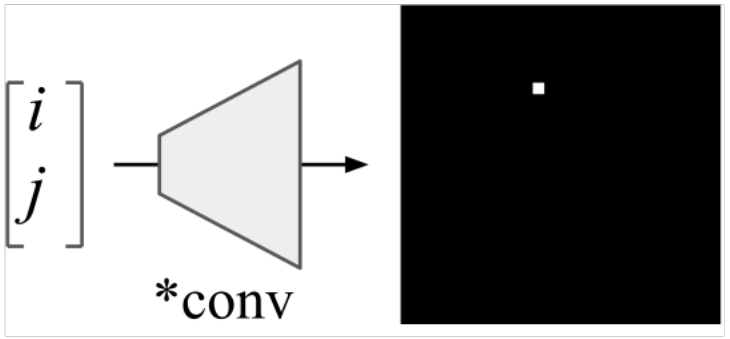
(a)



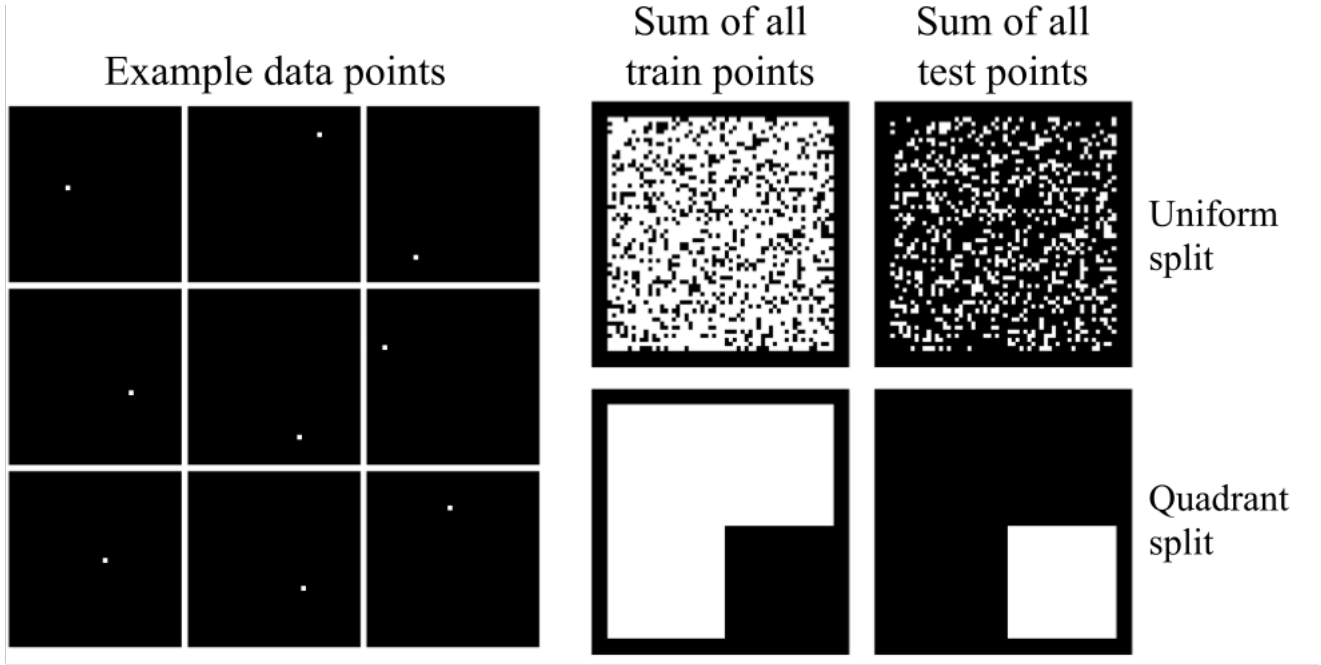
(b)

A little simpler problem

Supervised Coordinate Classification
Output: softmax
Loss: cross entropy

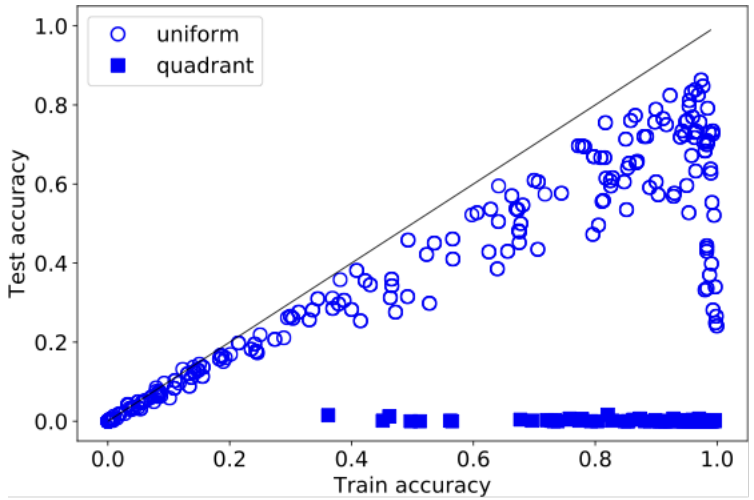


(a)

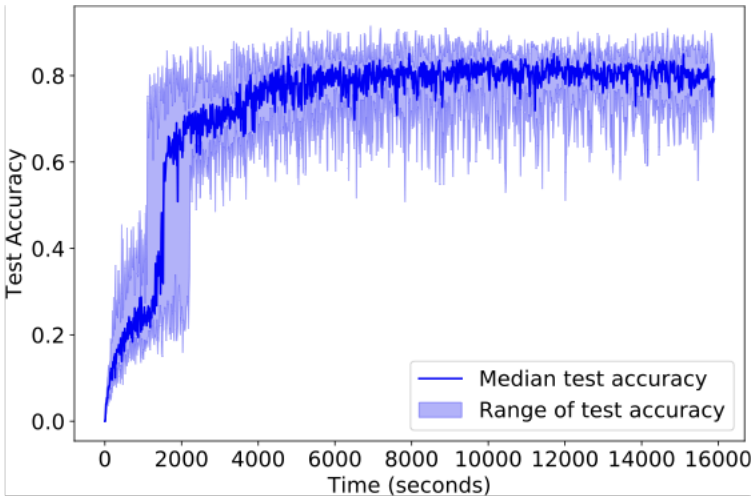


(b) (c)

How is the performance now?

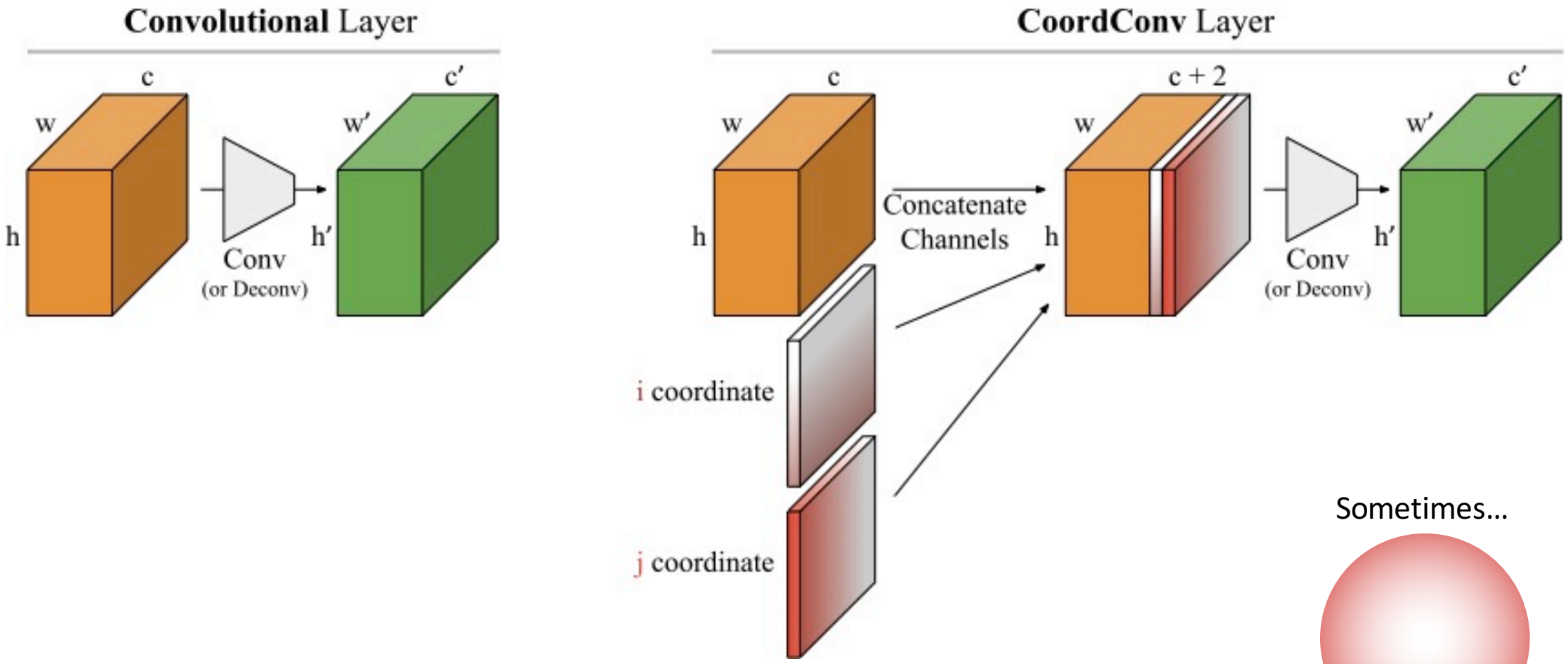


(a)



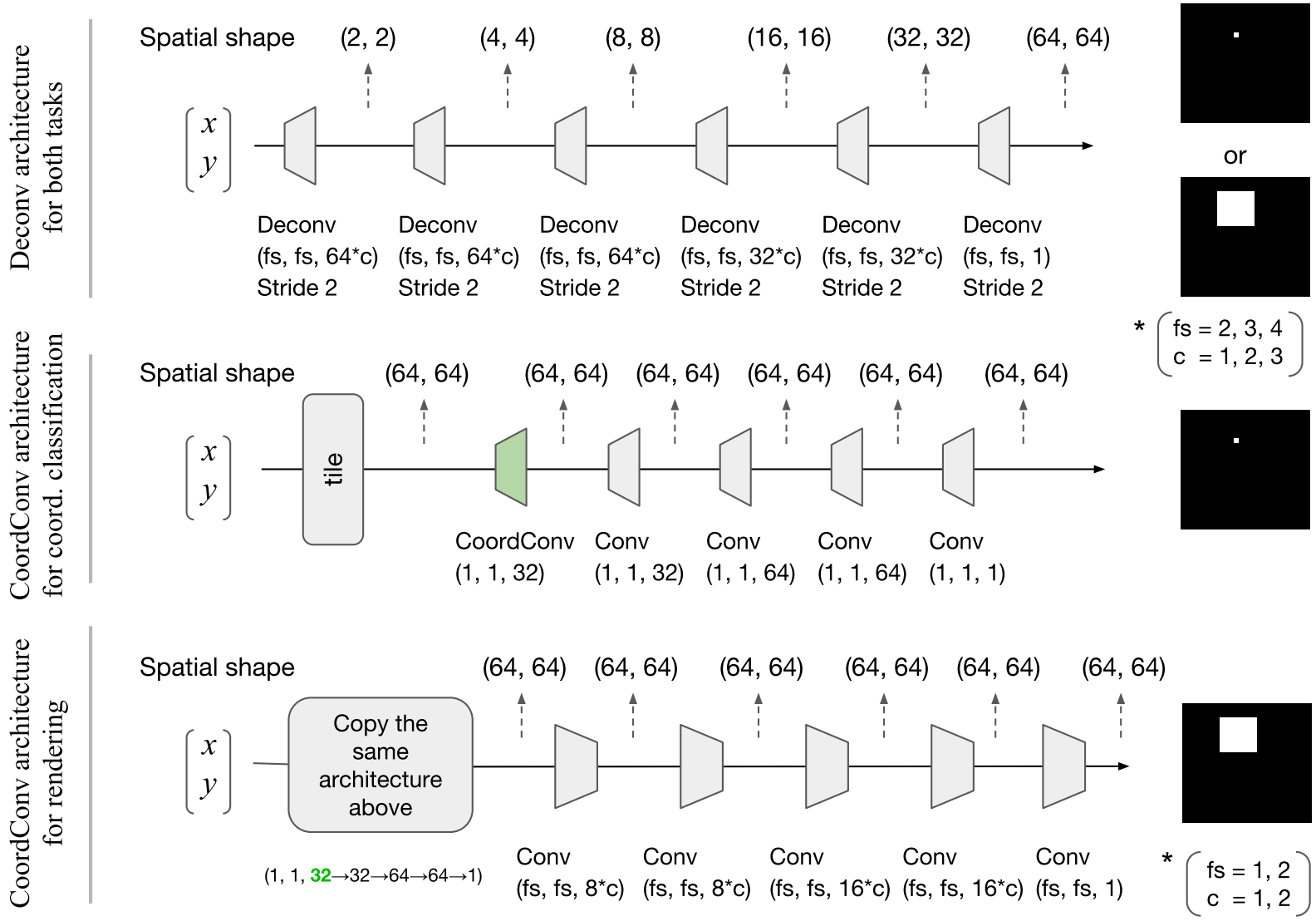
(b)

Solution → CoordConv



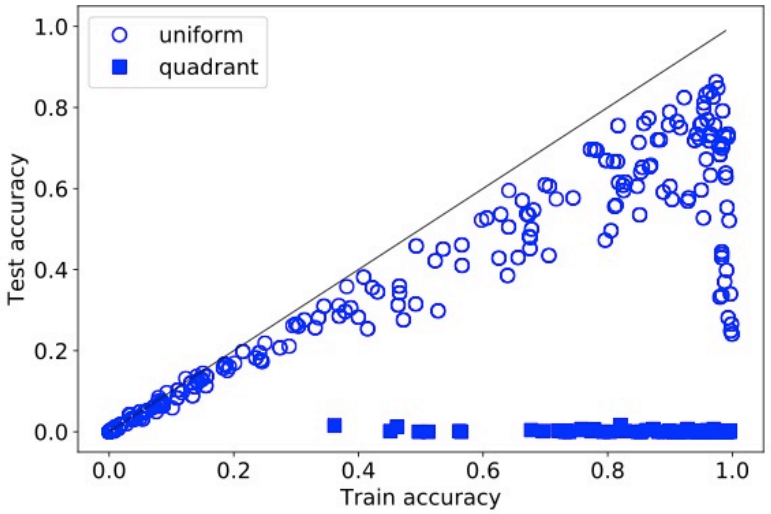
Sometimes…

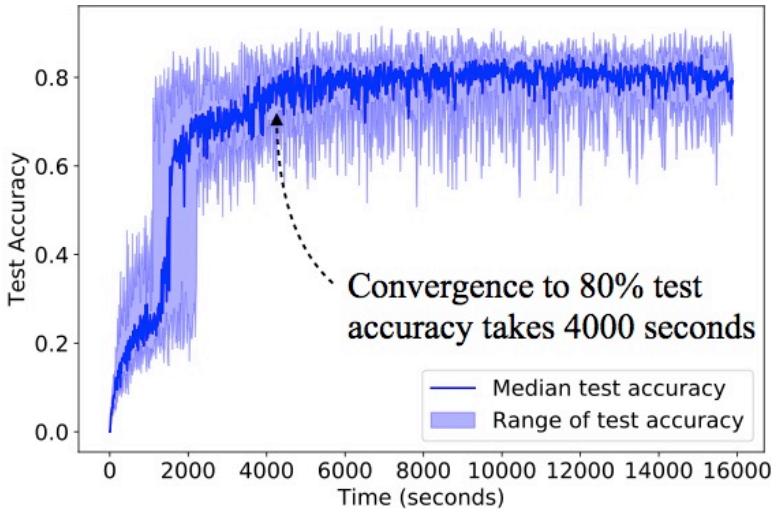
Network Architecture



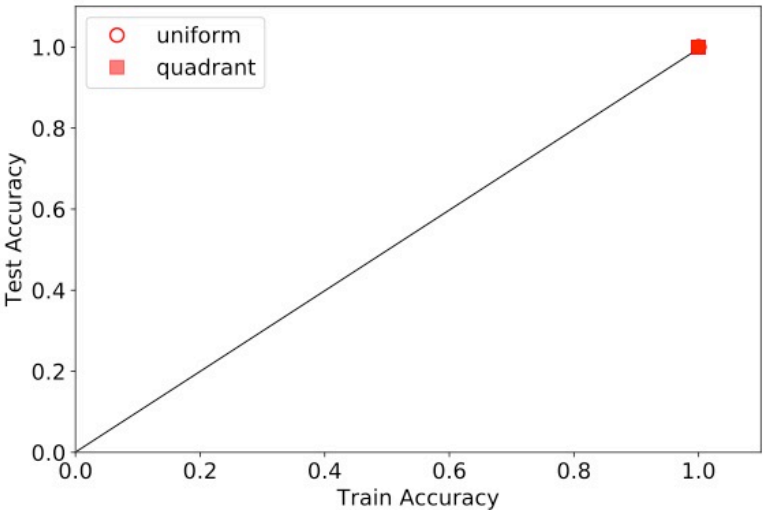
Performance on the toy tasks

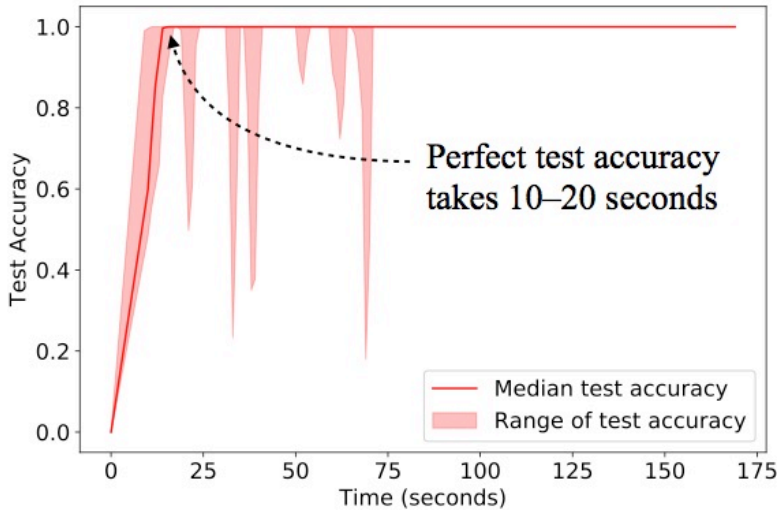
Convolution



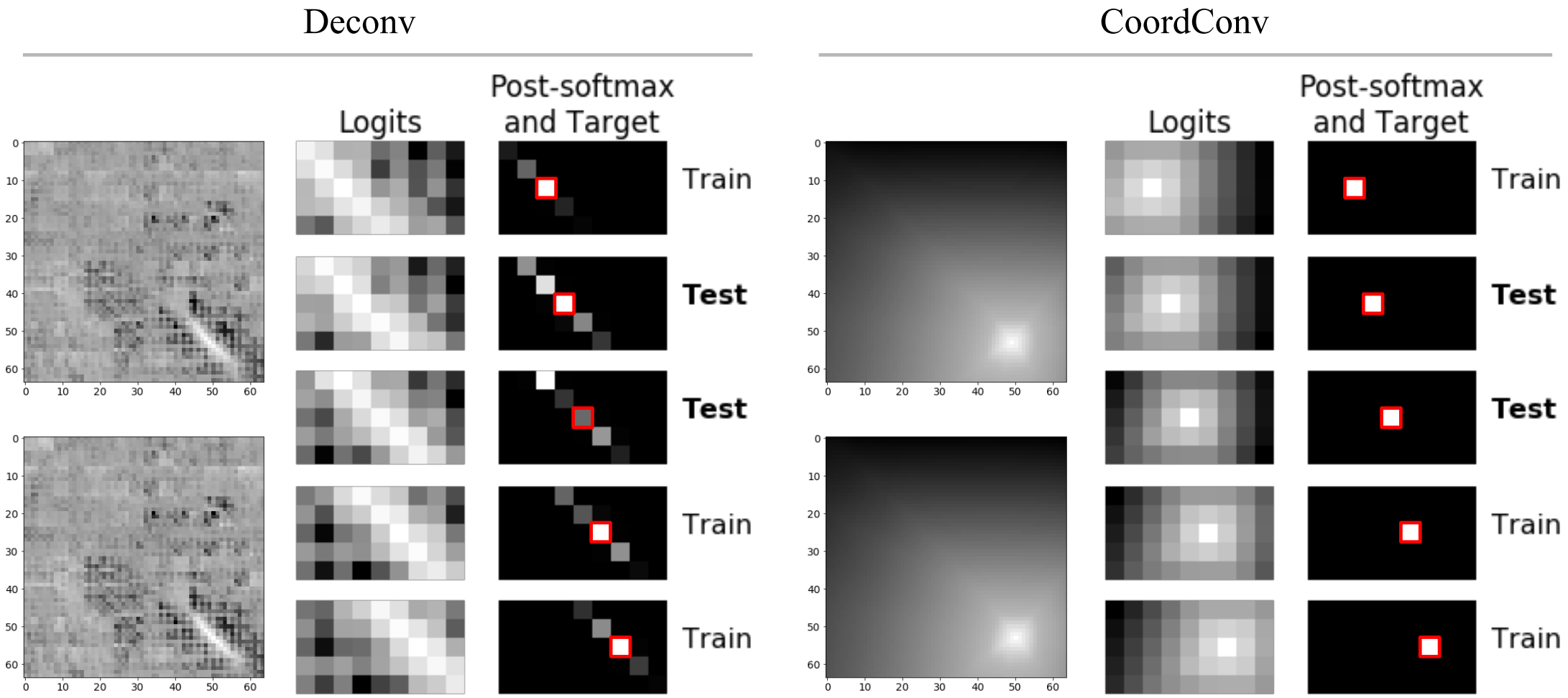


CoordConv

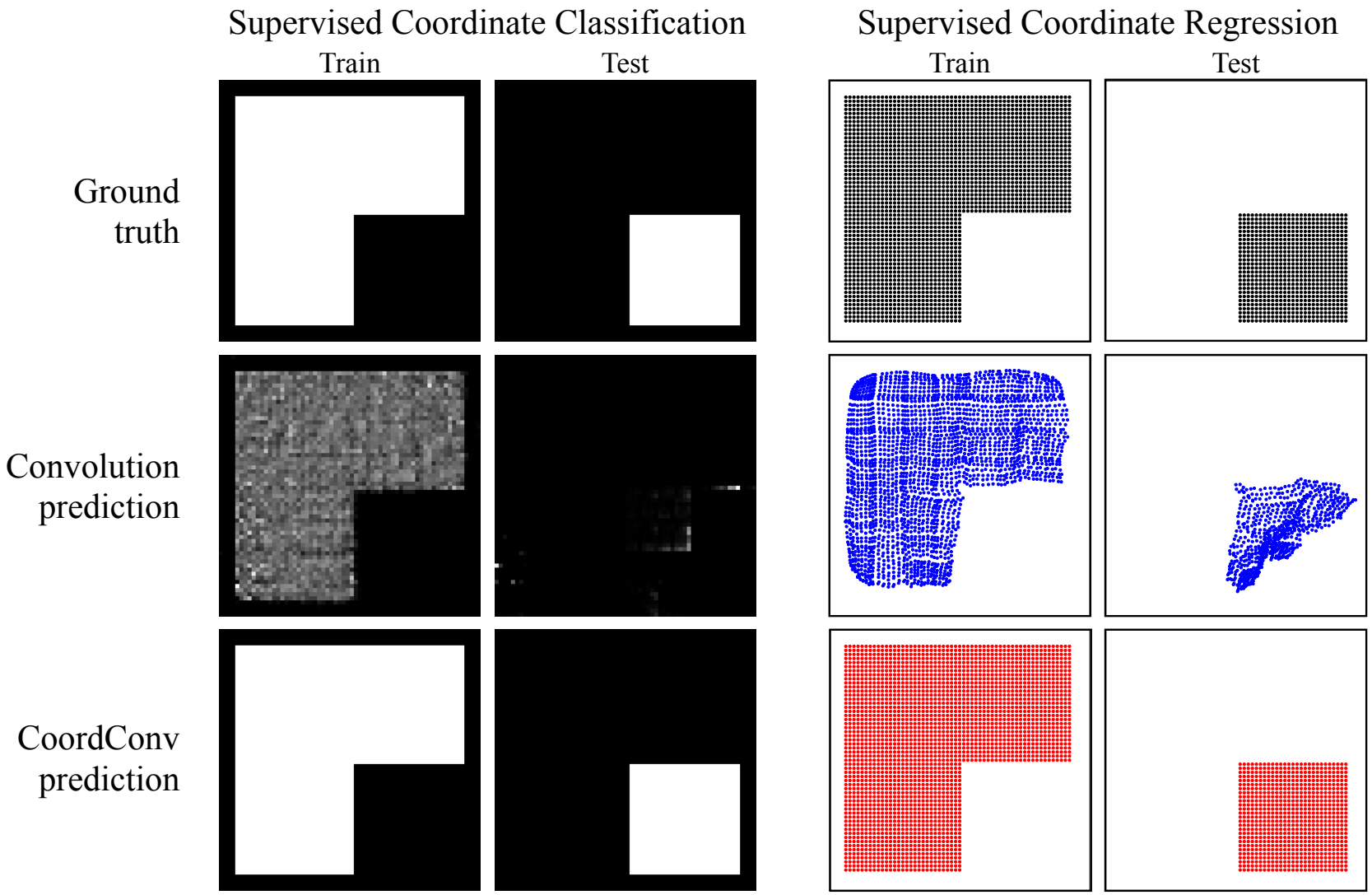




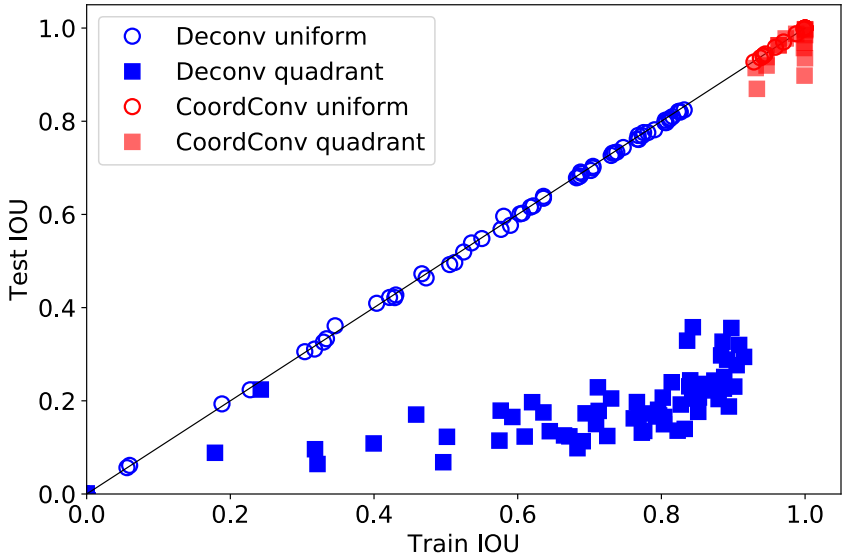
New Results

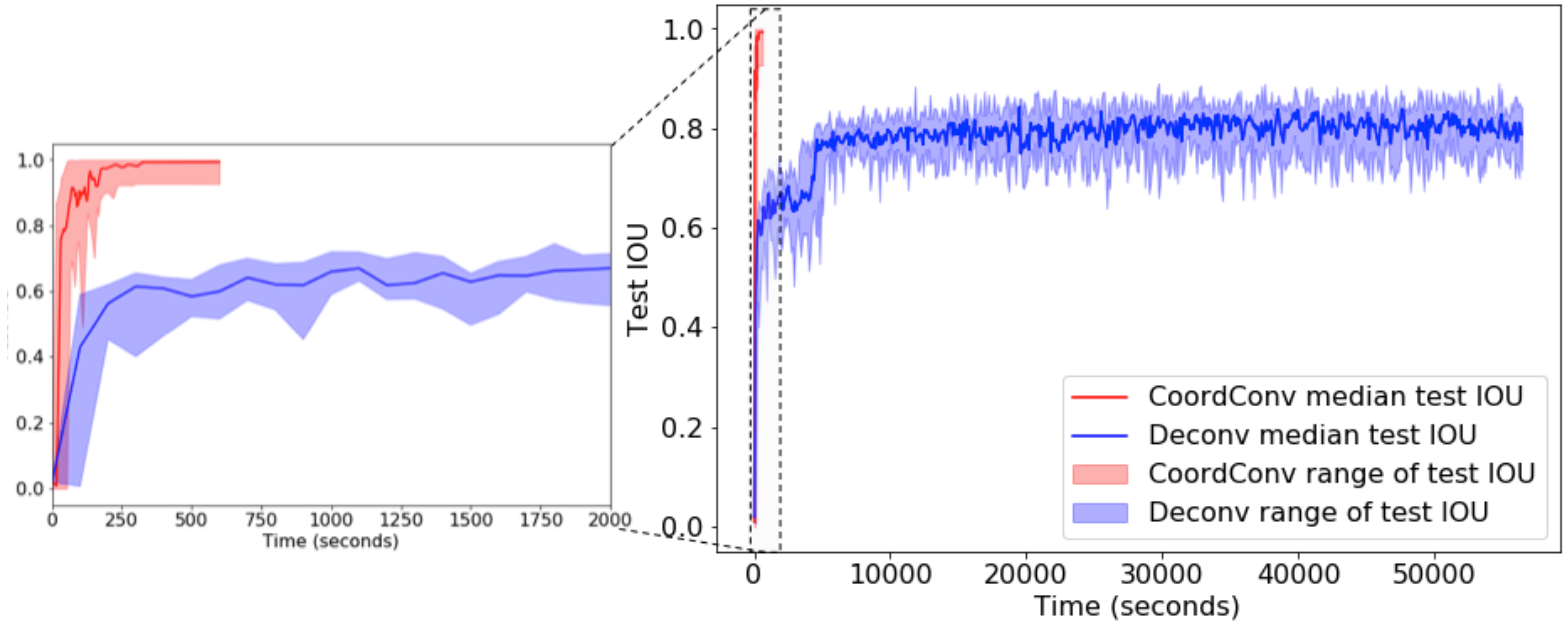

Deconv
CoordConv
Logits
Post-softmax and Target
Logits
Post-softmax and Target
Train
Test
Test
Train
Train
Train
Test
Test
Train
Train

On quadrant split



More results on supervised rendering

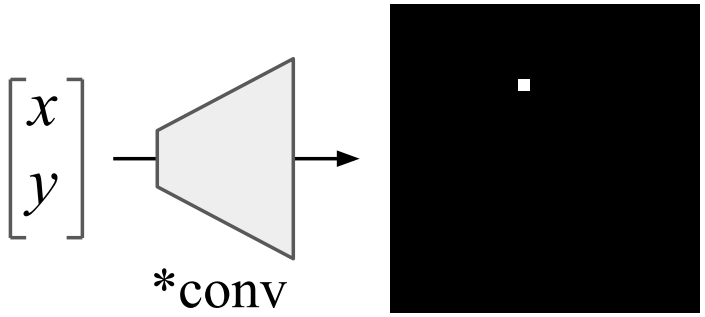




Performance on more “real” problems

- Claims that CoordConv helps in a wide range of domains
- Object detection
  - on a simple problem of detecting MNIST digits scattered on a canvas, we found the IOU of a Faster-RCNN network improved by about **24 percent** when using CoordConv.
- Image classification
  - Of all vision tasks, we might expect image classification to show the least performance change when using CoordConv instead of convolution, as classification is more about *what* is in the image than *where* it is.
  - Indeed, when we add a CoordConv layer to the bottom of ResNet-50 and train on ImageNet, we find only a microscopic improvement.

Toy Tasks

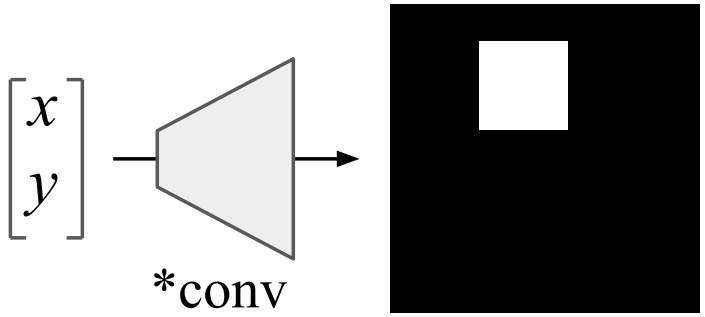


Supervised Coordinate Classification

Output: softmax Loss: cross entropy

(With reverse arrow) **Supervised Coordinate Regression**

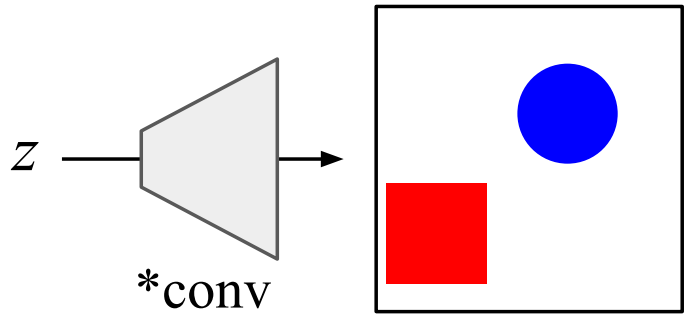
Output: linear Loss: MSE



Supervised Rendering

Output: per-pixel sigmoid

Loss: cross entropy

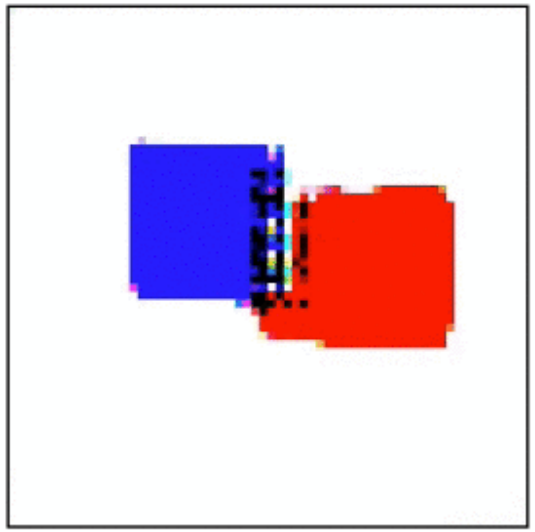


Unsupervised Density Learning (GAN)

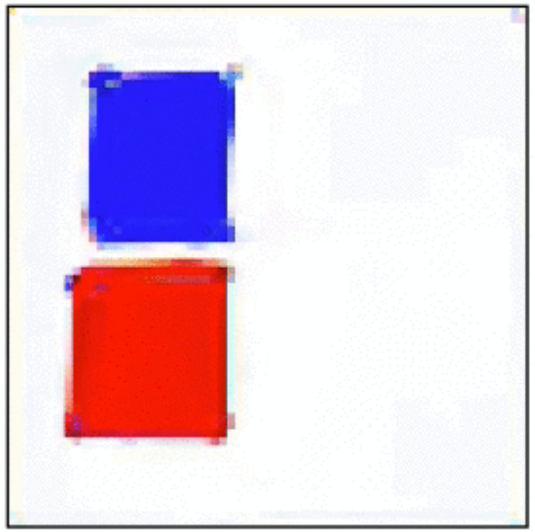
Output: per-pixel, per-channel sigmoid

Loss: learned GAN discriminator

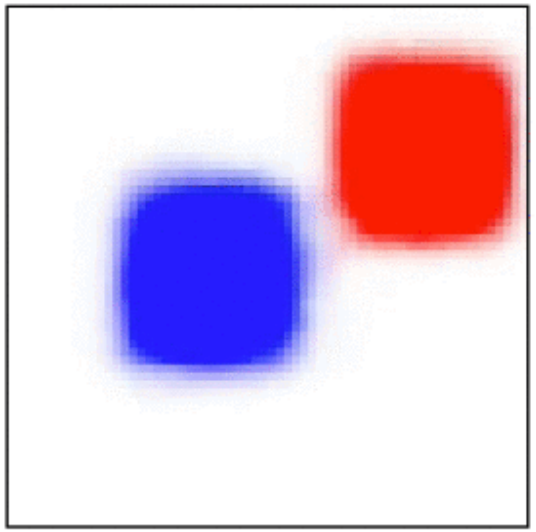
Generative models



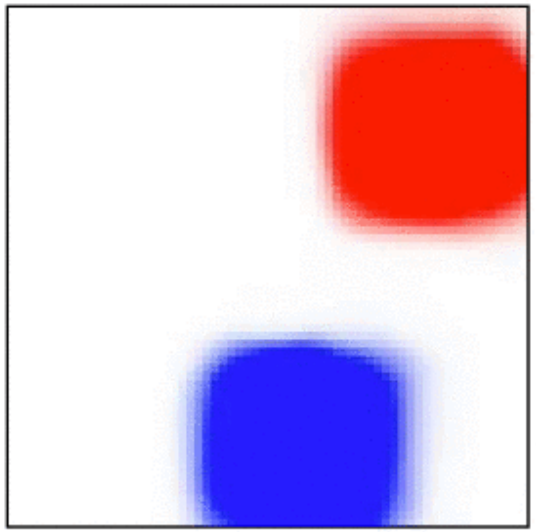
Conv GAN



CoordConv GAN



Conv VAE



CoordConv VAE

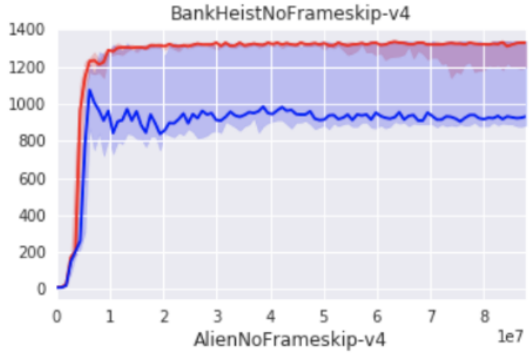
More on GAN

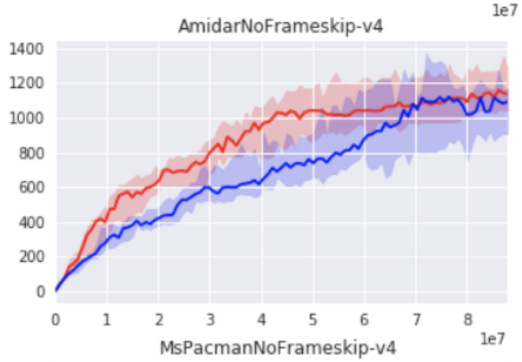
Conv GAN

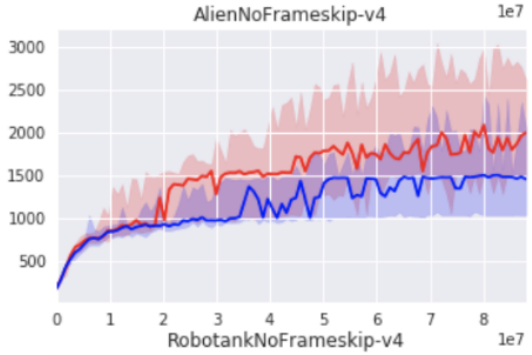
CoordConv GAN

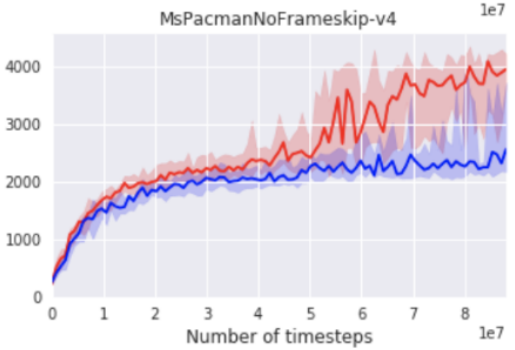
Results on Reinforcement Learning

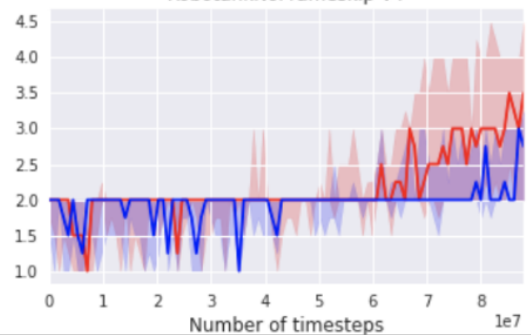
Works better or faster



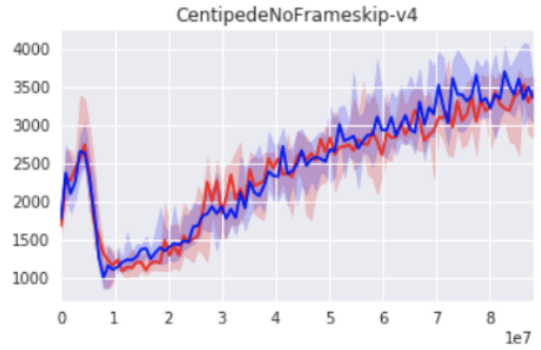


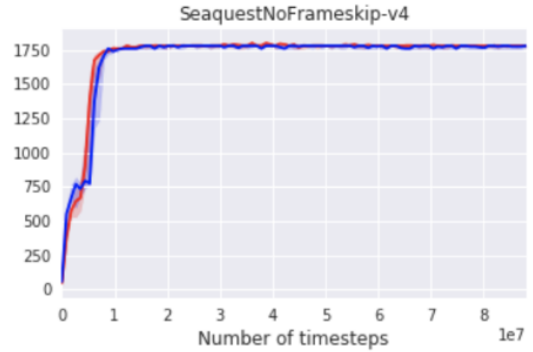




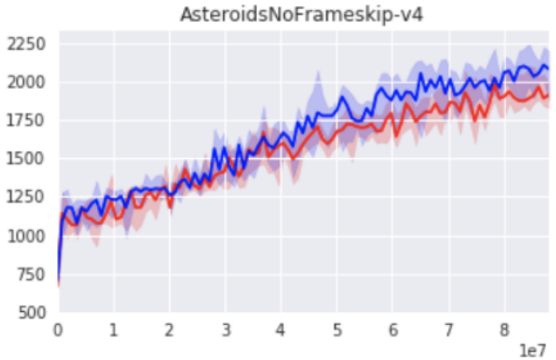


Similar





Slightly worse



Summary

- Extremely simple idea → just add two additional channels to the input image!
- Seems to be effective → can boost performance for a number of applications
- Next steps
  - Evaluate the benefits of CoordConv in large-scale datasets
  - Detection
  - Language tasks
  - Video prediction
  - Spatial transformer networks → image registration?