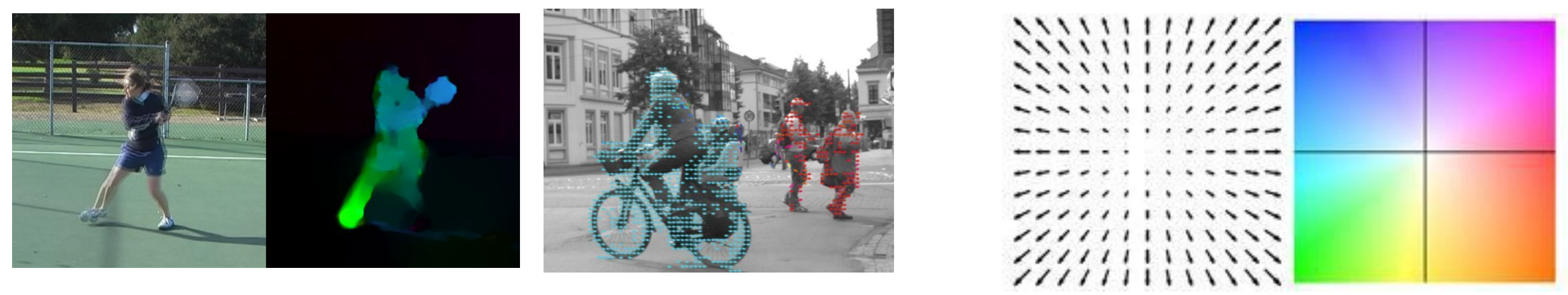
How Do Deep Neural Nets Compute Optical Flow?

Raphi Kang[1,2], Carsen Stringer[1]

[1] Janelia Research Campus, HHMI, [2] Massachusetts Institute of Technology

Optical Flow

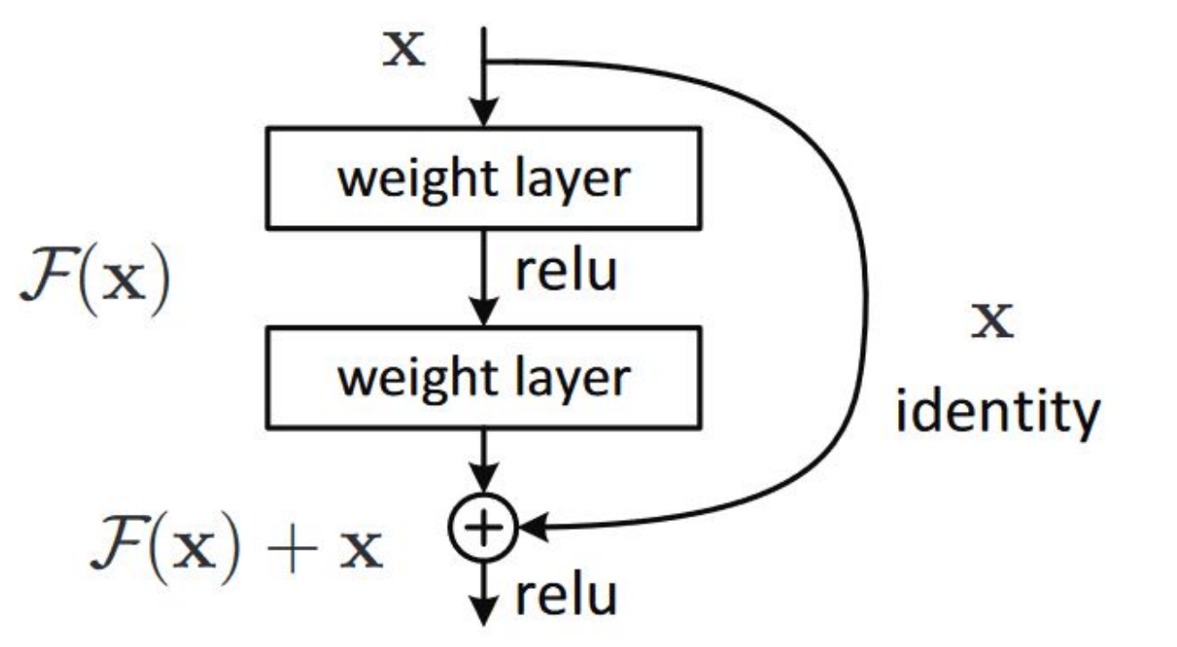
Tracking motion of objects between consecutive frames



Useful for: object/pedestrian detection, camera motion estimation, motion boundaries in scene, video compression.
Requires per pixel localization

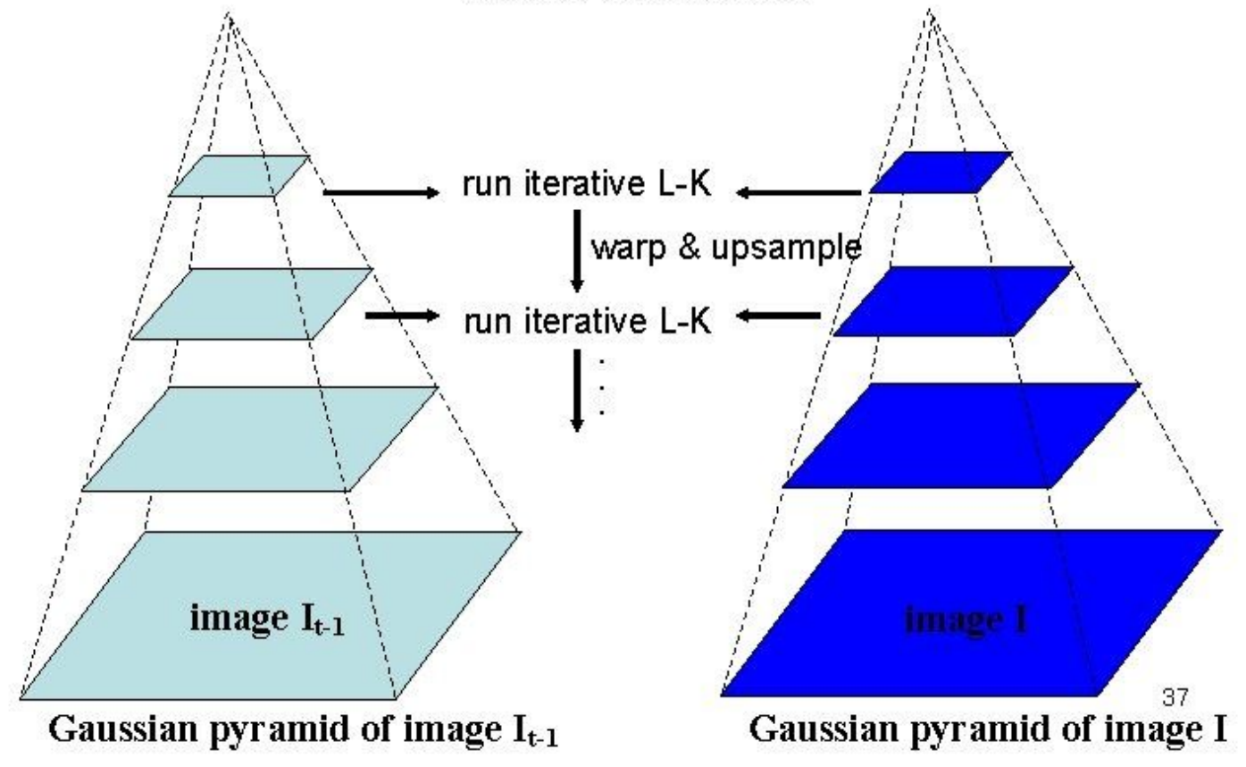
DNN Architectures: Skip Connections & Image Pyramids

Residual connection


x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

- re-introduce global information

Coarse-to-fine optical flow estimation

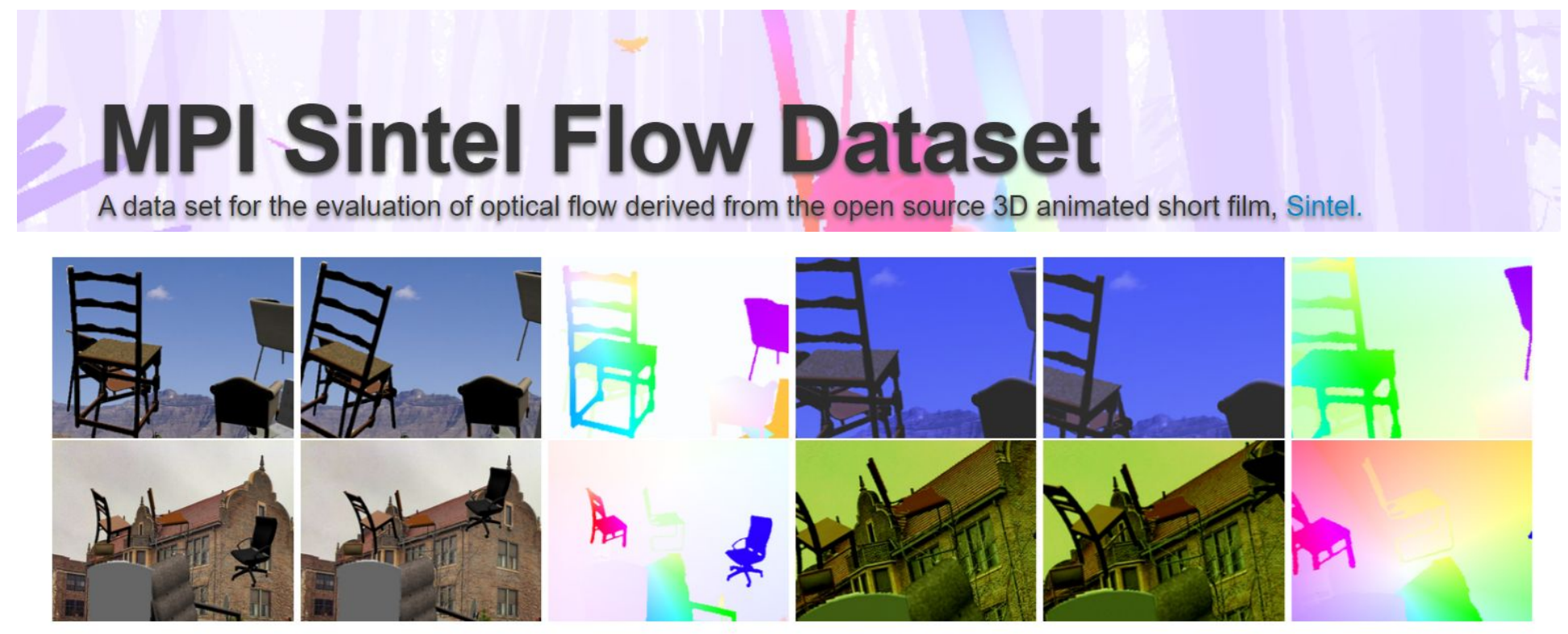


- detect large motions as pixel shifts

Brightness Constancy Constraint: $I(x,y,t) = I(x+\Delta x, y+\Delta y, t+\Delta t)$
(small flow field restriction)

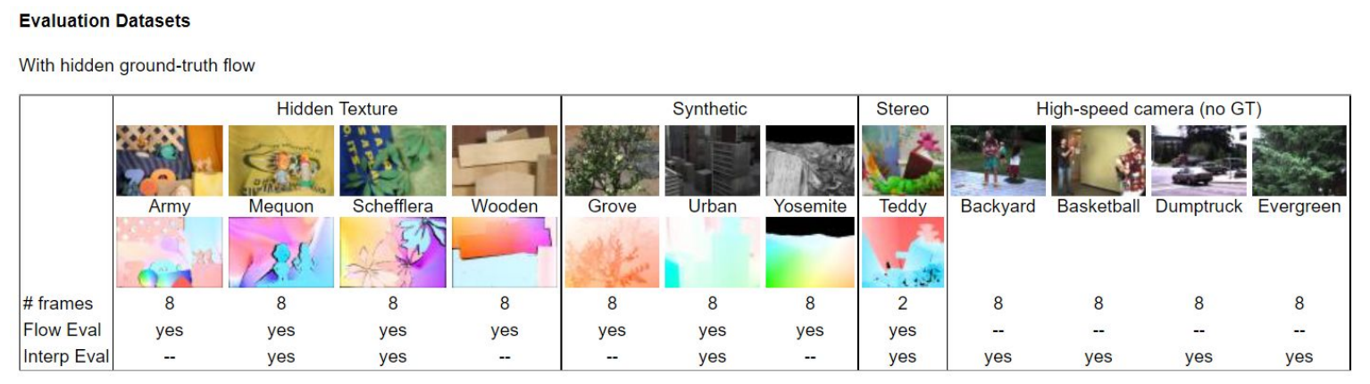
Optical Flow Datasets & Preprocessing + Augmentation

State of the Art datasets with ground truth labels:
MPI Sintel (open source animated film) & Flying Chairs


MPI Sintel Flow Dataset
A data set for the evaluation of optical flow derived from the open source 3D animated short film, Sintel.

Unused non-digital data:
KITTI - (too specific) Middlebury - (too small)

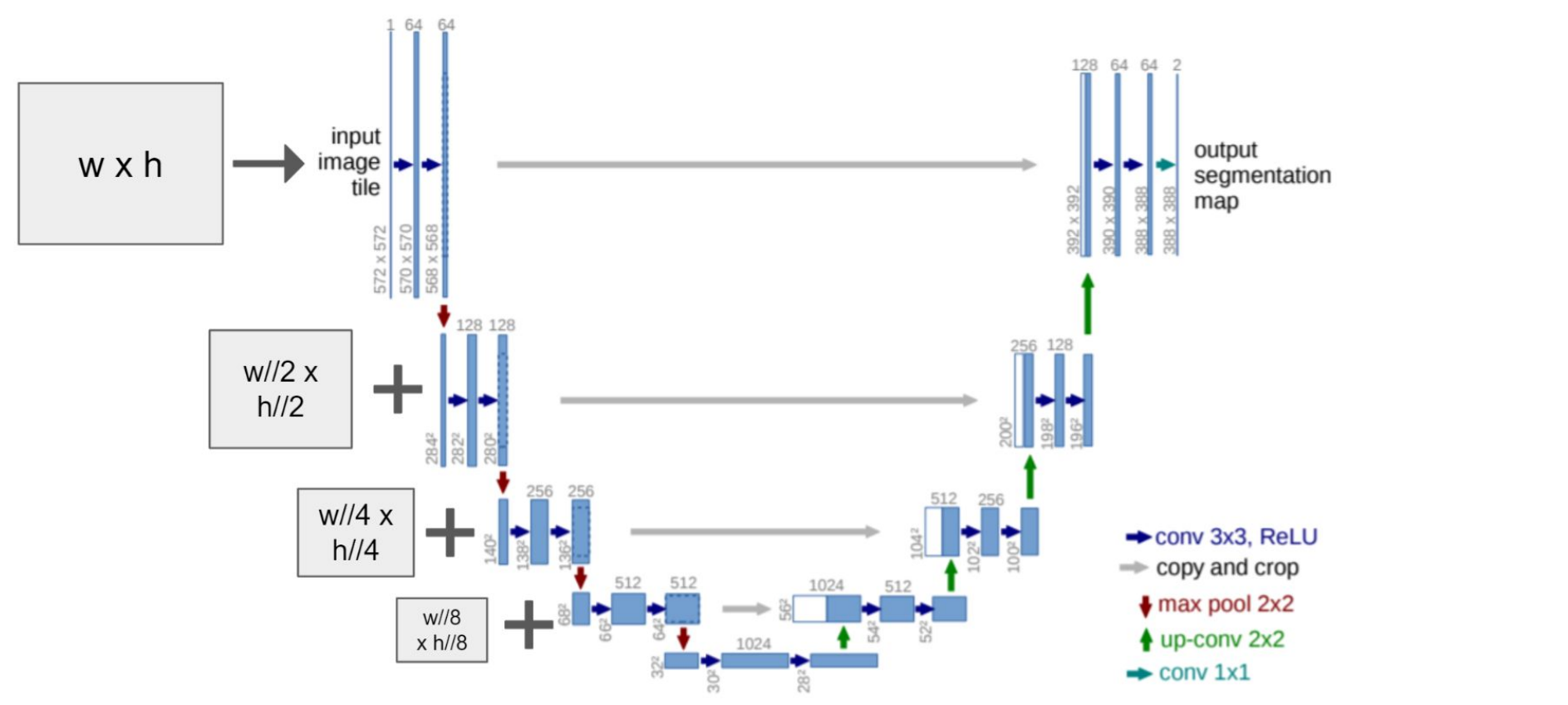
360° Velodyne Laserscanner
Stereo Camera Rig
GPS
KITTI-360
http://www.cvlibs.net/datasets/kitti-360


Evaluation Datasets
With hidden ground-truth flow
Hidden Texture
Synthetic
Stereo
High-speed camera (no GT)
Army
Mequon
Schefflera
Wooden
Grove
Urban
Yosemite
Teddy
Backyard
Basketball
Dumptruck
Evergreen

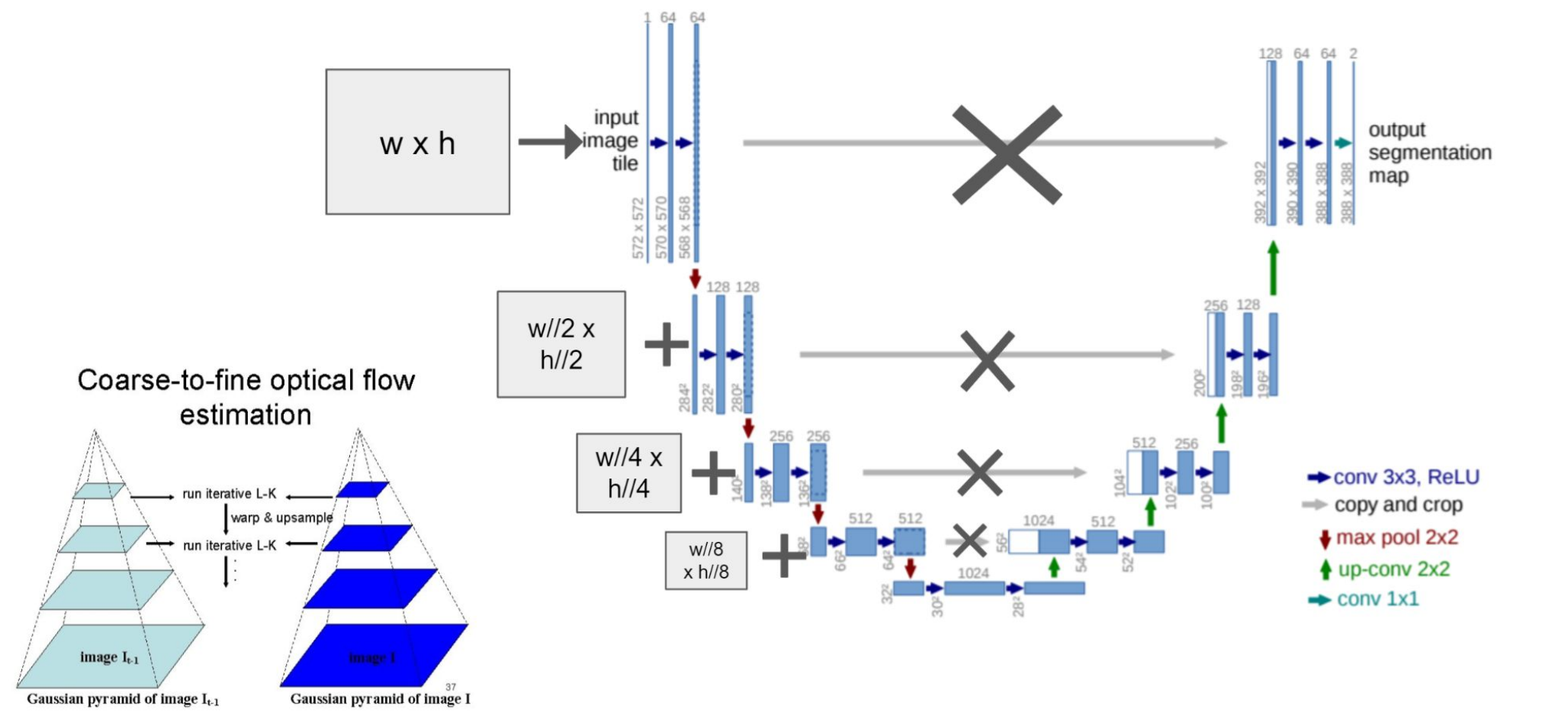
Preprocessing: normalize pixel values, tanh to rid OF of outliers and make more vivid, downsample input
Data Augmentation: random rotate, crop, zoom, translate data & OF.

DNN Architectures

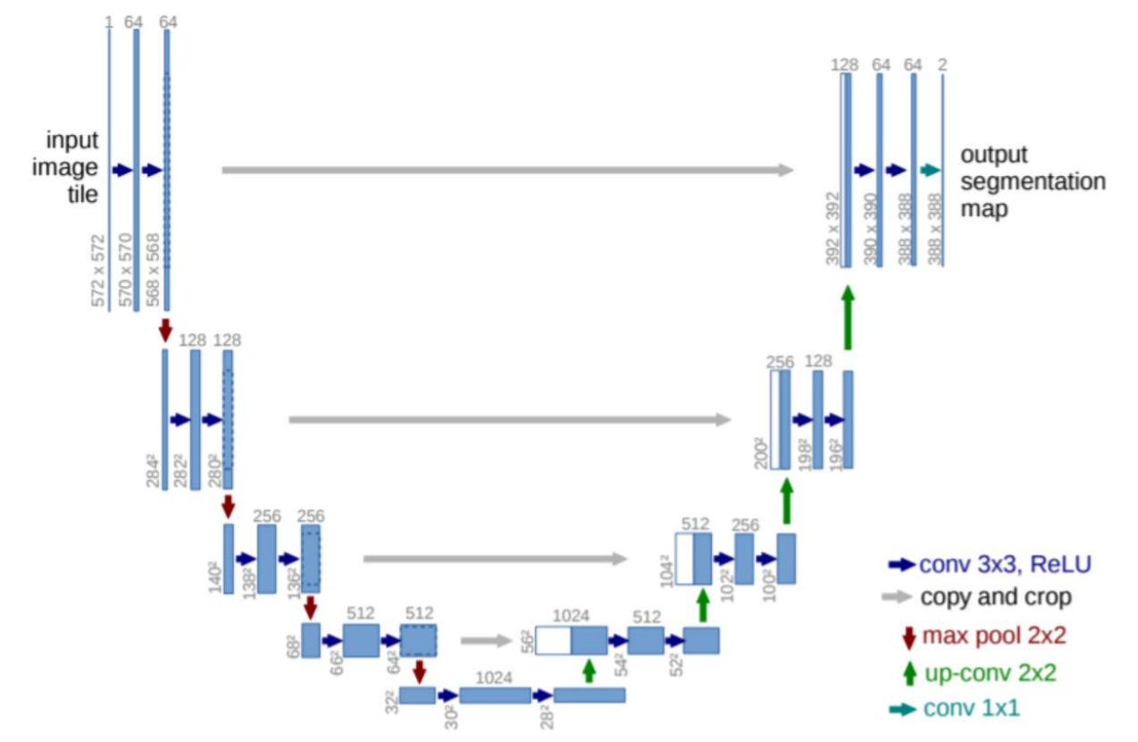
1. UNet Architecture + Image Pyramid



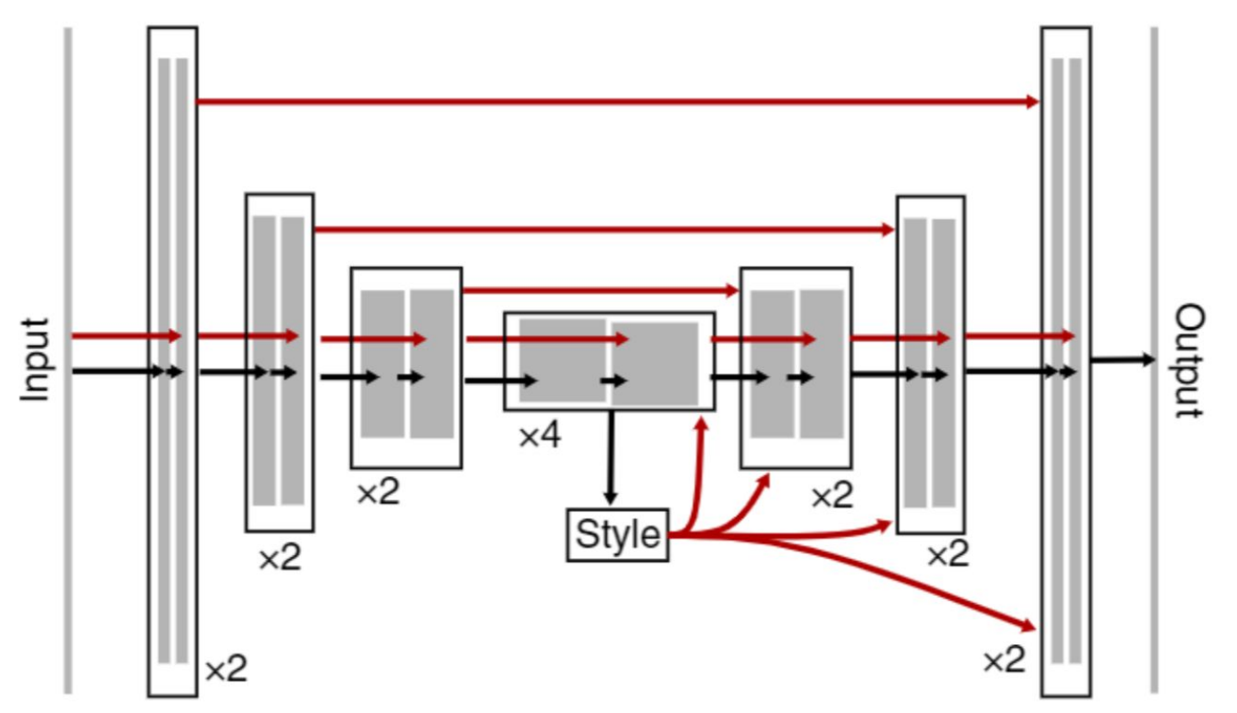
2. Image Pyramid Only



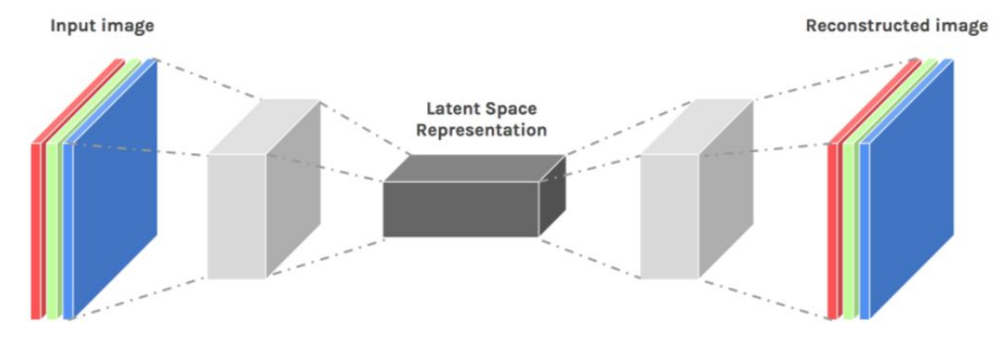
3. Simple Unet (No Image Pyramid)



4. ResNet Architecture



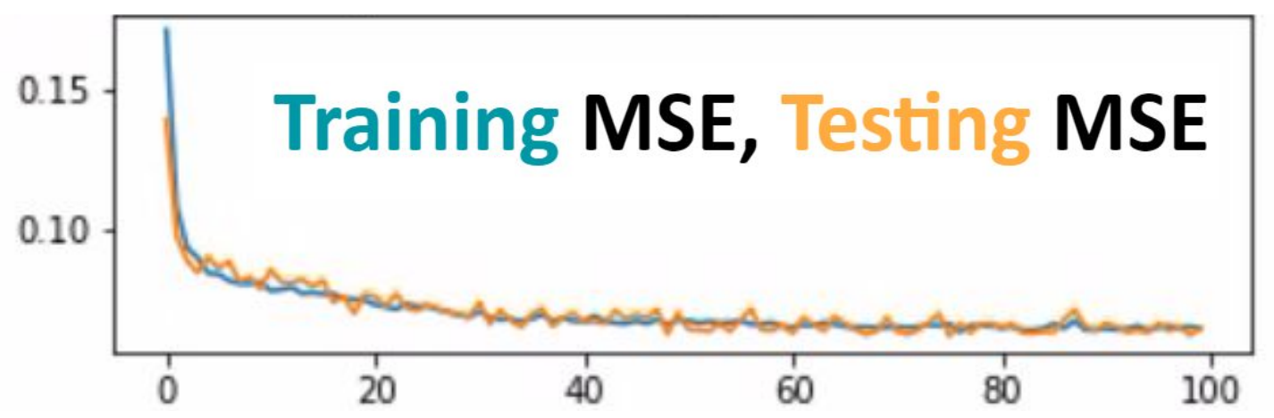
5. Encoder-Decoder Network



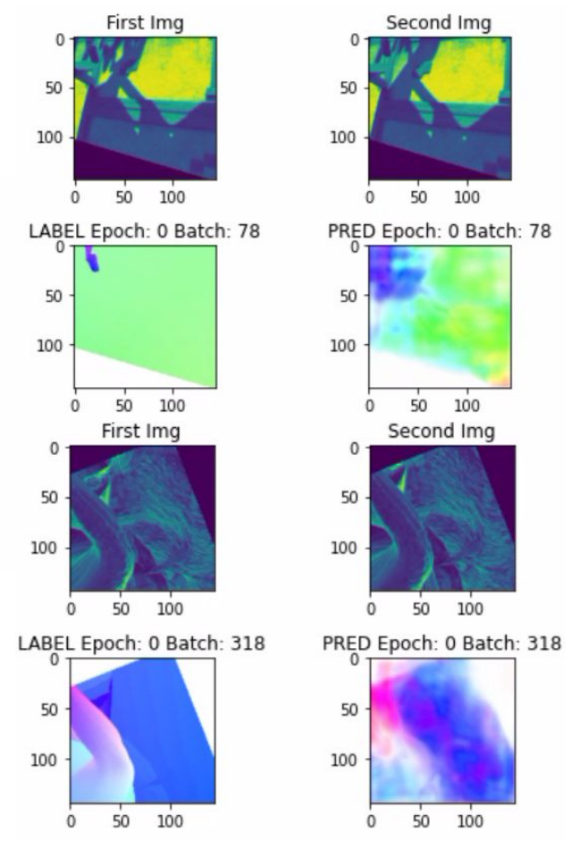
DNN performance on OF

- Training Loss: minimize MSE - Optimization Method: Adam - Set aside 10% of input data for testing batch

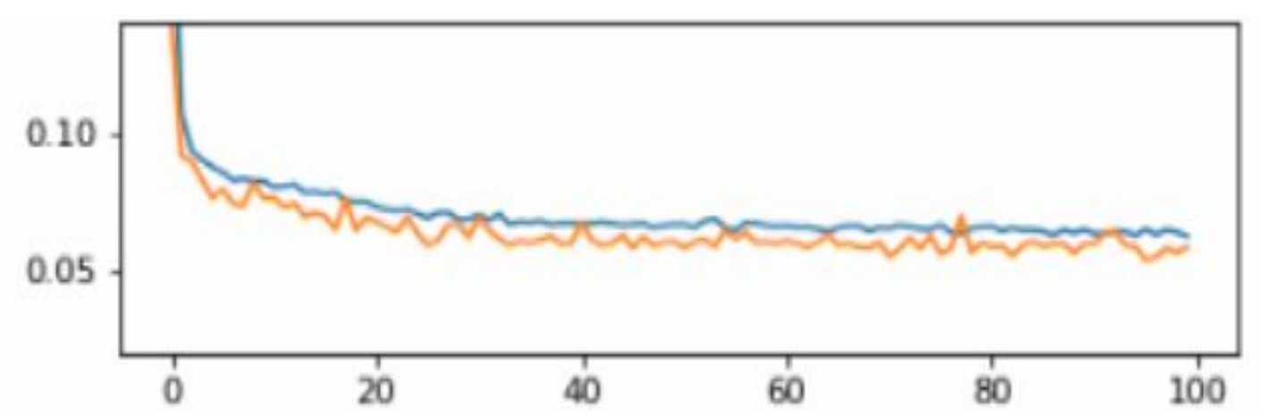
1. Unet + Img Pyramid performance



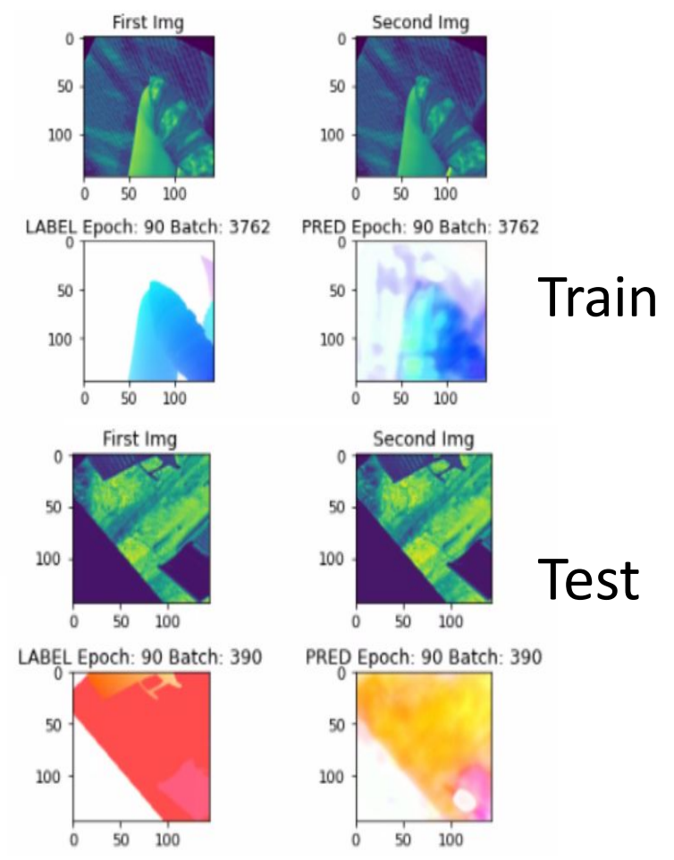
Avg testing MSE at last 5 epochs (of 100): **0.06376**



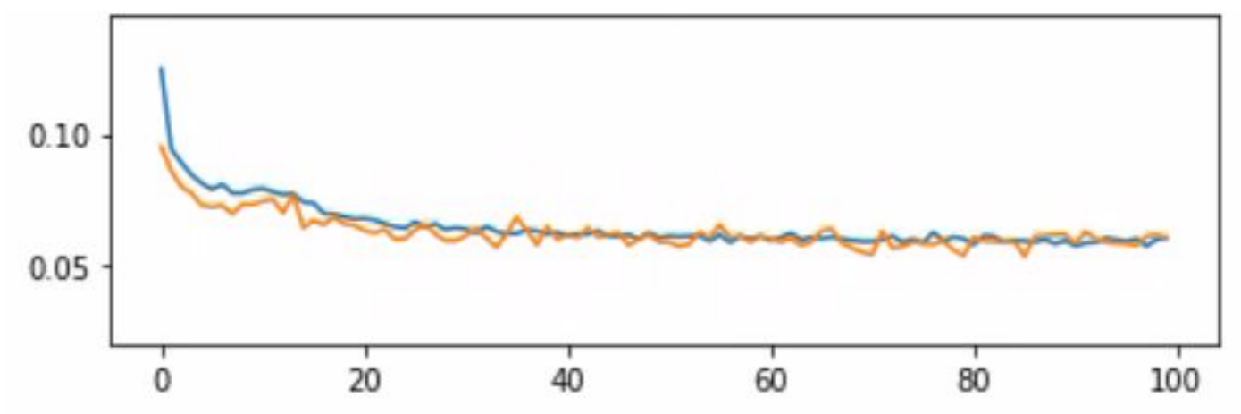
2. Simple UNet Performance



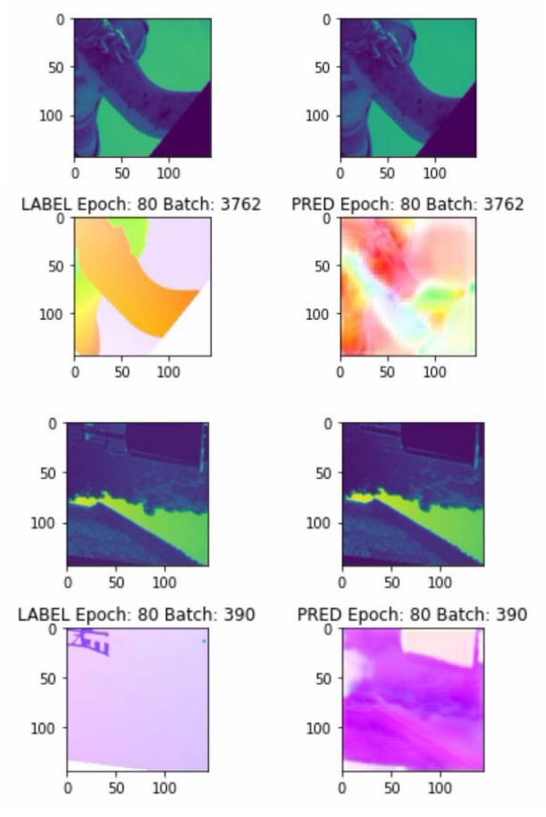
Avg MSE: **0.05630**



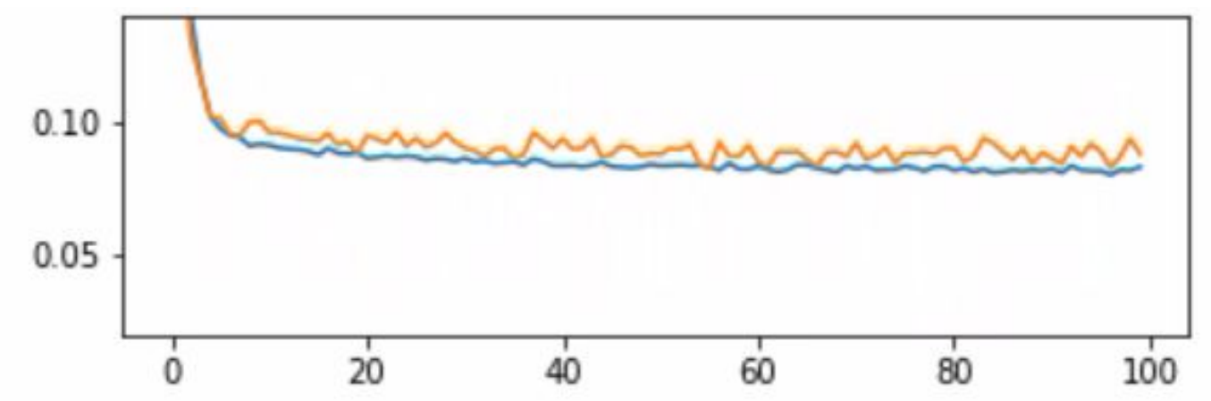
3. ResNet Performance



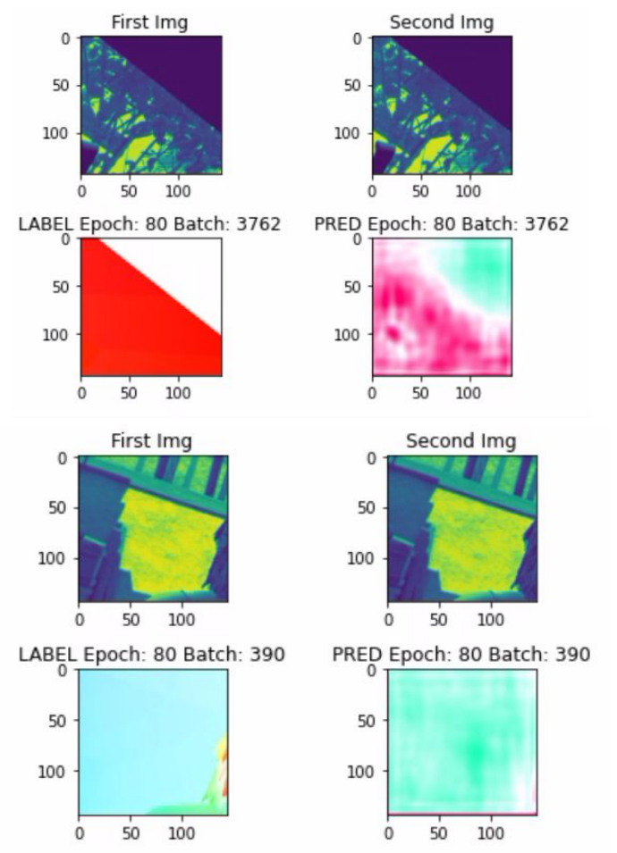
Avg test MSE: **0.05977**



4. Image Pyramid Performance

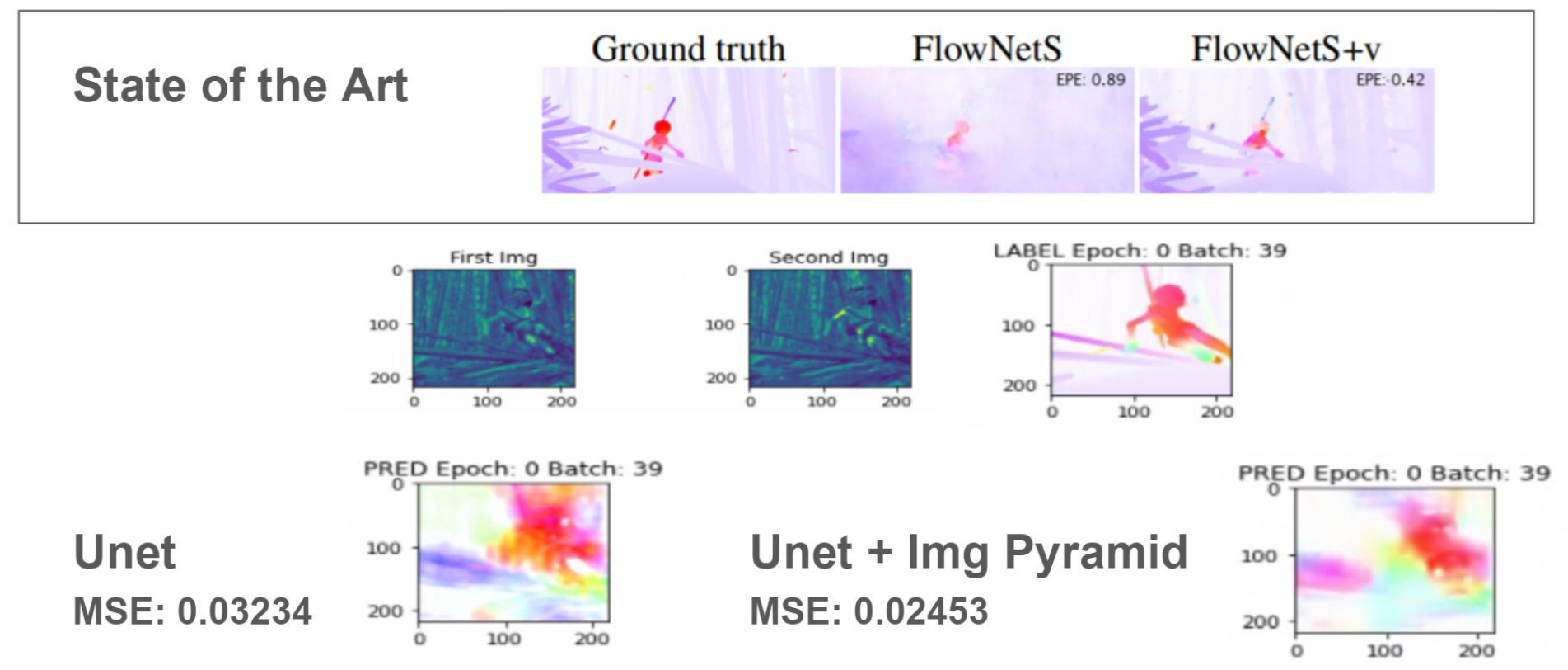


Avg MSE: **0.08097**



	Encoder-Decoder	Image Pyramid + Encoder-Decoder	CPNet	UNet	Image Pyramid + UNet
Skip Connections	X	X	Yes	Yes	Yes
Image Pyramid	X	Yes	X	X	Yes
Avg. Testing MSE of last 5 epochs	>0.09805	0.08097	0.05977	**0.05630**	0.06376

Qualitative Comparisons



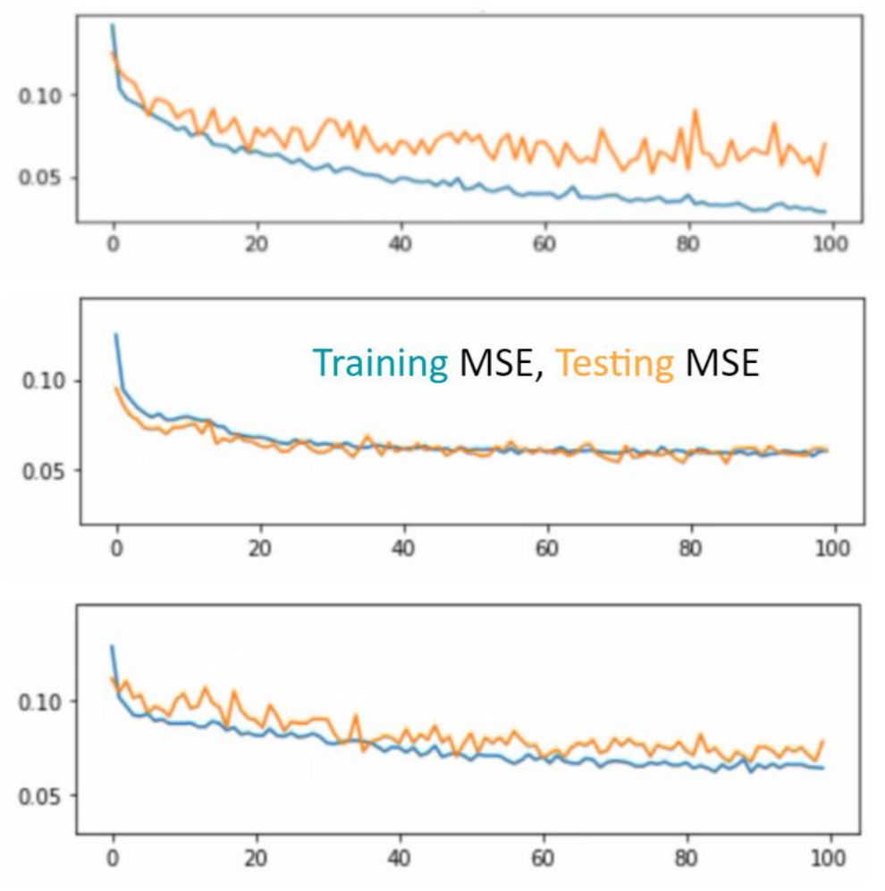
Flownet (2017, “state of the art”) uses variational refinement for edge detection + uses all of flying chairs to train (10x Sintel data size)
Image Pyramid makes overall prediction smoother.

Discussion - Effect of Data Limitations

Sintel Only
Size ~2000

(Sintel + Chairs)
Size ~4000

(Sintel+Chairs)//2
Size ~2000



Overfitting

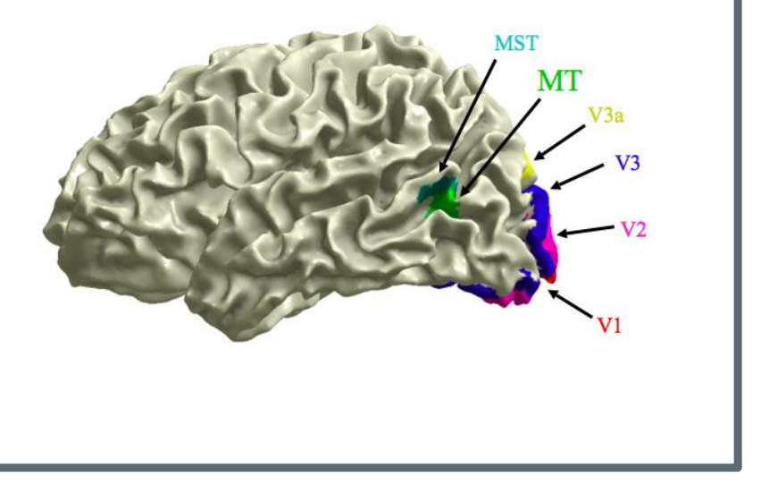
MSE (avg of 5): 0.05977

MSE (avg of 5): 0.07282

Discussion - Biological Analogs to Architectures

Skip Connections: different cortical areas supply local/global information

Image Pyramid: generated conv. filters resemble biological motion processing filters



Conclusion & Future Direction

- Image Pyramid may not be best performing architecture to mimic multiple instances of temporal information relay in cortex
- While simple Unet is best performing, further refining needed to reach state of the art OF detection
- Statistical variation in dataset allows for more robust model (regardless of size)

In the Future:

- Improve data augmentation + increase dataset size + perform trials multiple times
- Representational Similarity Analysis to compare mouse neural encoding of images to NN representations