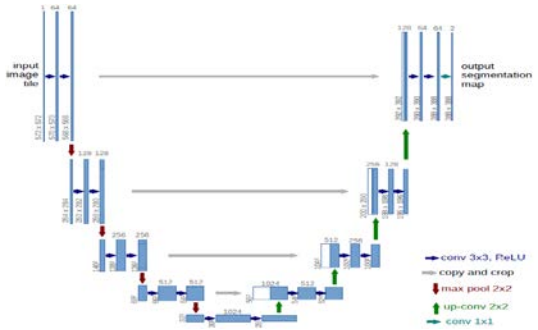
Dense correspondence estimation with deep learning

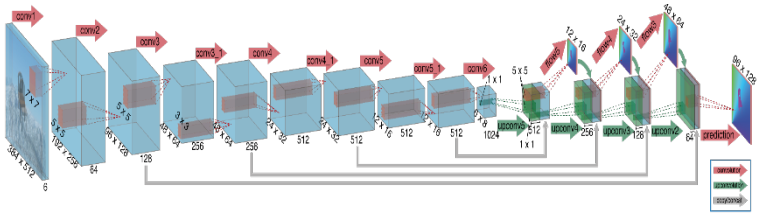
Cross-dataset generalization

Thomas Brox
University of Freiburg, Germany

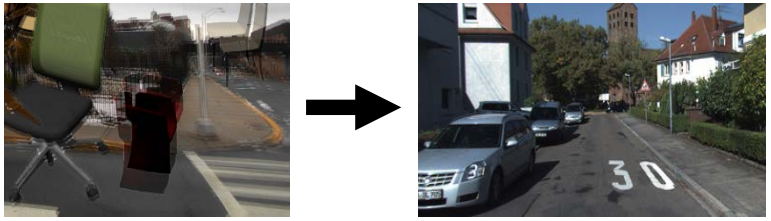
Overview



- Part I: Encoder-decoder networks

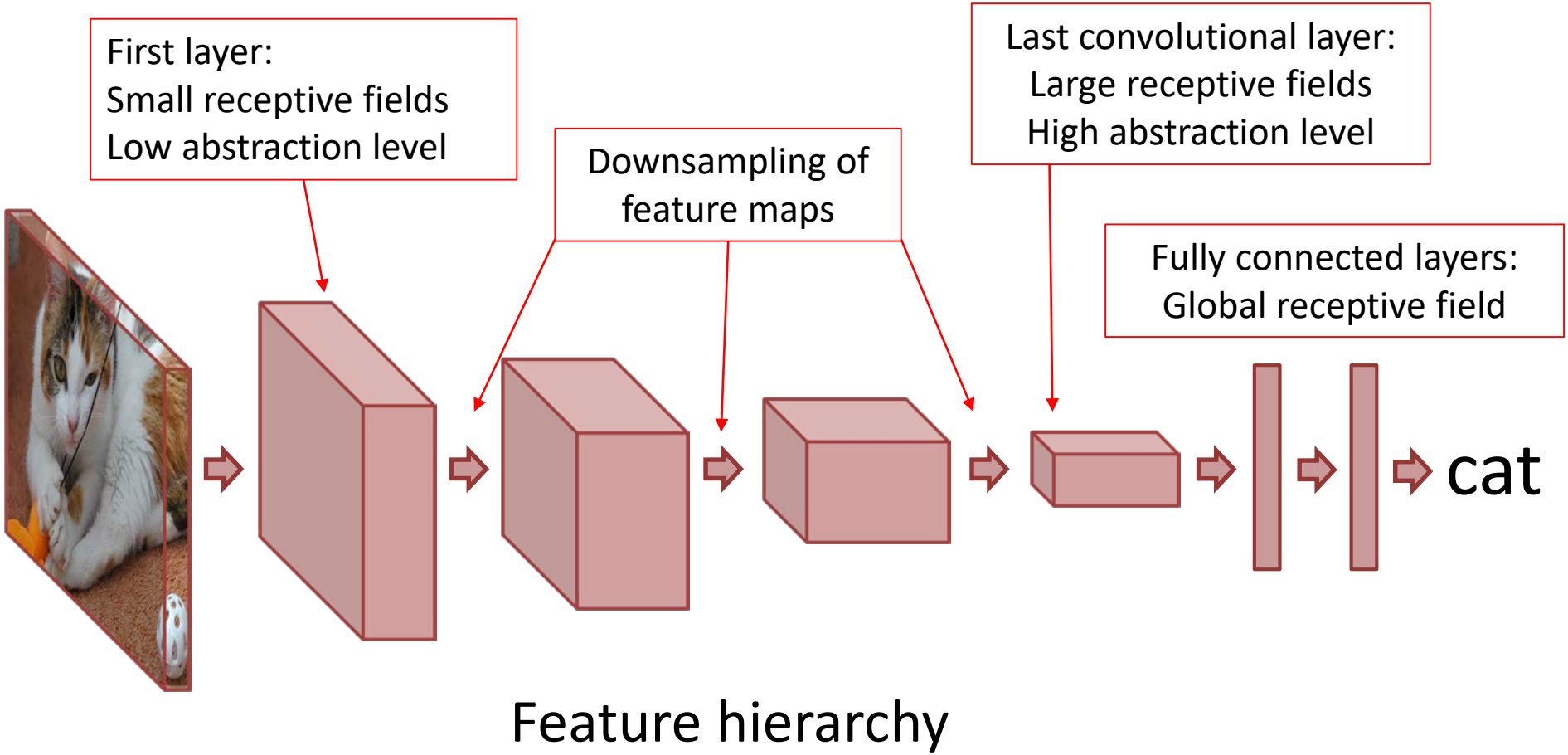


- Part II: Correspondence estimation with FlowNet



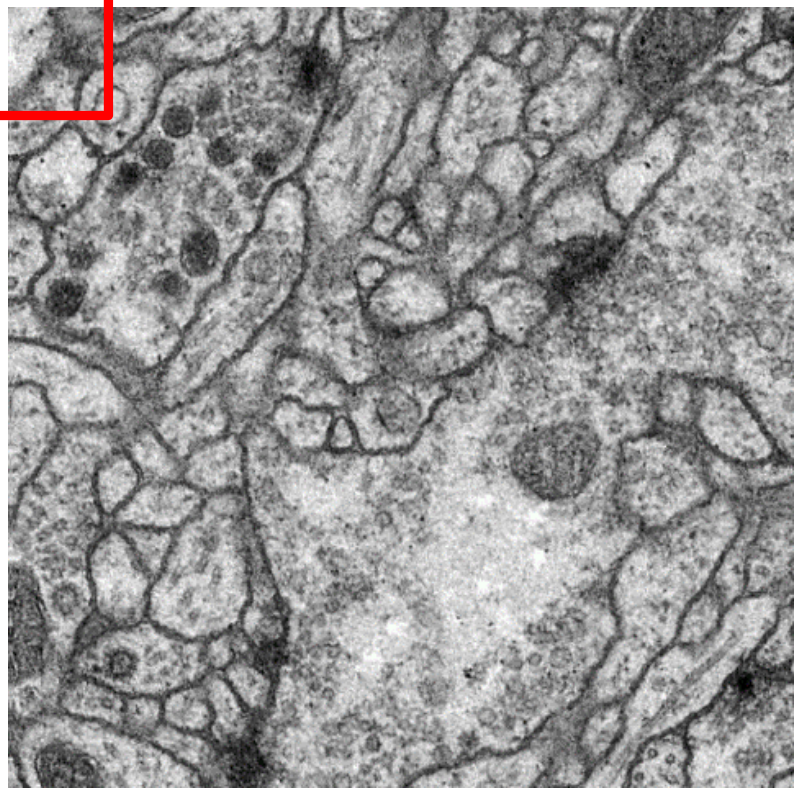
- Part III: Cross-dataset generalization

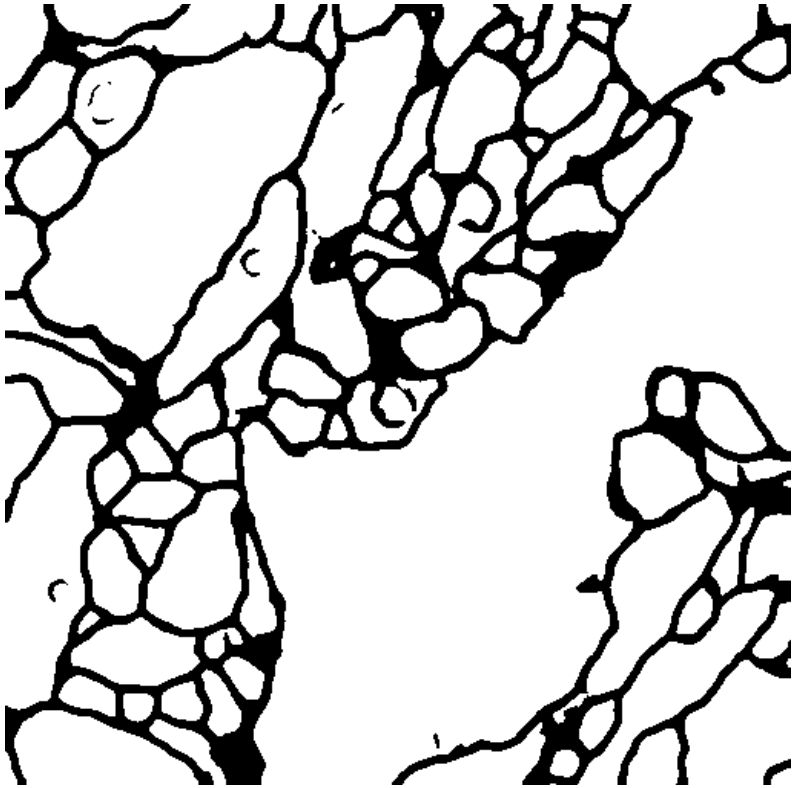
Typical image classification network



Feature hierarchy

Semantic segmentation as pixel classification



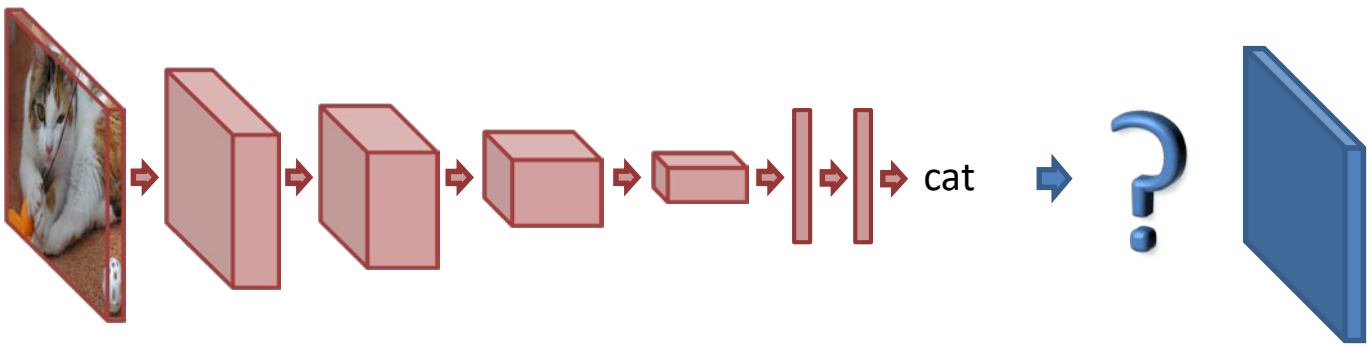


- Sliding window classification (Ciresan et al. 2012)

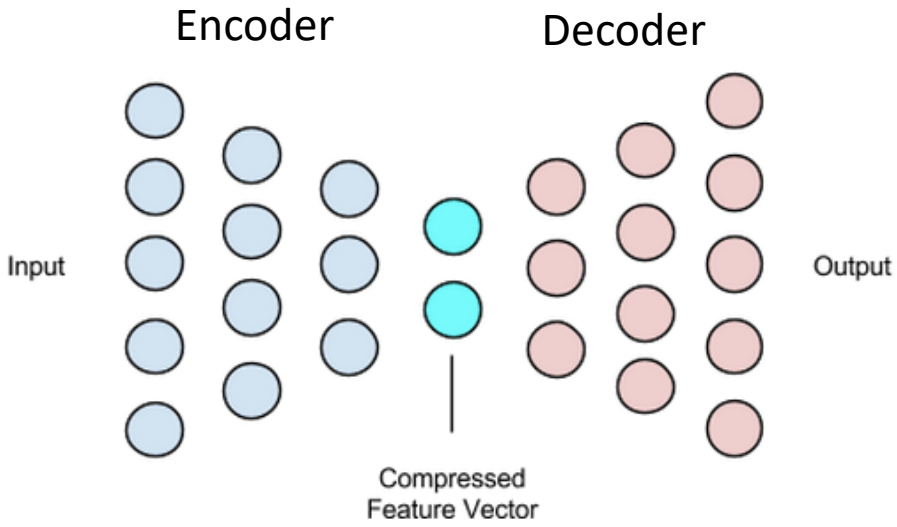
Limitations of patch-based analysis

- Each decision is taken independently
- Slow, since a network must be run for each pixel
  (can be avoided with a clever implementation)
- Patch size is an important parameter
  (trade-off between localization accuracy and context)

How to get back the resolution?

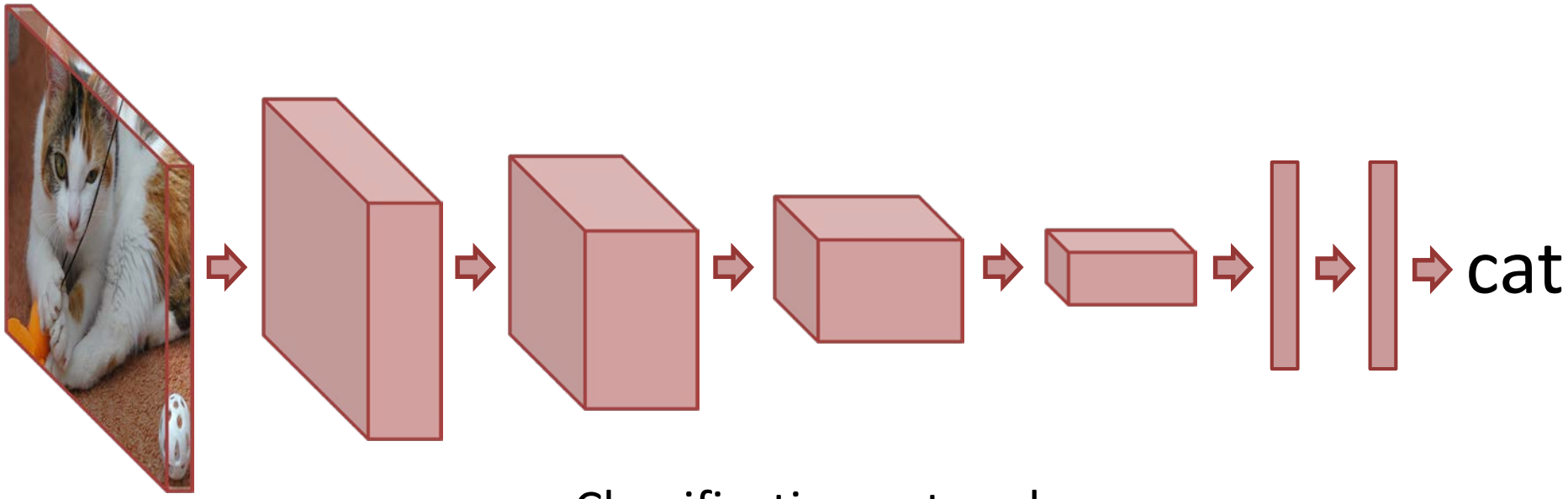


Auto-encoder



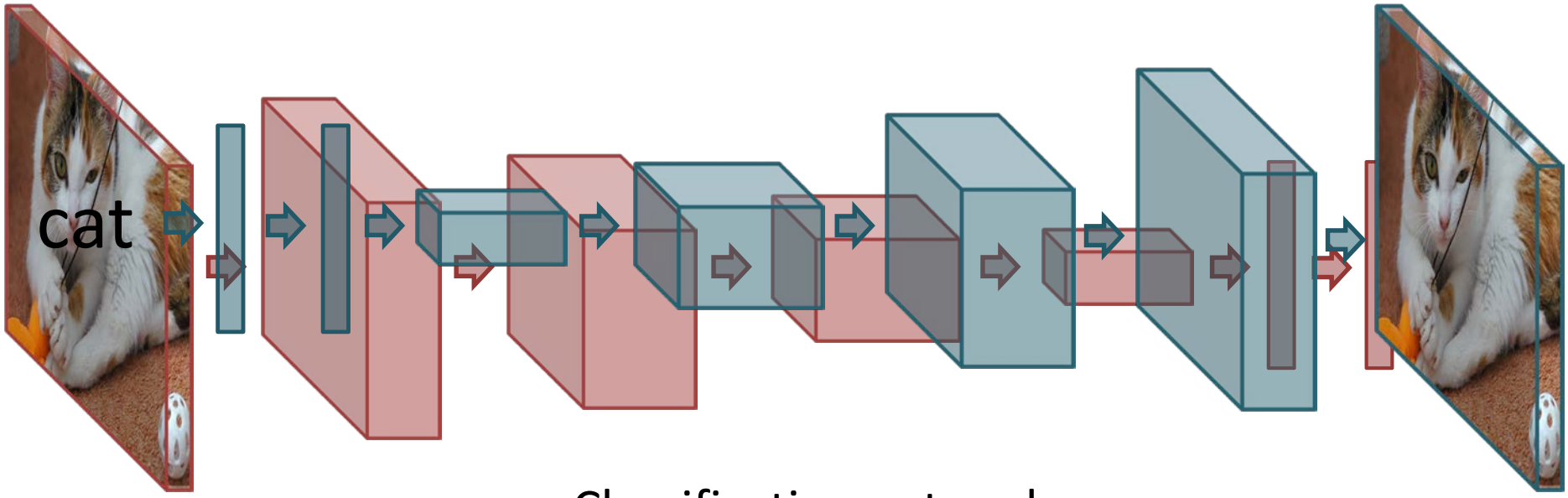
→ We need a convolutional decoder

Encoder



Classification network

Encoder



Classification network

Up-convolutional network (decoder)

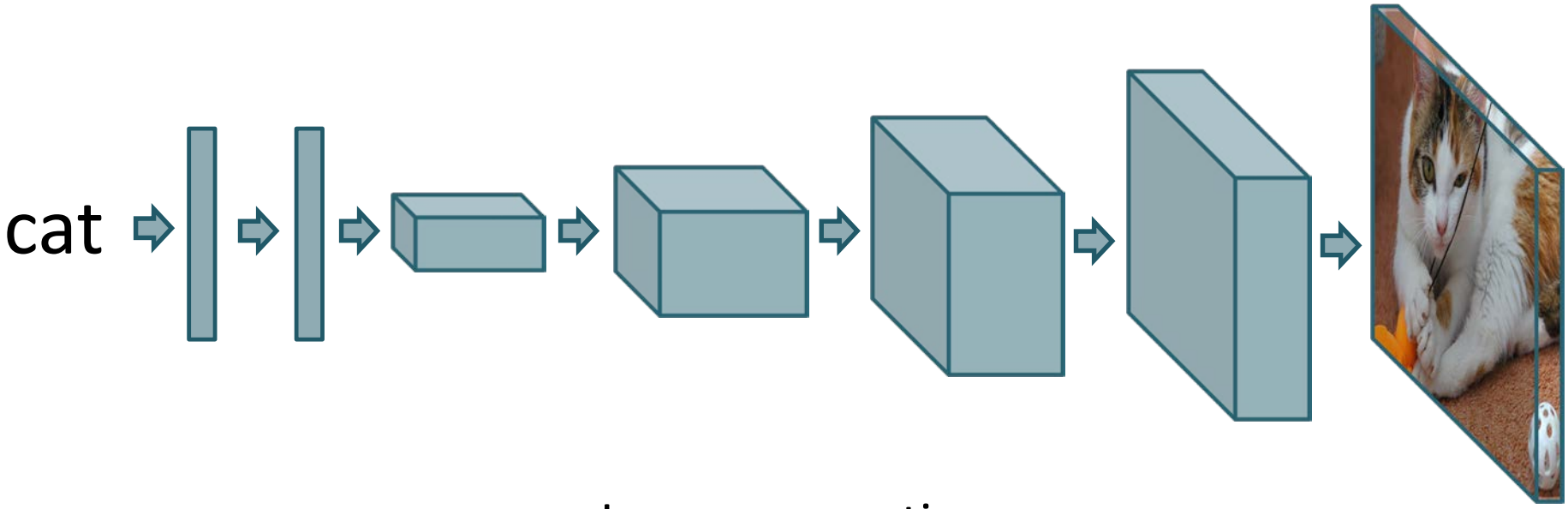
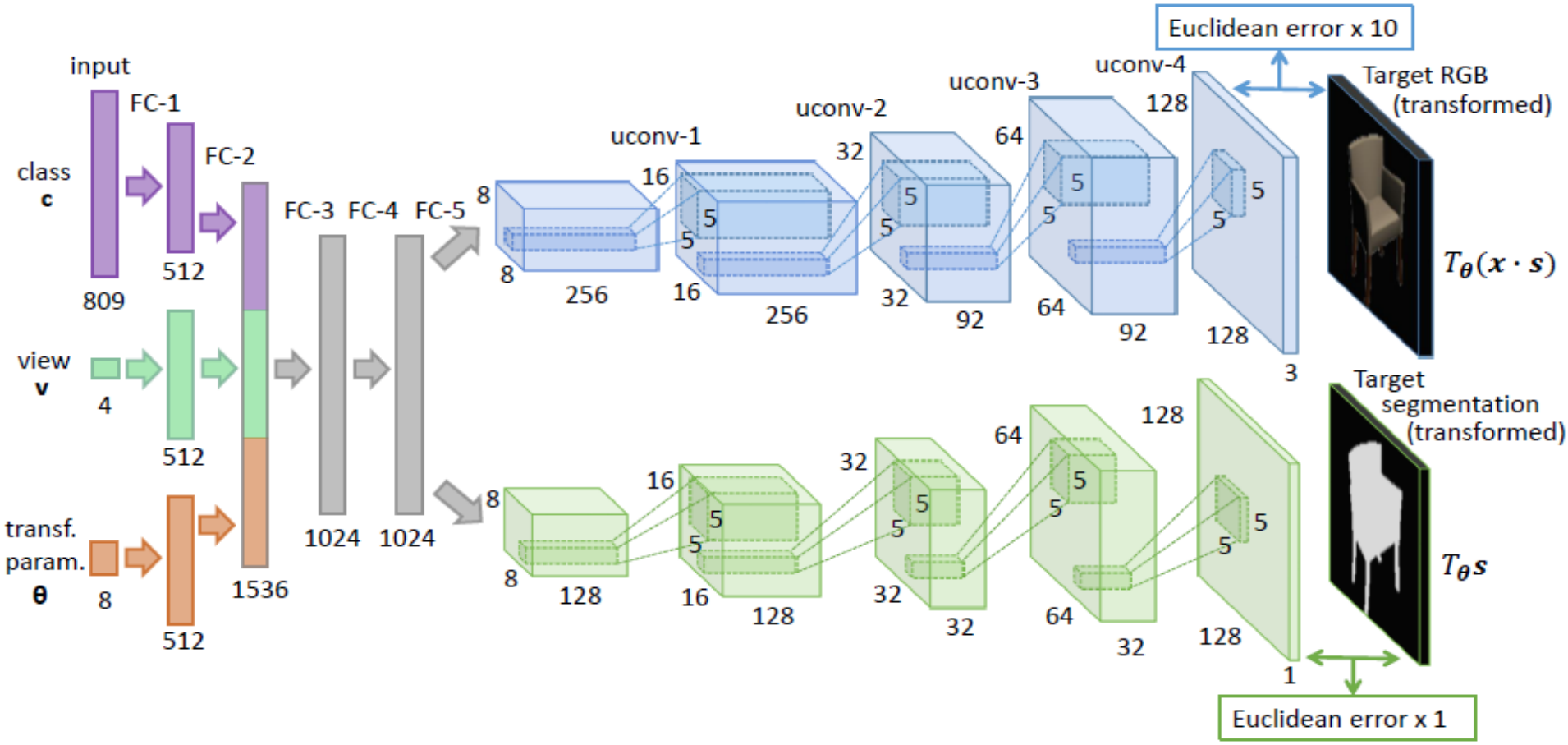


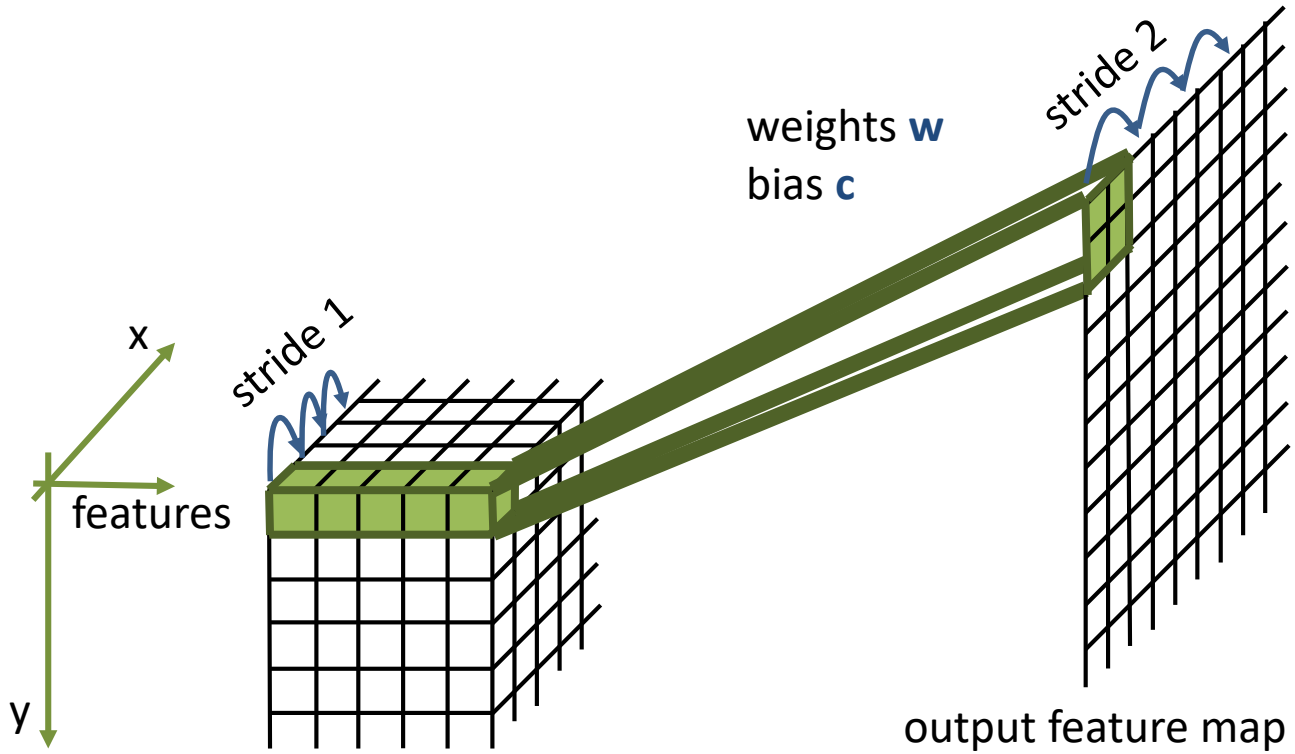
Image generation

Generating chair images with a network



Dosovitskiy et al. 2015

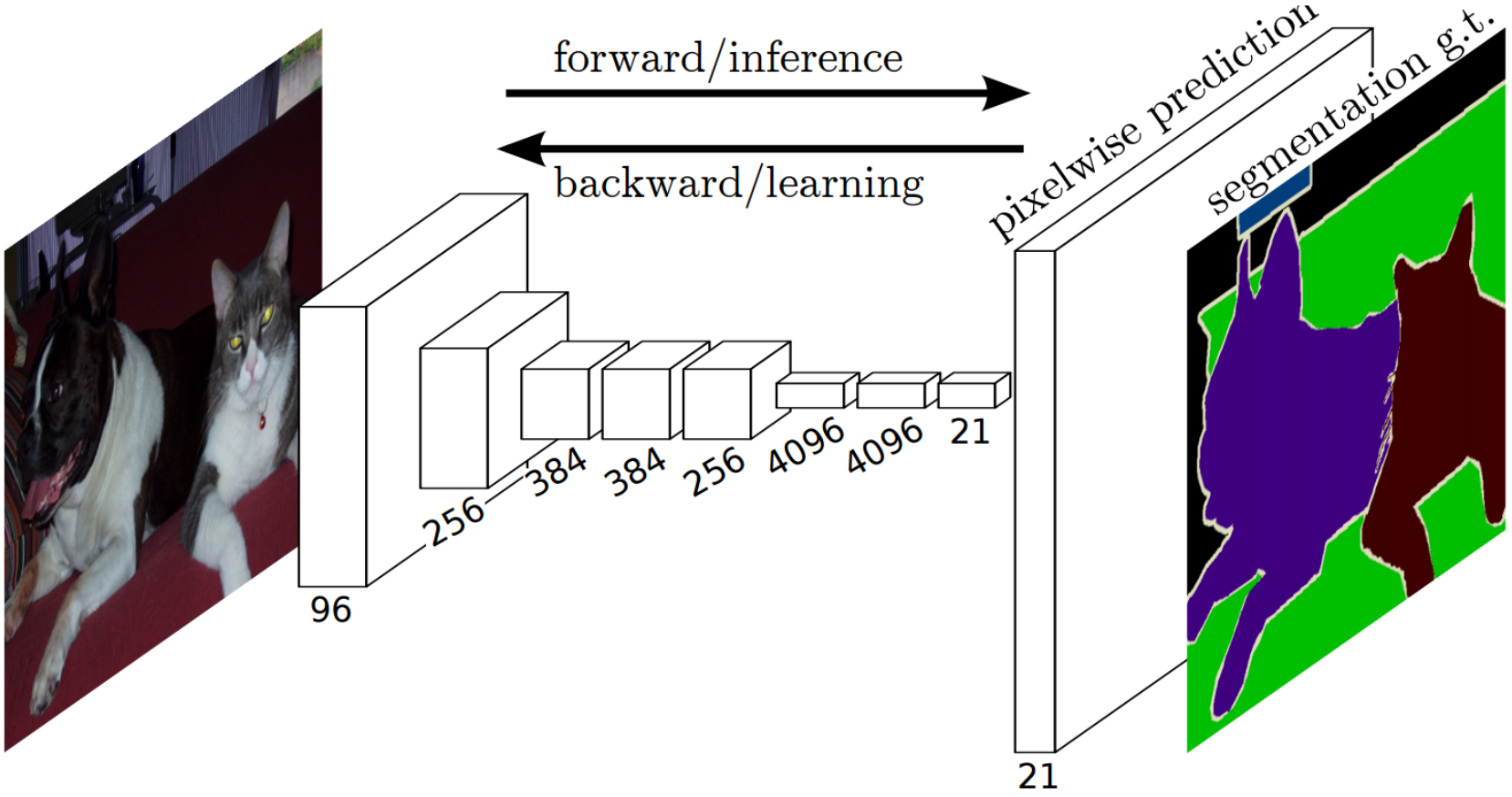
Up-convolution: convolution with stride



Intuition: learned upsampling operator

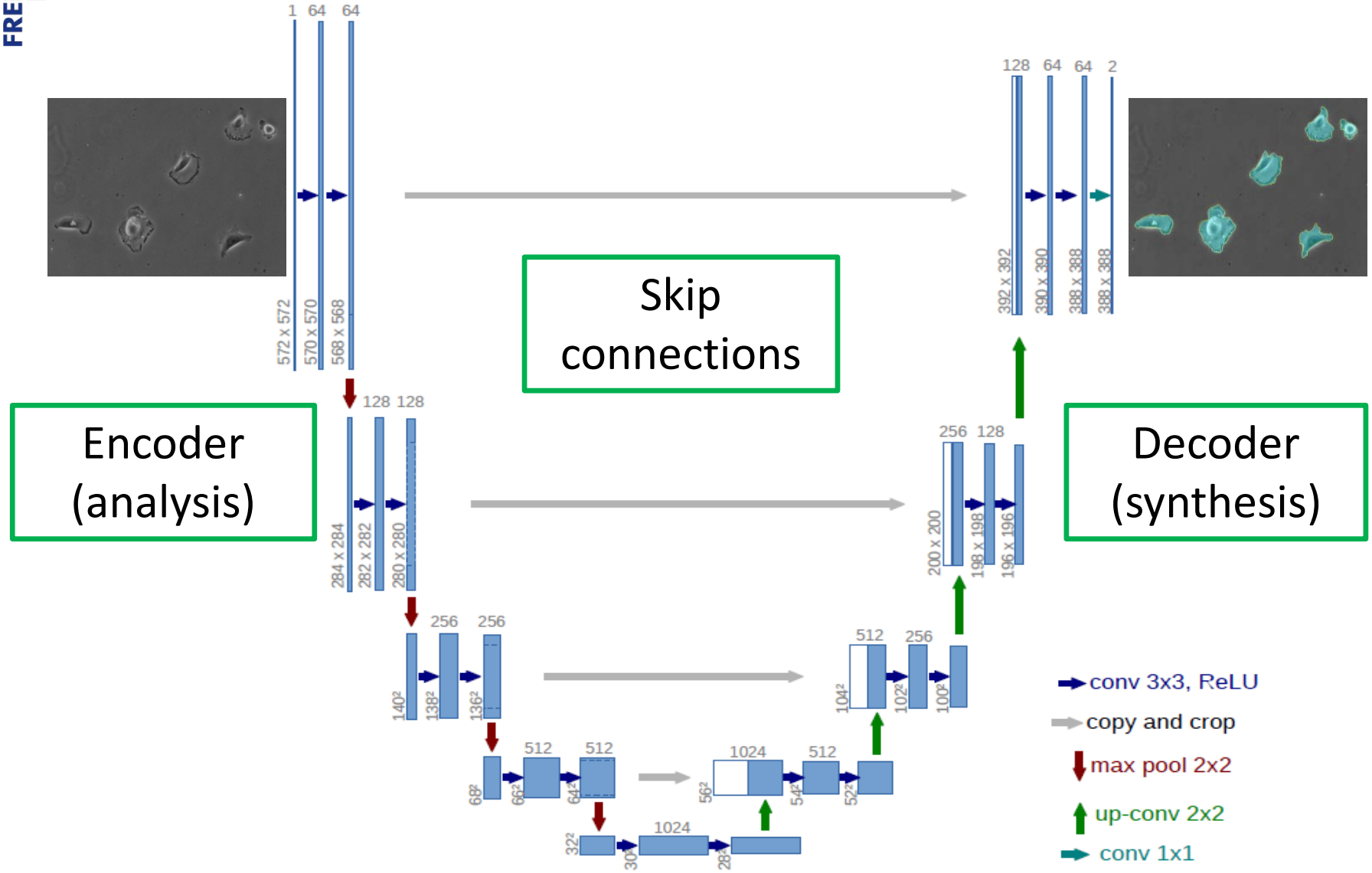
→ Synthesize high-res features from low-res features

Fully convolutional networks



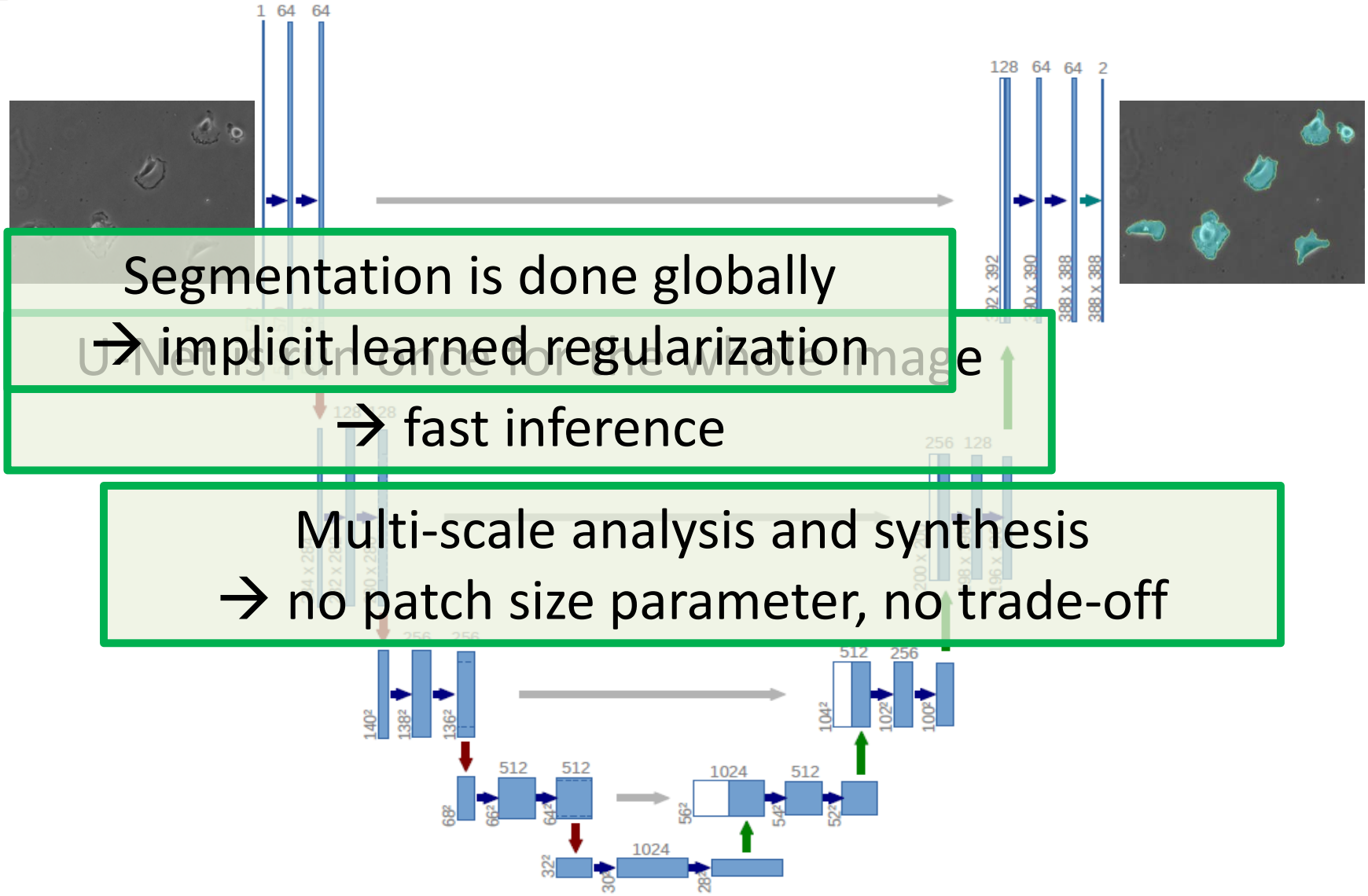
Long et al. 2015

U-Net



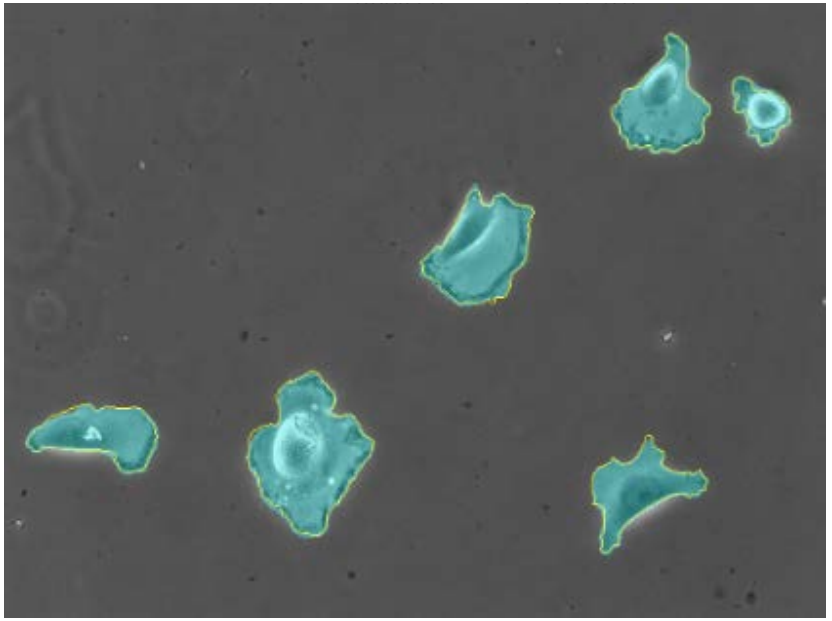
Ronneberger et al. 2015

Encoder-decoder vs. patch-based analysis



Binary segmentation

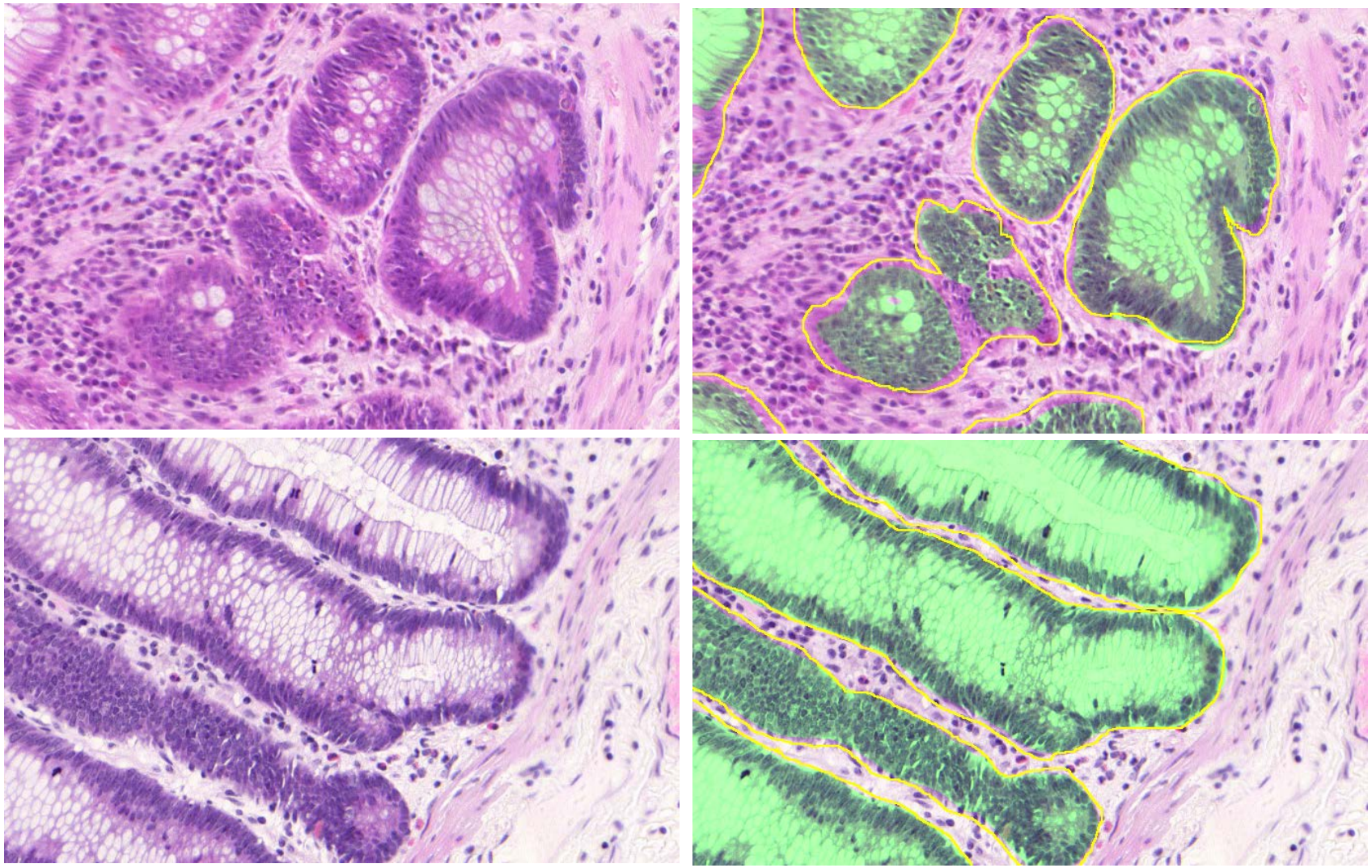
Electron Microscopy



Light microscopy cell tracking

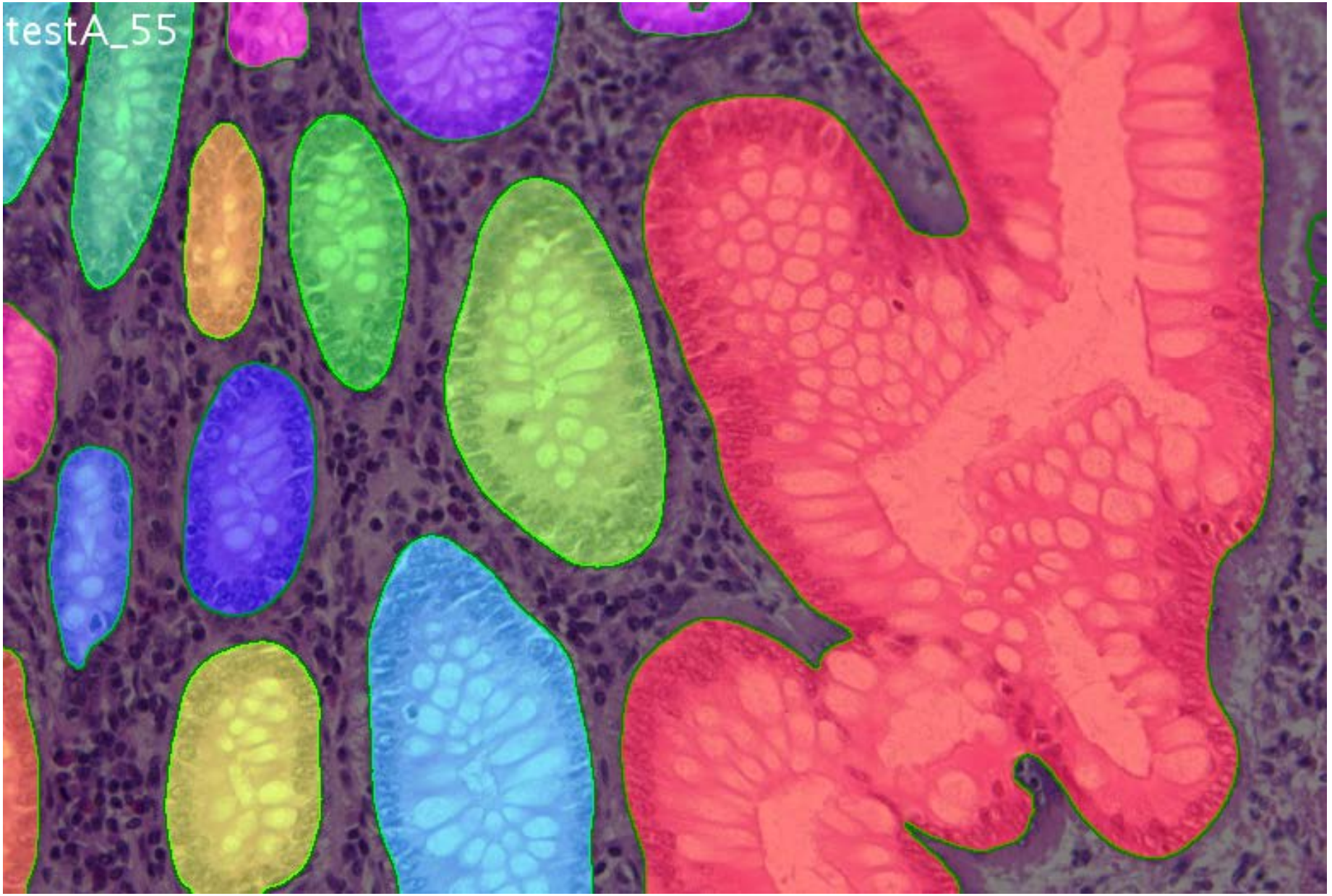
Data from ISBI challenge

Large variety of application domains

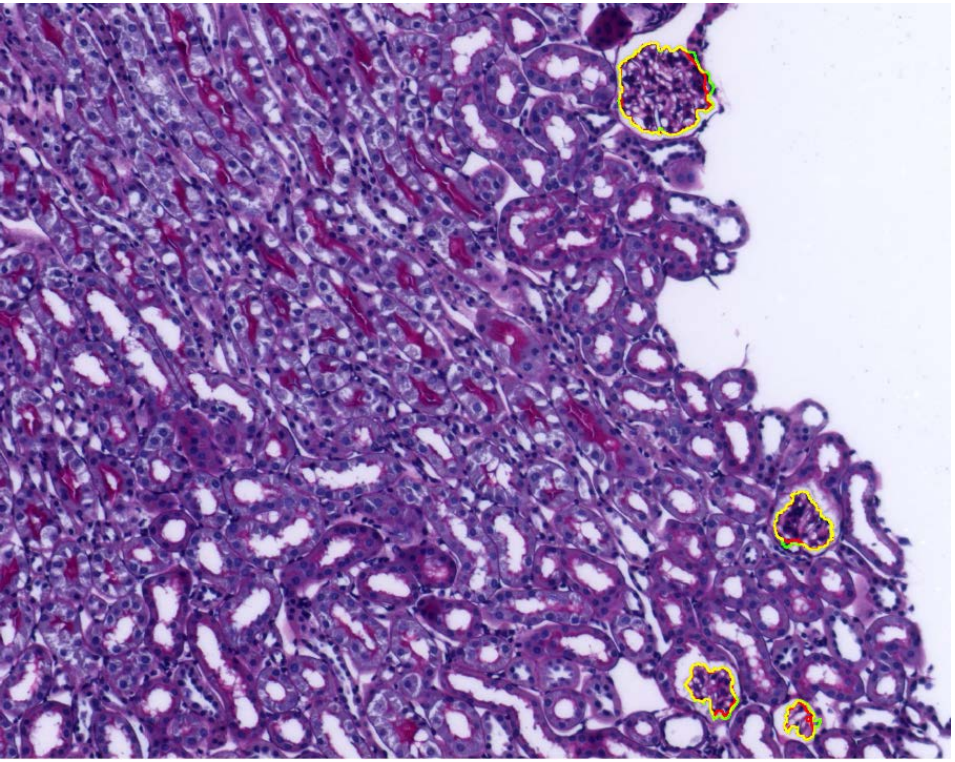


Histopathological data from Gland Segmentation Challenge Contest

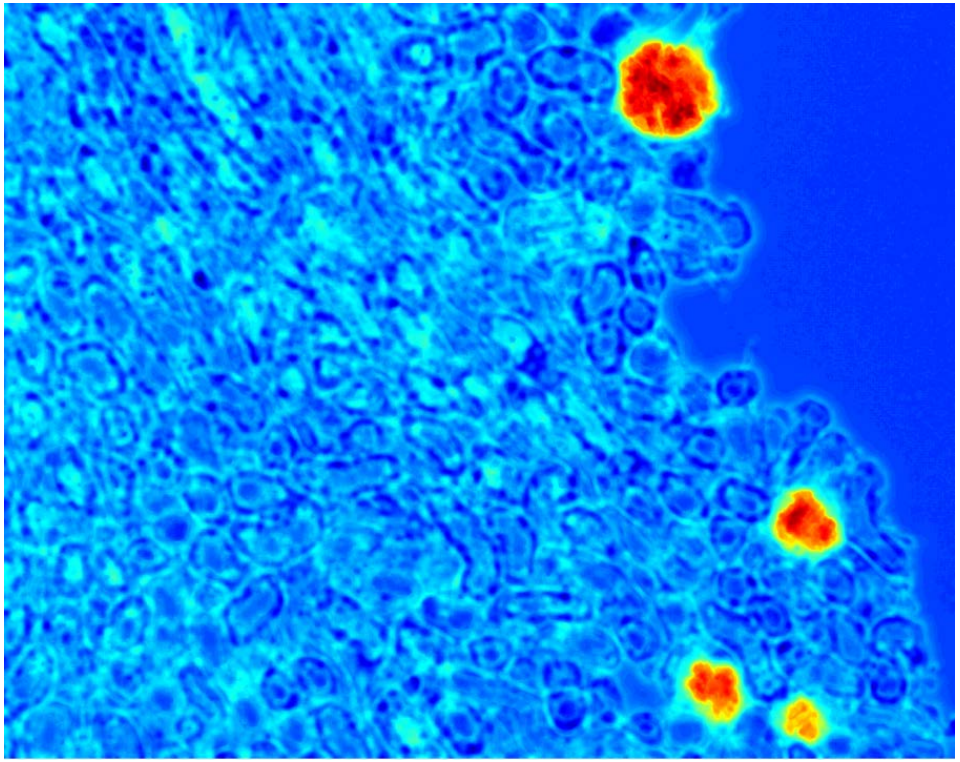
Gland Segmentation Challenge Contest



Segmentation of Glomeruli (Kidney)



green: ground truth,
red: U-Net

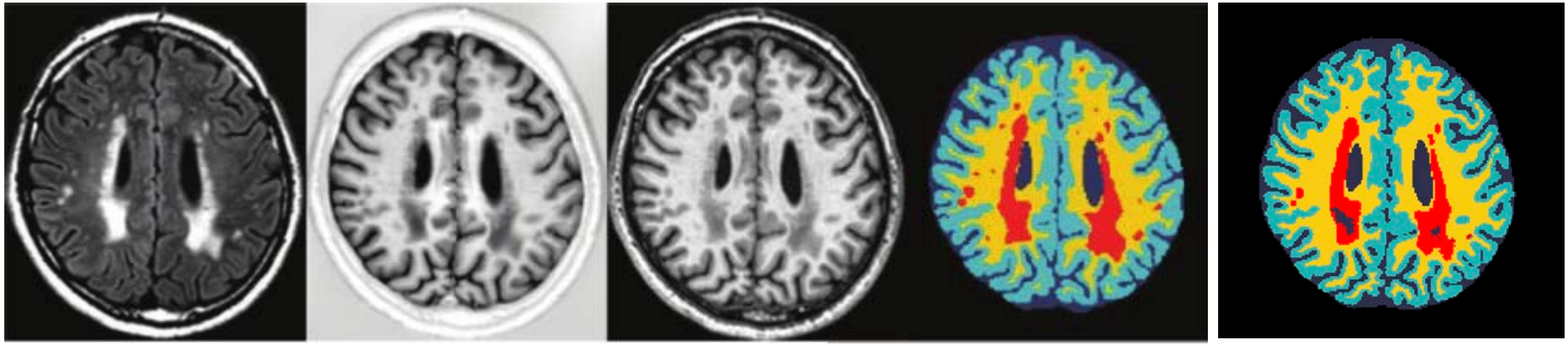


Score map

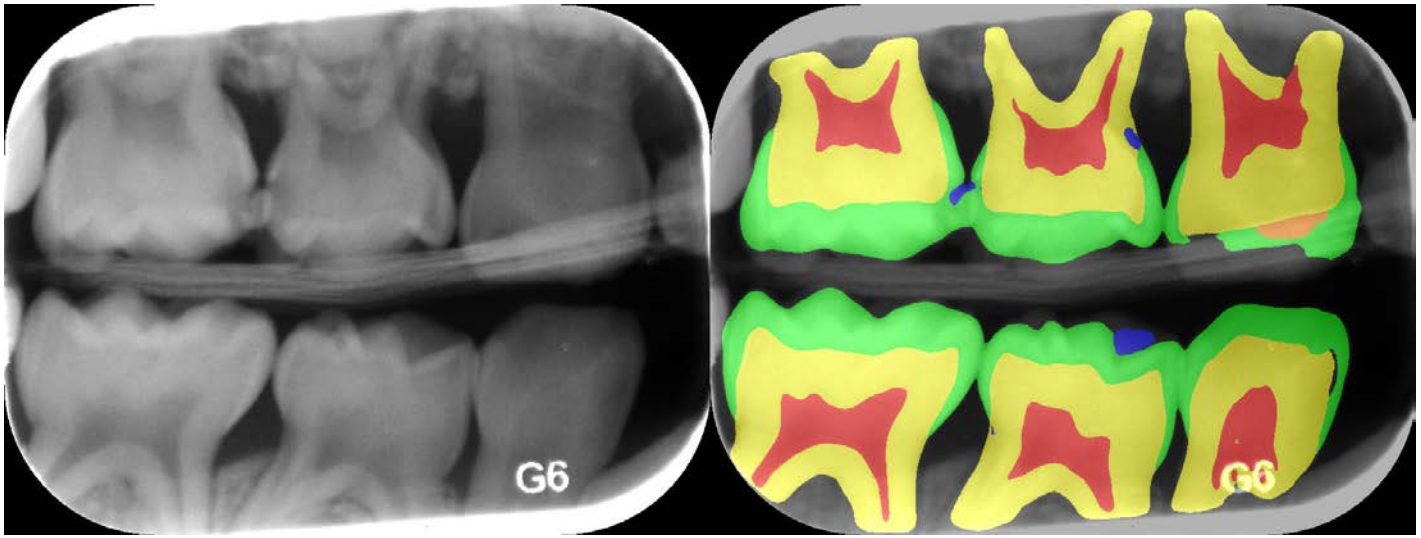
Data from Huber group (Nephrology, Freiburg)

Multi-class segmentation

T2-FLAIR T1-IR T1 Manual seg. U-Net

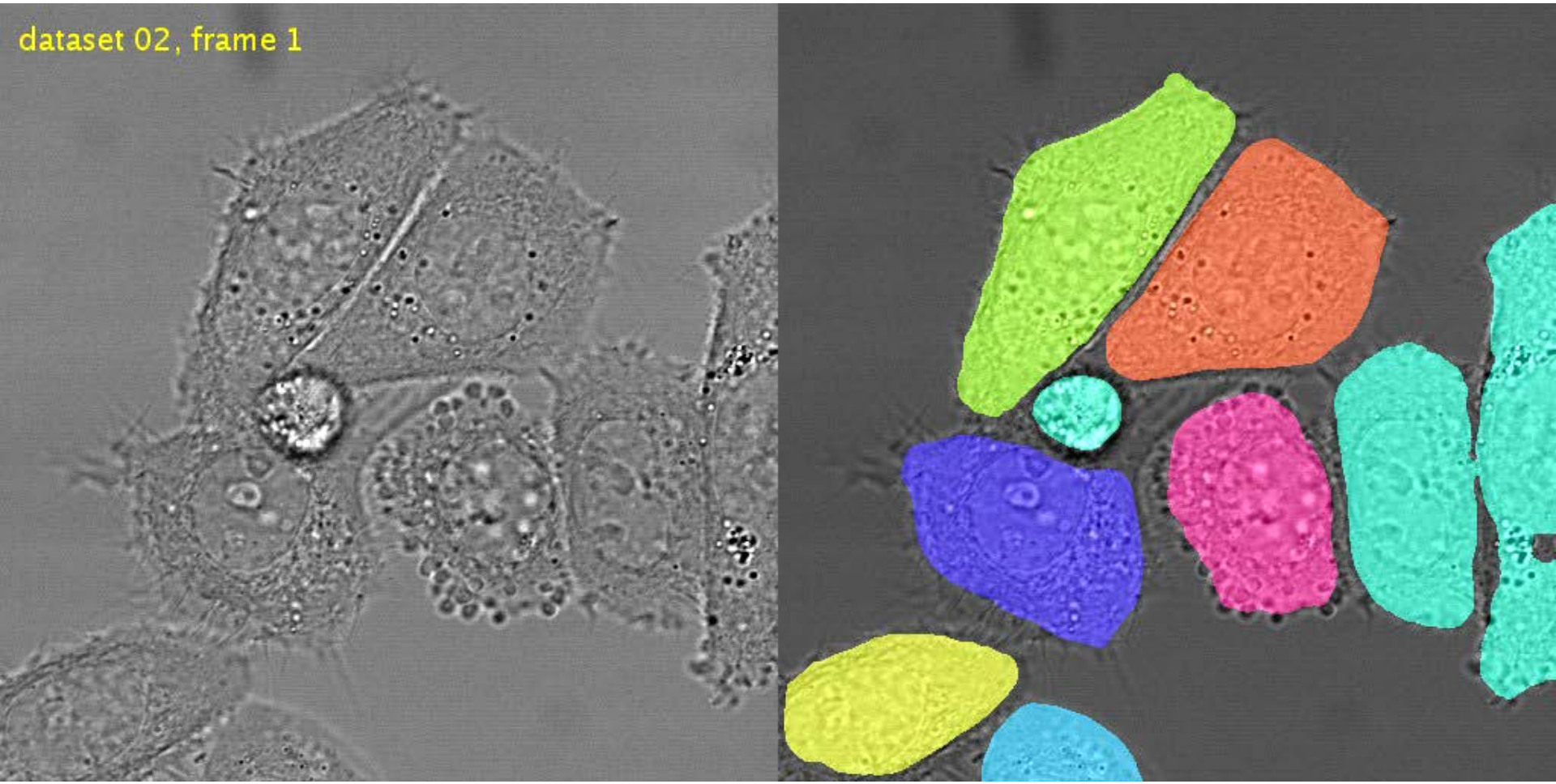


Brain Segmentation (MRBrainS Challenge)



X-ray dental segmentation (ISBI challenge)

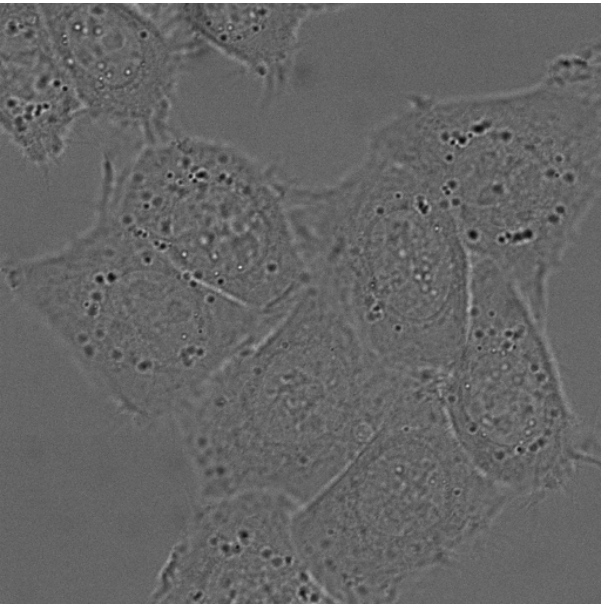
Instance segmentation



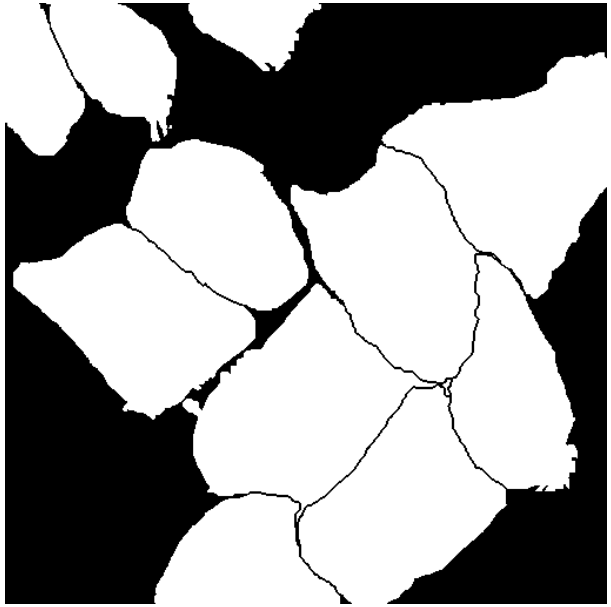
DIC-HeLa cell tracking

Data from ISBI challenge

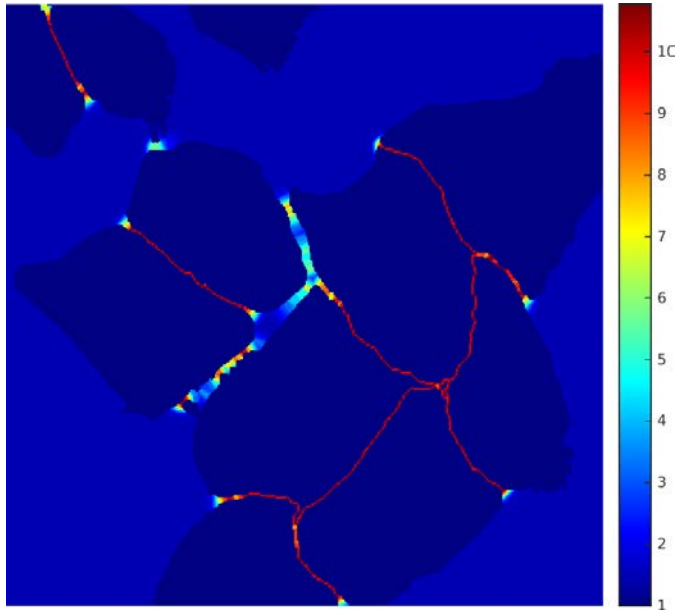
Weighted loss for instance segmentation



Image

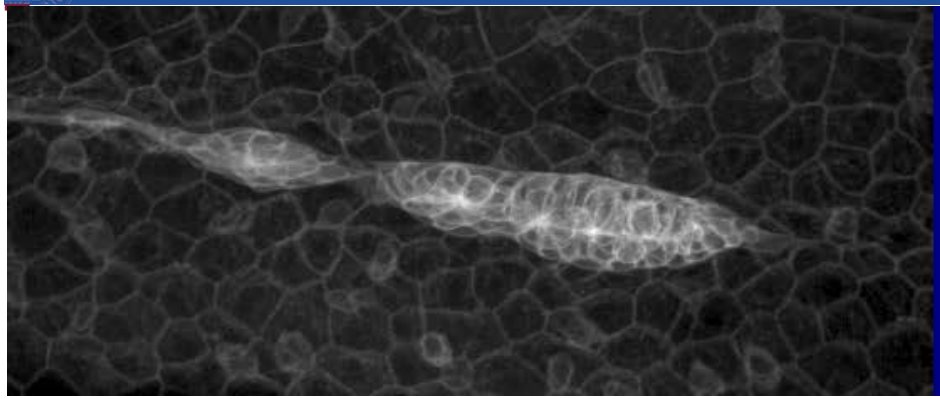
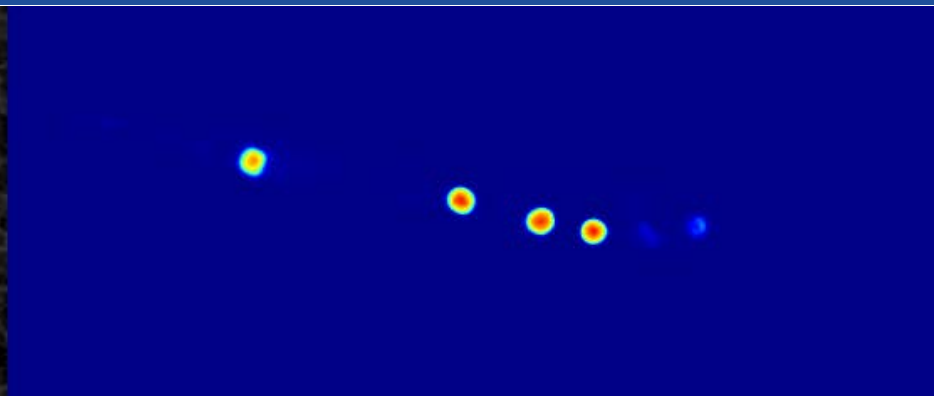
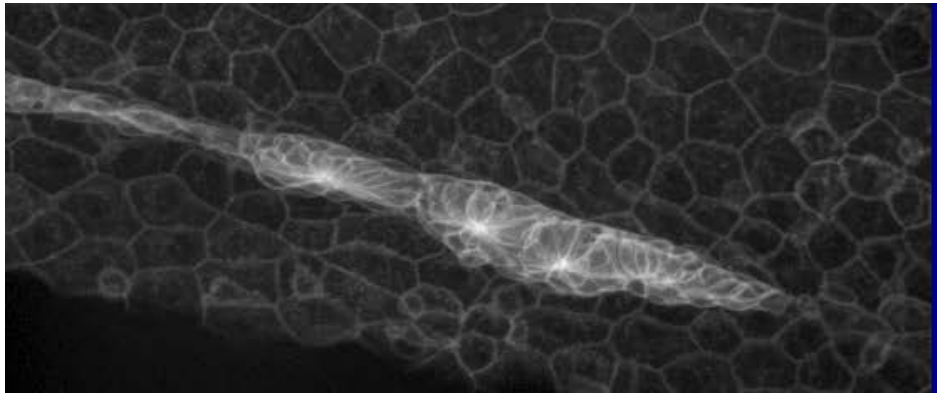
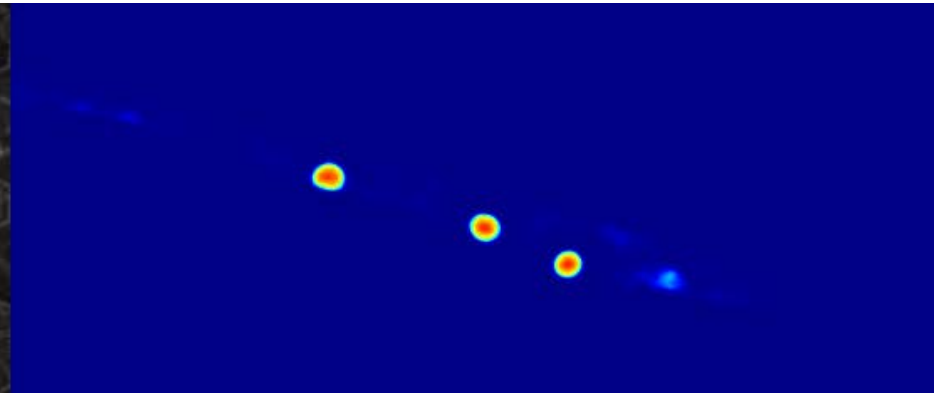
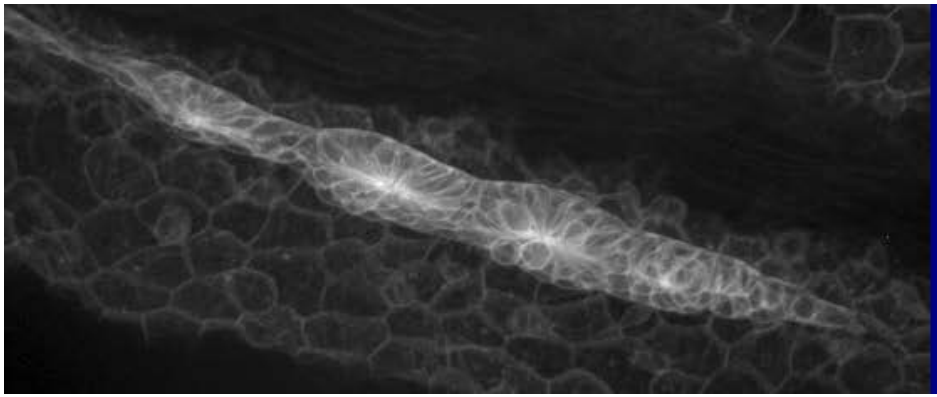
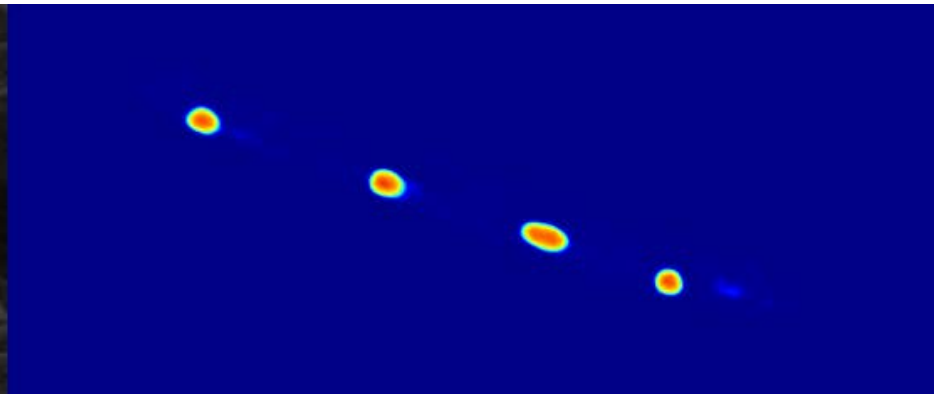


Segmentation mask (background between touching objects)



Weights for the loss

Detection of cell rosettes in the zebrafish primordium

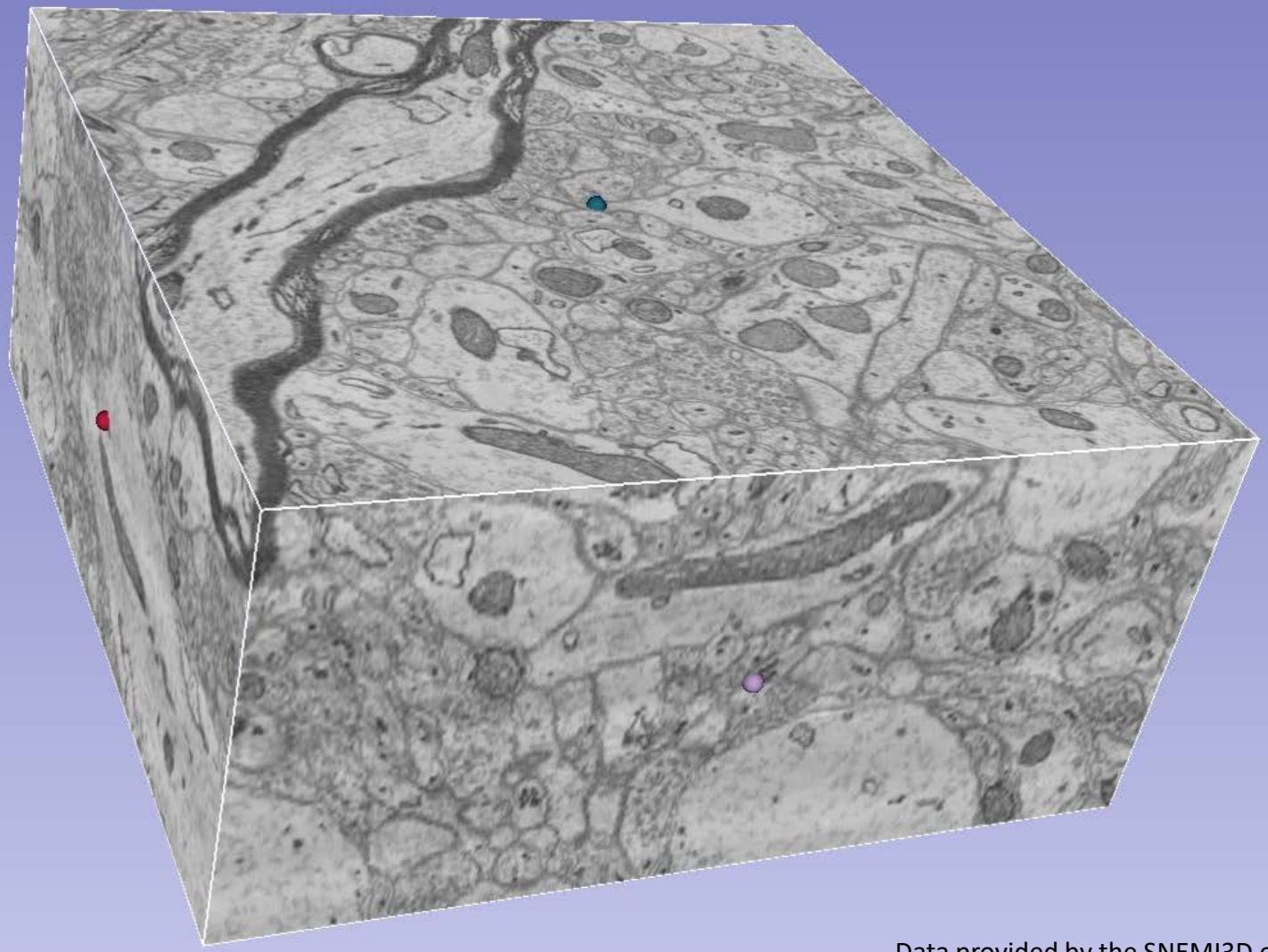







Raw images

Detection score maps

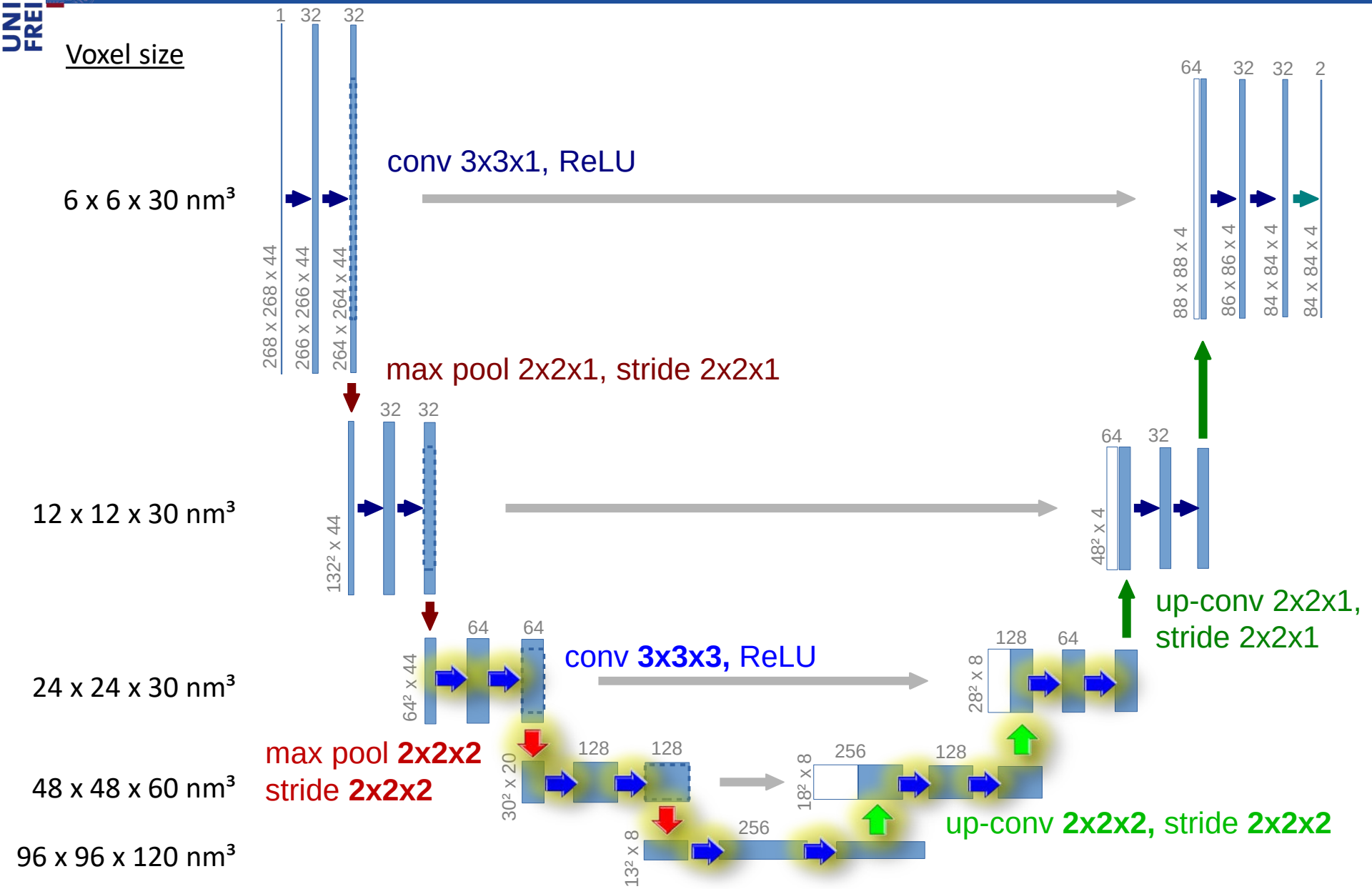
Extension to volumetric data

Serial section electron microscopy of mouse cortex.
1024 x 1024 x 100 voxels (element size: 6 x 6 x 30 nm^3)

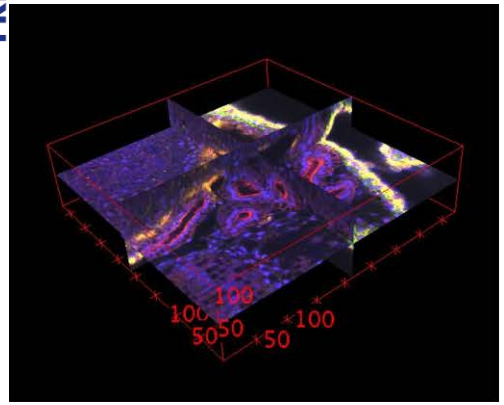


Data provided by the SNEMI3D challenge

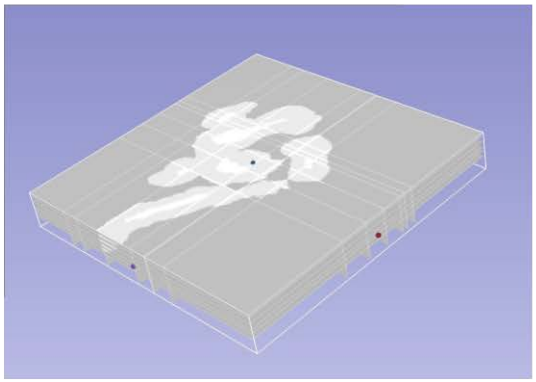
3D U-Net


Voxel size
6 x 6 x 30 nm³
12 x 12 x 30 nm³
24 x 24 x 30 nm³
48 x 48 x 60 nm³
96 x 96 x 120 nm³
conv 3x3x1, ReLU
max pool 2x2x1, stride 2x2x1
conv 3x3x3, ReLU
max pool 2x2x2 stride 2x2x2
up-conv 2x2x1, stride 2x2x1
up-conv 2x2x2, stride 2x2x2
1
32
32
64
32
32
2
268 x 268 x 44
266 x 266 x 44
264 x 264 x 44
88 x 88 x 4
86 x 86 x 4
84 x 84 x 4
84 x 84 x 4
132² x 44
48² x 4
64² x 44
28² x 8
30² x 20
18² x 8
13² x 8
64
128
256

Semi-automated or automated segmentation

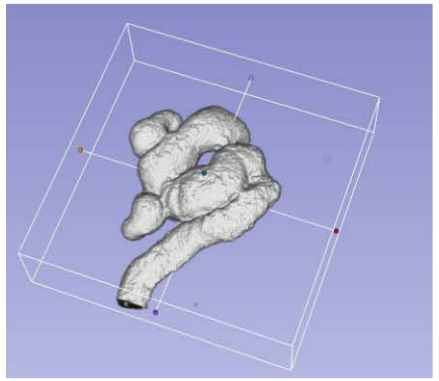


raw image

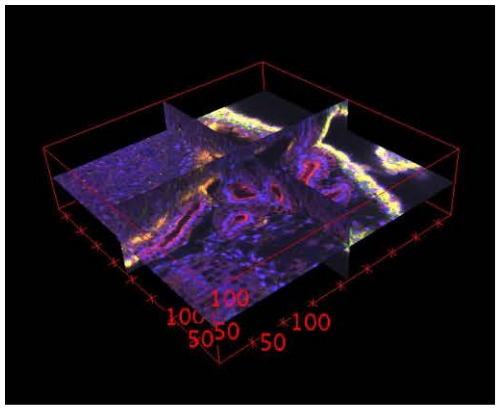


manual sparse annotation

train and apply 3D u-net

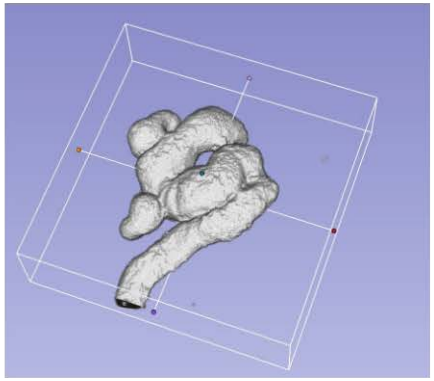


dense segmentation



raw image

apply trained 3D u-net

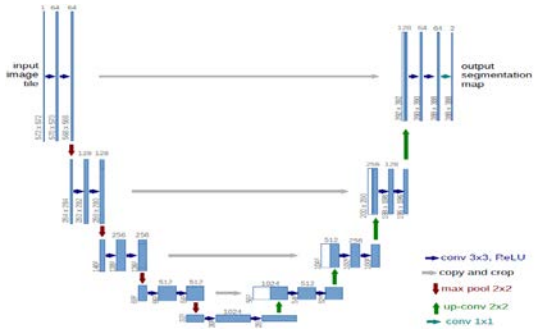


dense segmentation

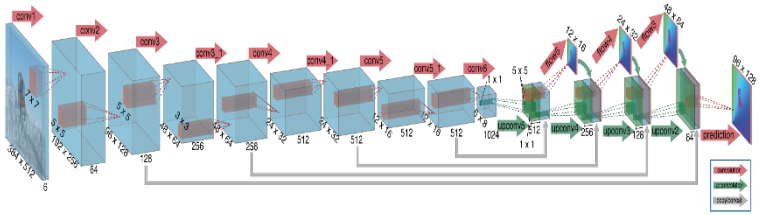
Cicek et al. 2016

End of Part I

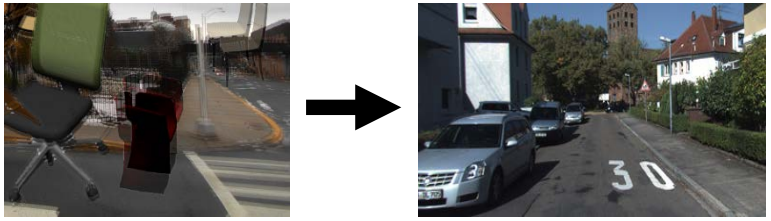
Overview



- Part I: Encoder-decoder networks

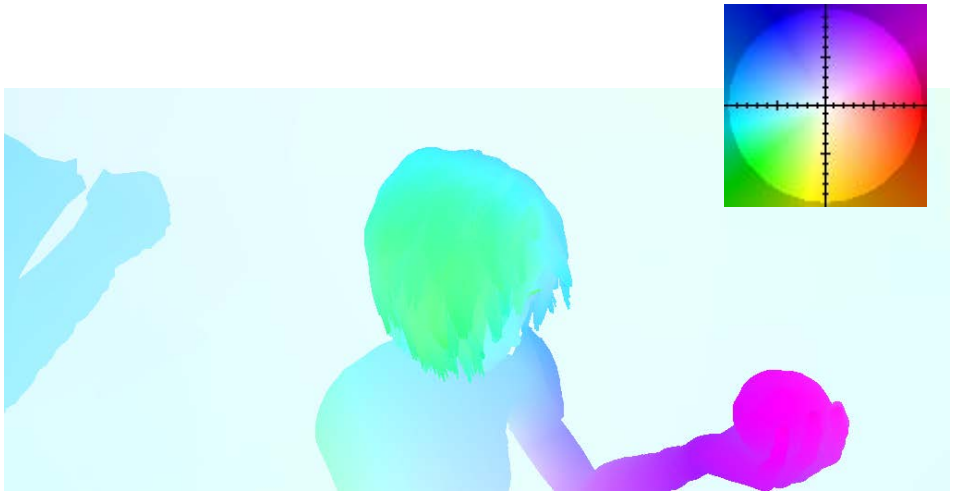


- Part II: Correspondence estimation with FlowNet



- Part III: Cross-dataset generalization

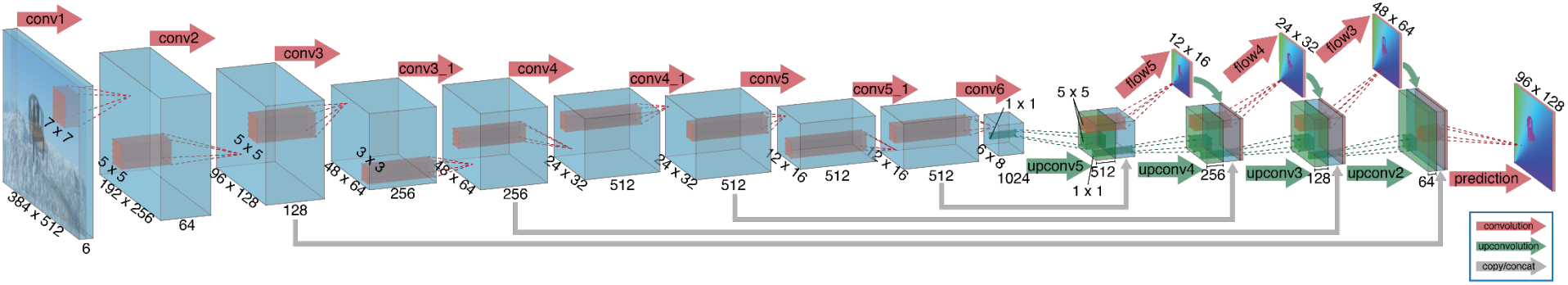
Correspondence estimation



Displacement fields appear in:

- Optical flow estimation (motion)
- Disparity estimation in stereo images (depth)
- Image registration

FlowNet: estimating optical flow with a ConvNet



- Can networks learn to find correspondences?
- Learning task different from semantic tasks
  (classification, segmentation, etc.)
- Same encoder-decoder architecture as U-Net

Dosovitskiy et al. 2015

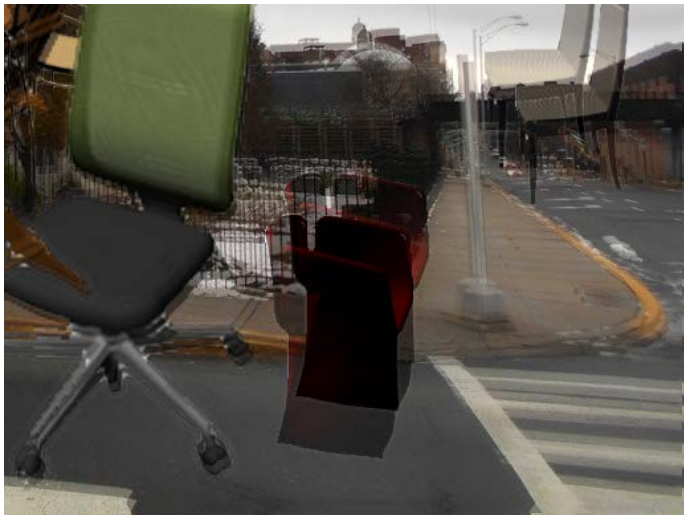
Enough data to train such a network?

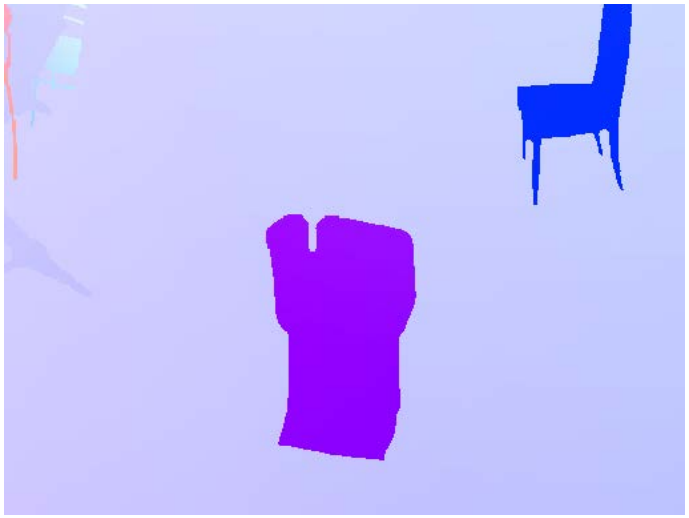
- Getting ground truth dense correspondences is hard

- Existing datasets are small:

	Frames with ground truth
Middlebury	8
KITTI	194
Sintel	1041
Wanted	>10000

Realism is overrated: the “Flying Chairs” dataset





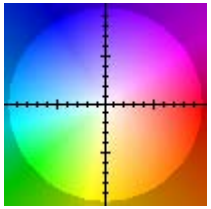


Image pair

Optical flow

Generalization to other images

FlowNetS

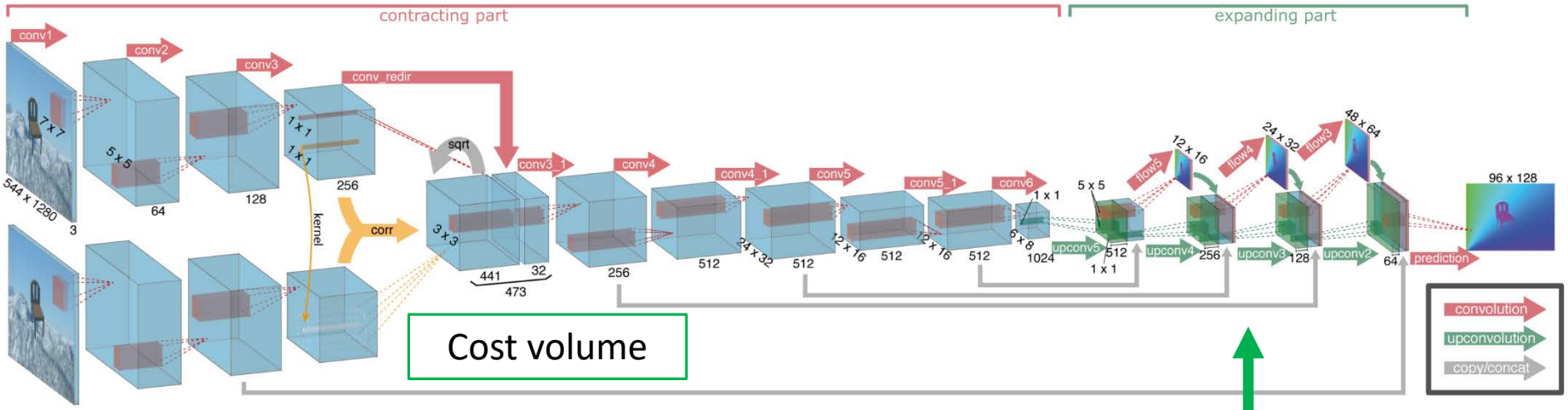
Ground truth

Although the network has only seen flying chairs for training, it predicts optical flow also on other data

FlowNetC: explicit correlation layer

Dosovitskiy et al. 2015

FlowNetC

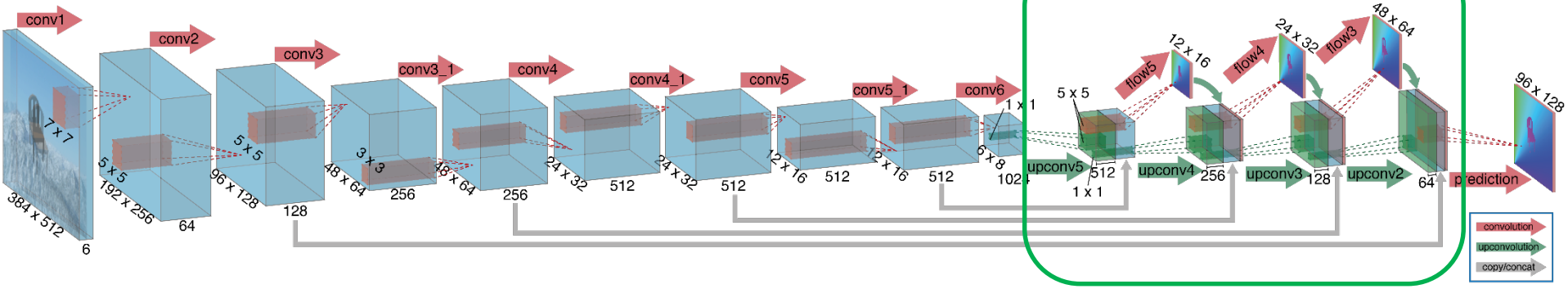


Cost volume

Siamese network

identical

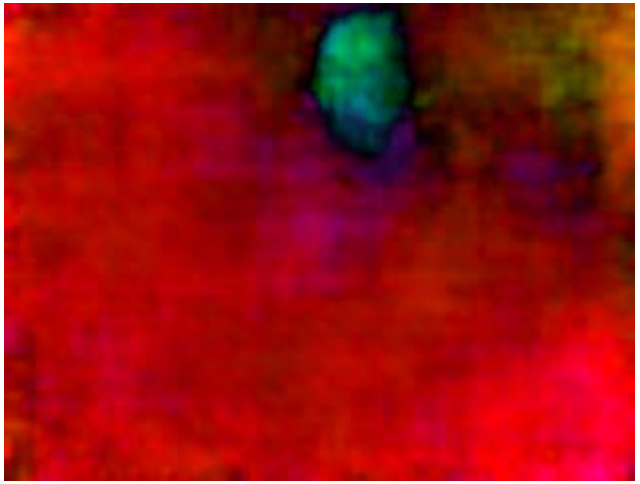
FlowNetS



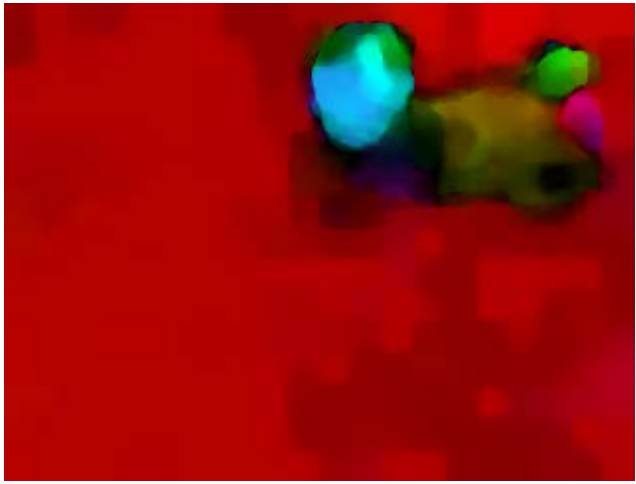
FlowNetS and FlowNetC in practice

Underperformer on small motion

Example from UCF 101

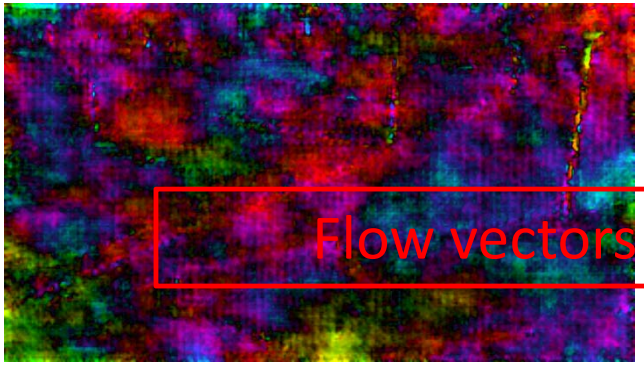


FlowNet



EpicFlow

Identical images

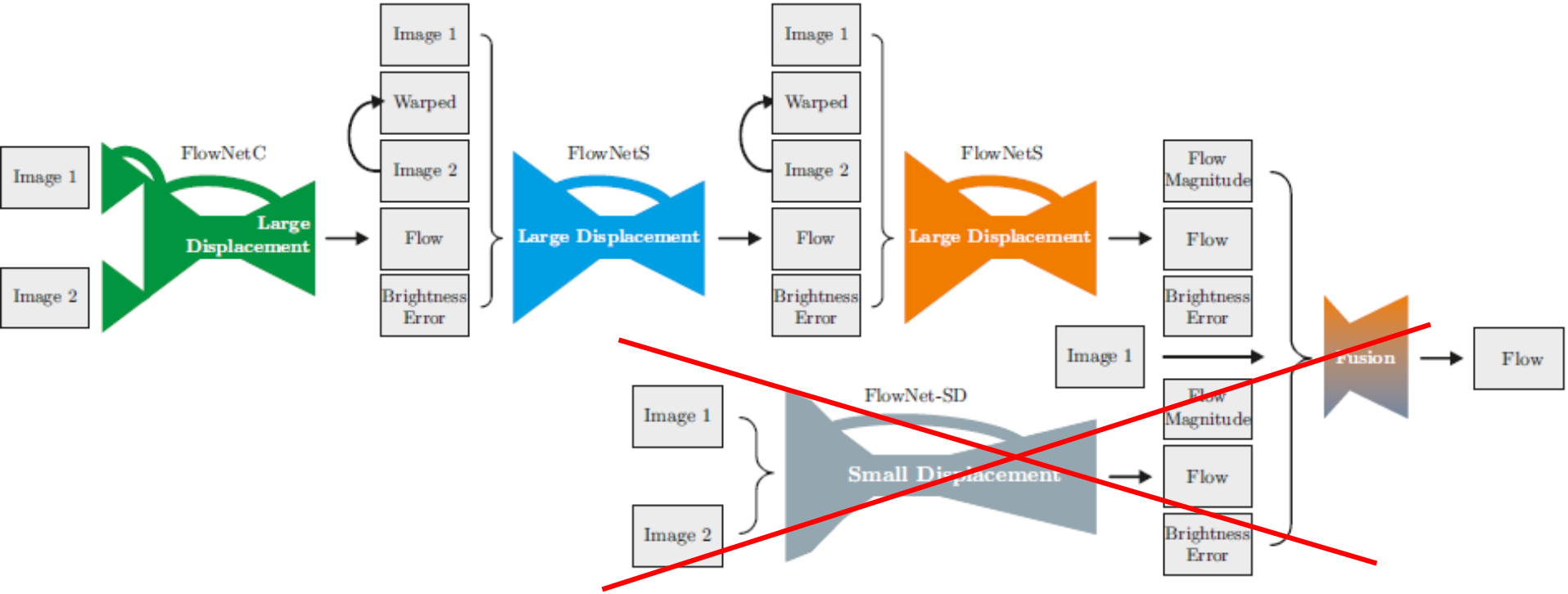


FlowNet

EpicFlow

FlowNet 2.0

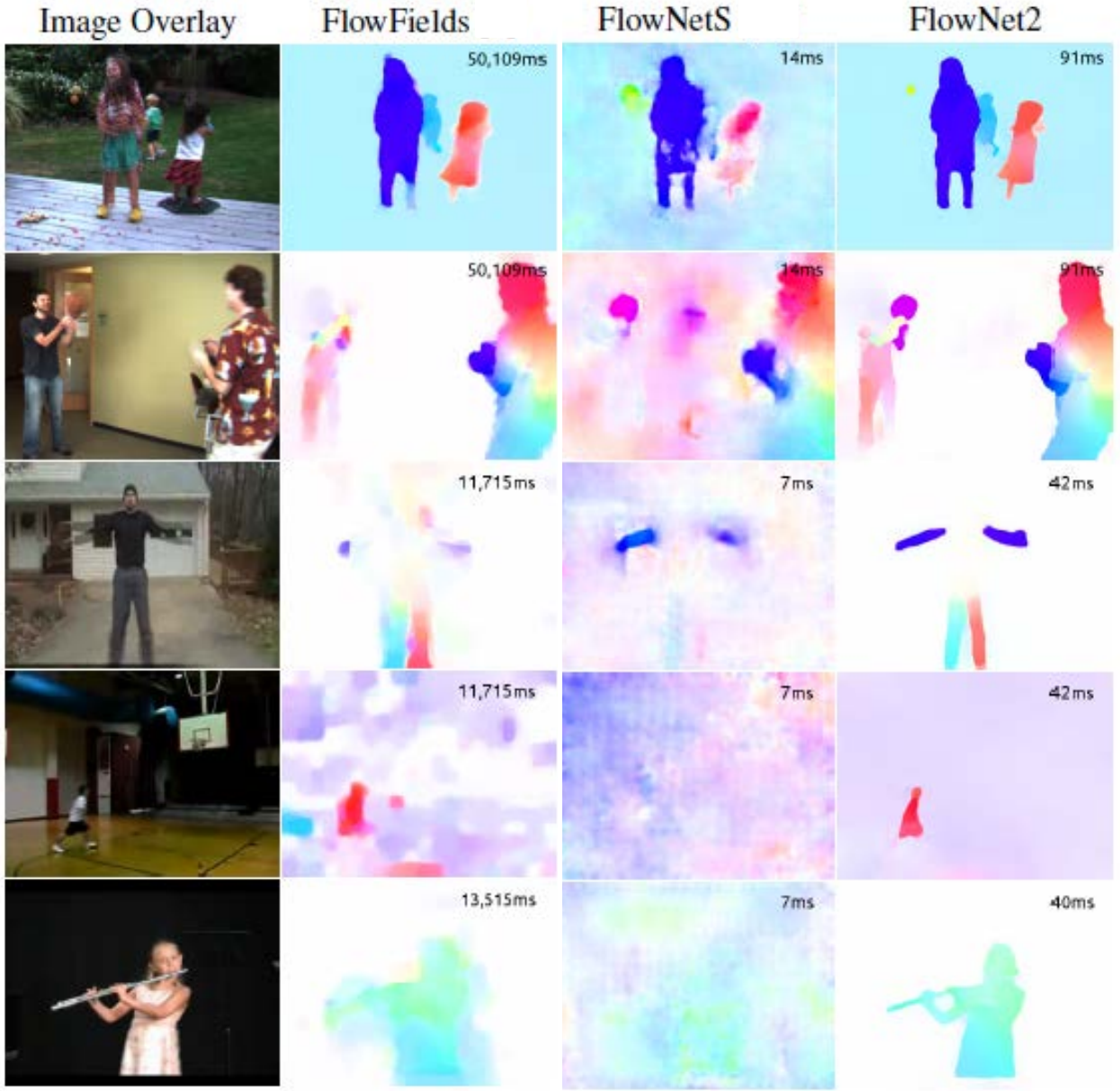
Ilg et al. 2017



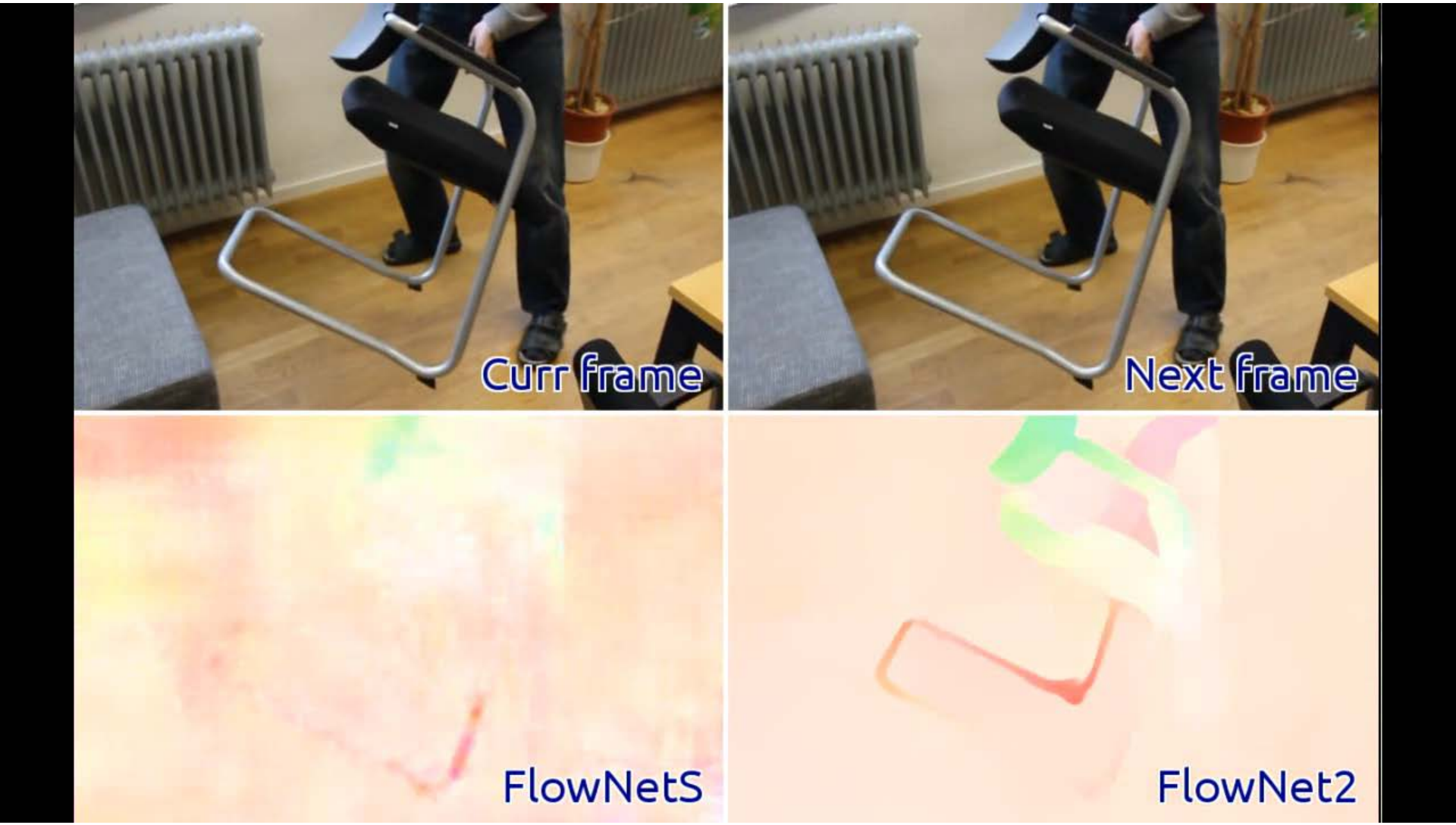
Major changes:

- Stacking of networks with motion compensation
- Improved data and training schedules

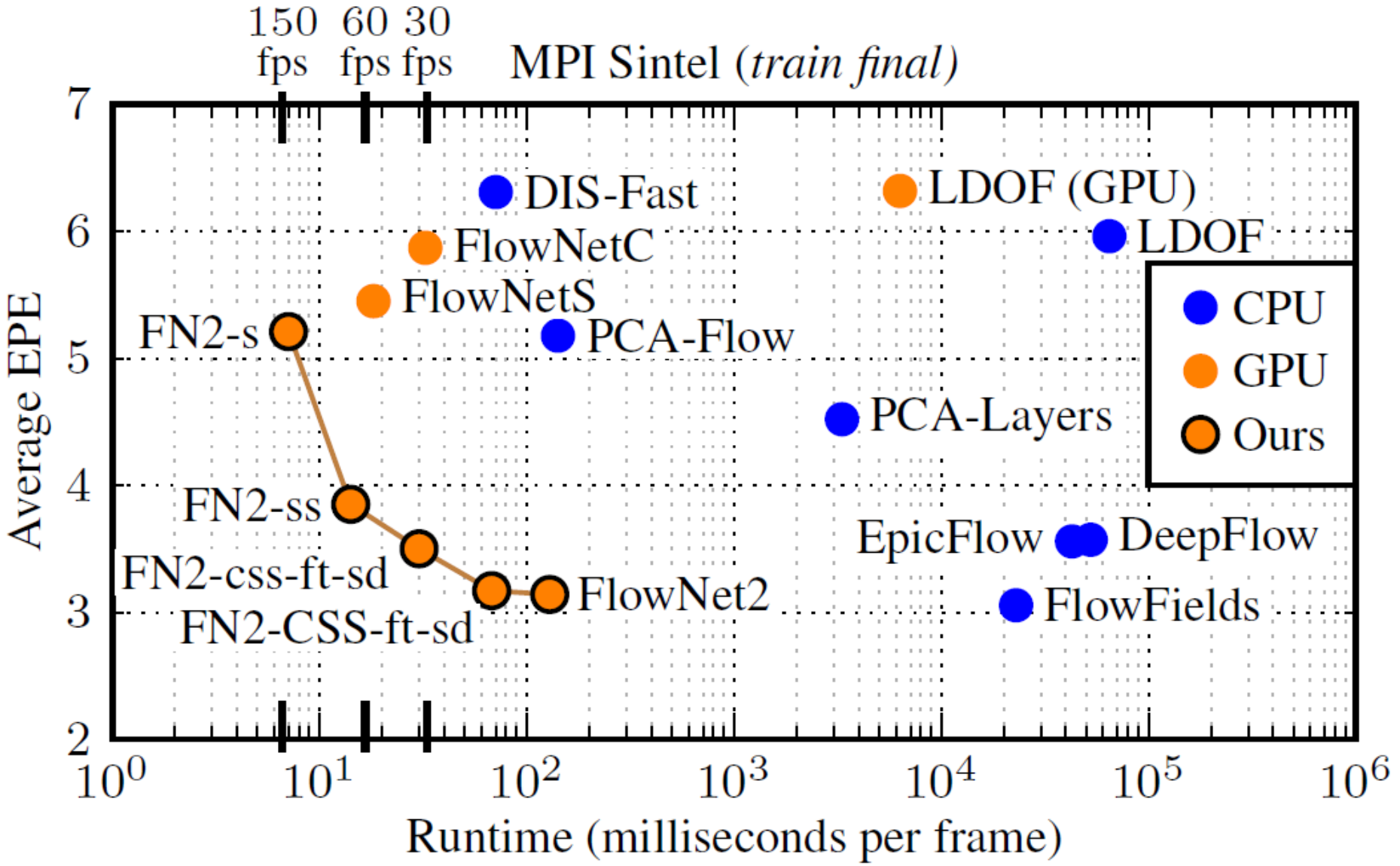
FlowNet vs. FlowNet 2.0



FlowNet vs. FlowNet 2.0



Accuracy-runtime tradeoff

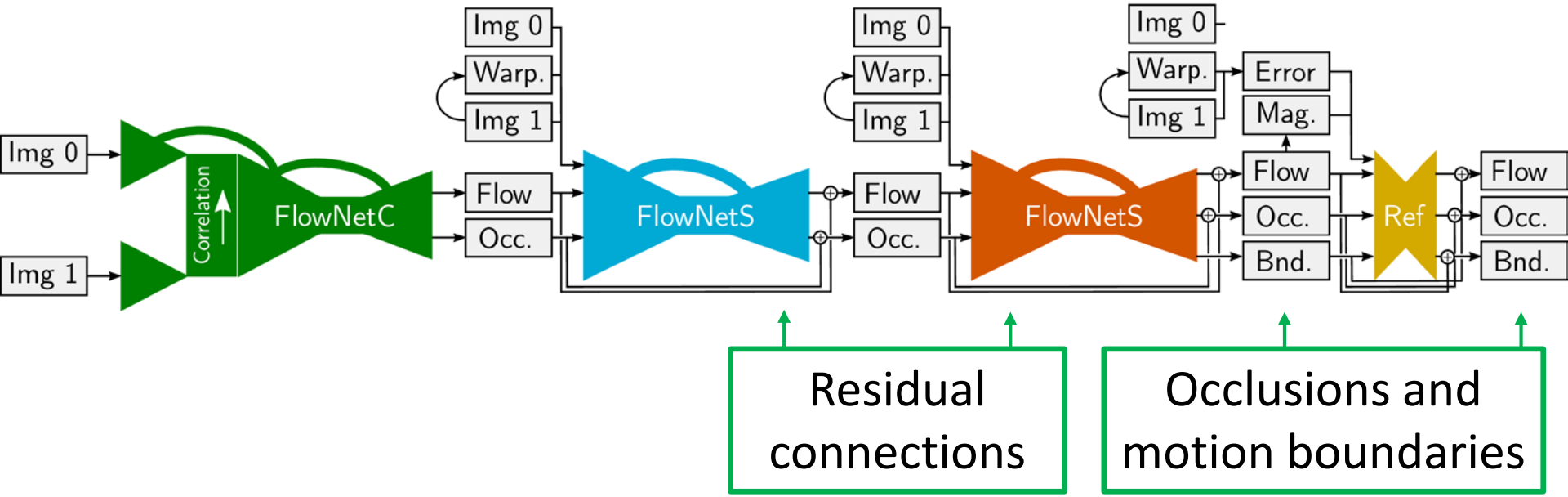


State-of-the-art optical flow estimation

	Sintel	KITTI	Runtime
FlowFields (Bailer et al. 2015)	5.8	18.7%	22810 ms
DC Flow (Xu et al. 2017)	5.1	14.9%	9000 ms
FlowNet (Dosovitskiy et al. 2015)	7.5	-	18 ms
FlowNet 2.0 (Ilg et al. 2017)	5.7	**8.6%**	77 ms
PWC-Net (Sun et al. 2018)	**5.0**	9.8%	30 ms

Occlusion and motion boundary estimation

Ilg et al. 2018



Residual connections

Occlusions and motion boundaries

Extra output for free!

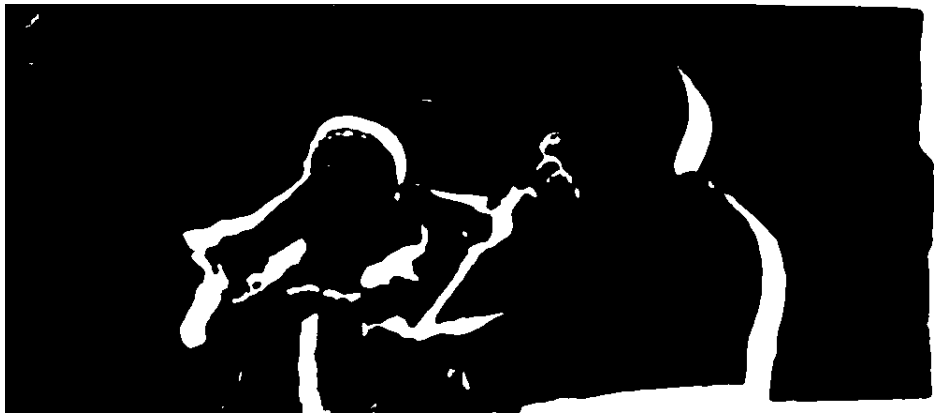
Occlusion estimation

Occlusion estimation

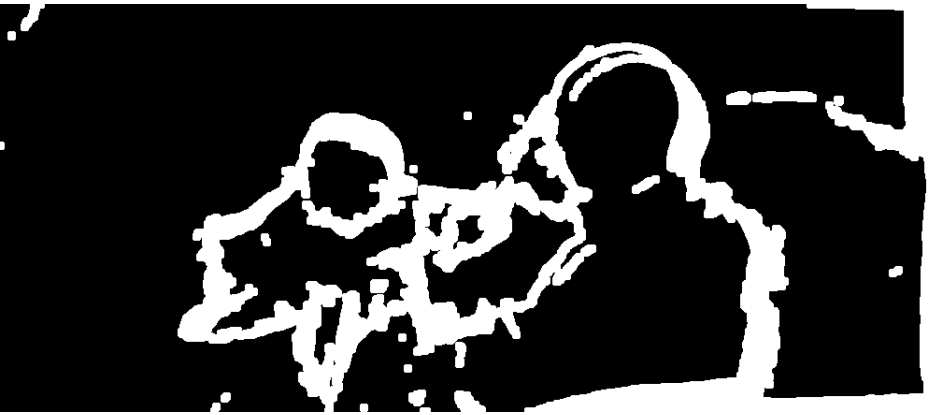
FlowNet-Occ

Occlusion estimation

Ground truth



FlowNet-Occ (Ilg et al. 2018)



Mirrorflow (Hur&Roth 2017)

S2DFlow (Leordeanu et al. 2013)

Occlusion estimation

	F-Measure
Forward-backward consistency (FlowNet)	0.38
MirrorFlow (Hur&Roth 2017)	0.39
S2D Flow (Leordeanu et al. 2013)	0.47
FlowNet with occlusion estimation (Ilg et al. 2018)	**0.70**

How reliable is the estimated optical flow?

Uncertainties and Bayesian deep learning

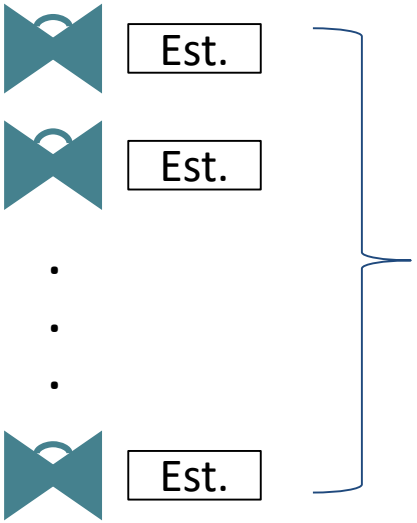
- Bayesian Neural Networks
  - MCMC [MacKay 1992, Neal 1996, Welling and Teh 2011, Chen et al. 2014]
  - Variational inference [Graves 2011, Blundell et al. 2015]
  - Probabilistic backprop [Hernandez-Lobato and Adams 2015]

Typically intractable with large networks

  - MC dropout
    [Gal and Ghahramani 2016, Kendall and Gal 2017]

Other options to estimate uncertainty efficiently?

Ensembles



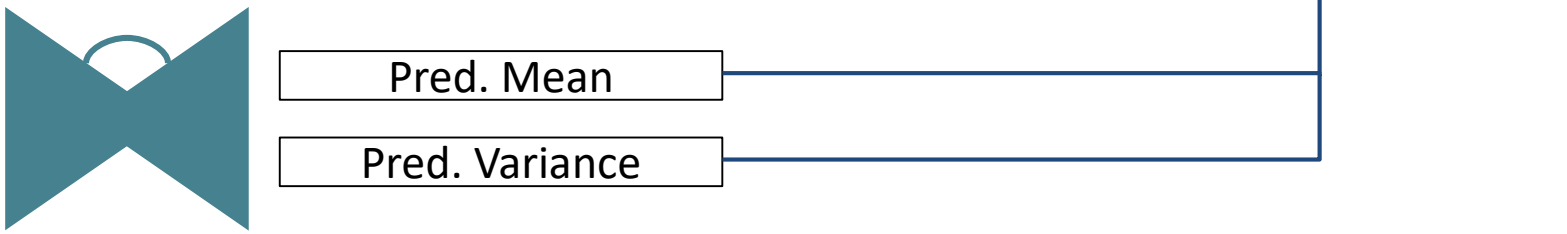
$$\mu(f(\mathbf{x})|\mathcal{D}) = \frac{1}{M}\sum_{i=1}^{M}\hat{f}(\mathbf{x}|\mathbf{w}_i,\mathcal{D})$$

$$\sigma^2(f(\mathbf{x})|\mathcal{D}) = \frac{1}{M}\sum_{i=1}^{M}(\hat{f}(\mathbf{x}|\mathbf{w}_i,\mathcal{D}) - \mu(f(\mathbf{x})|\mathcal{D}))^2$$

Train and run M networks to get a mean and a variance estimate

Predictive uncertainty

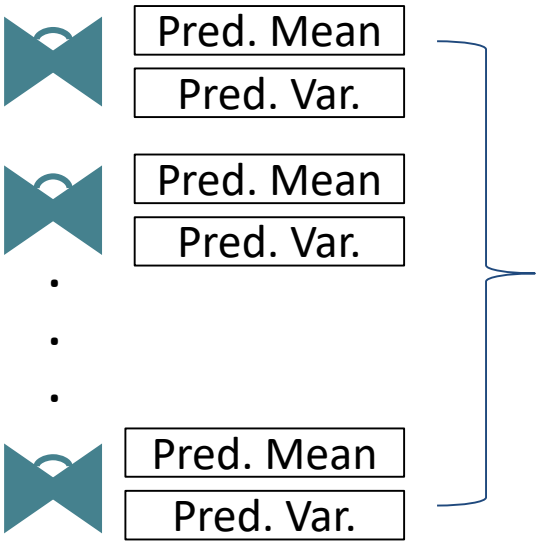
$$p(f(\mathbf{x}) \mid \mathbf{w}) = p(f(\mathbf{x}) \mid \boldsymbol{\theta}(\mathbf{x}, \mathbf{w}))$$



Train a single network that predicts parameters (mean and variance) of a parametric distribution

→ Maximize the log-likelihood of the training set

Predictive ensembles



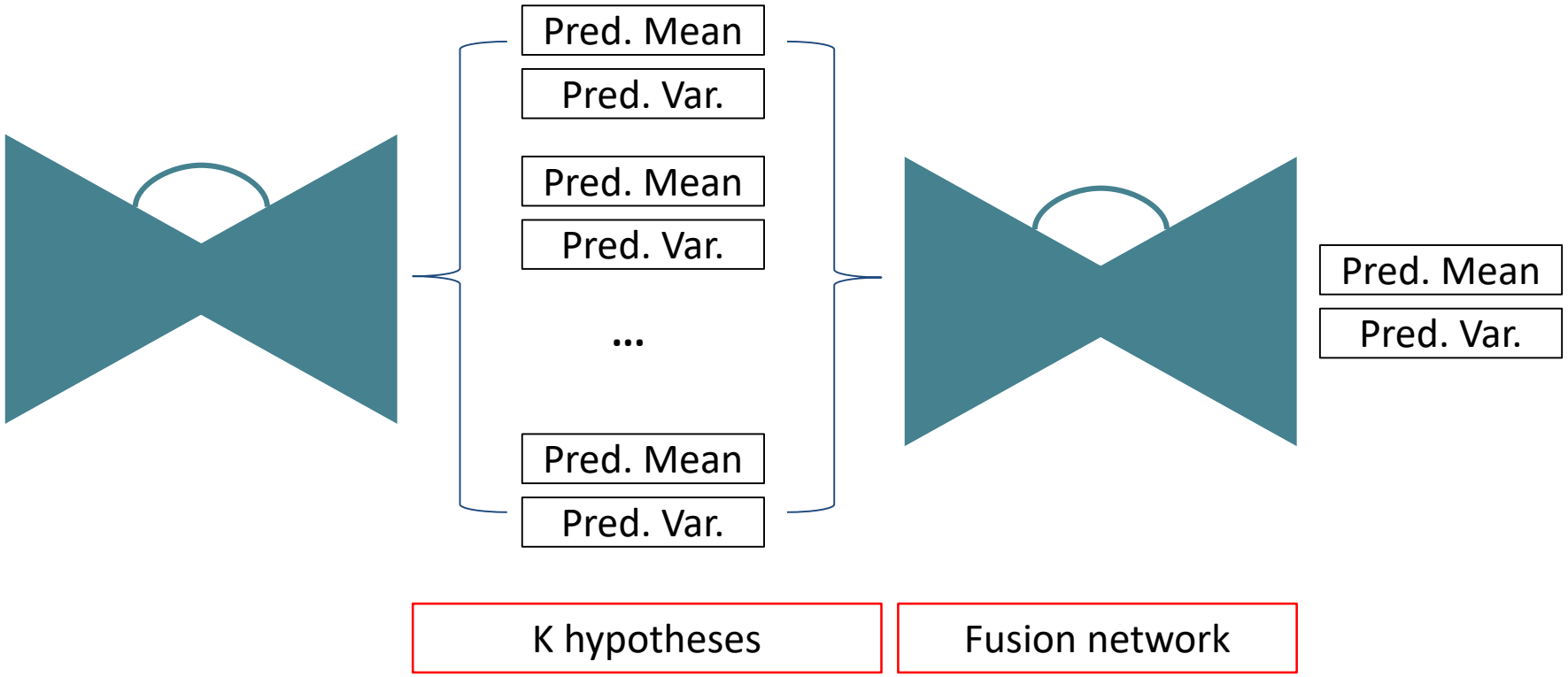
$$\mu(f(\mathbf{x})|\mathcal{D}) = \frac{1}{M}\sum_{i=1}^{M}\mu_i$$

$$\sigma^2(f(\mathbf{x})|\mathcal{D}) = \frac{1}{M}\sum_{i=1}^{M}(\mu_i - \mu(f(\mathbf{x})|\mathcal{D}))^2 + \sigma_i^2$$

Train and run M predictive networks each trained on the log-likelihood

FlowNetH – multiple hypotheses in one network

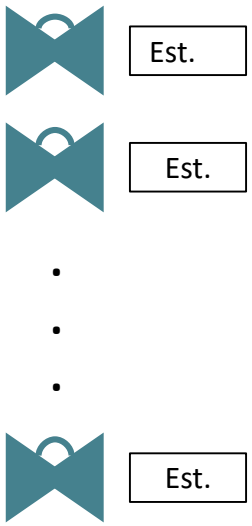
Ilg et al. 2018



FlowNetH produces multiple diverse hypotheses, from which it locally selects the best → no ensemble needed

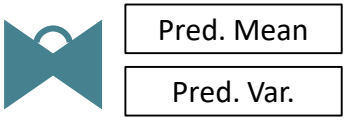
Overview

Empirical Ensembles

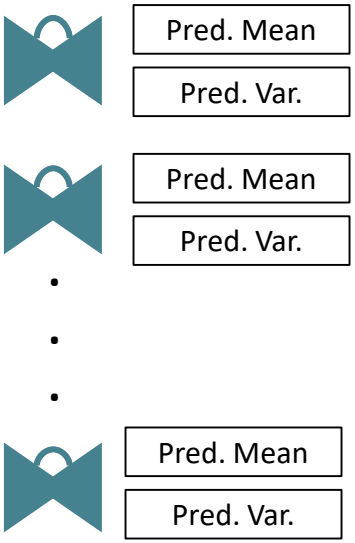


MC Dropout
Bootstrapped Ensembles
Snapshot Ensembles

Predictive Distribution

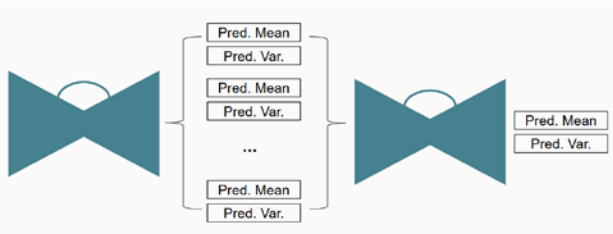


Predictive Ensembles

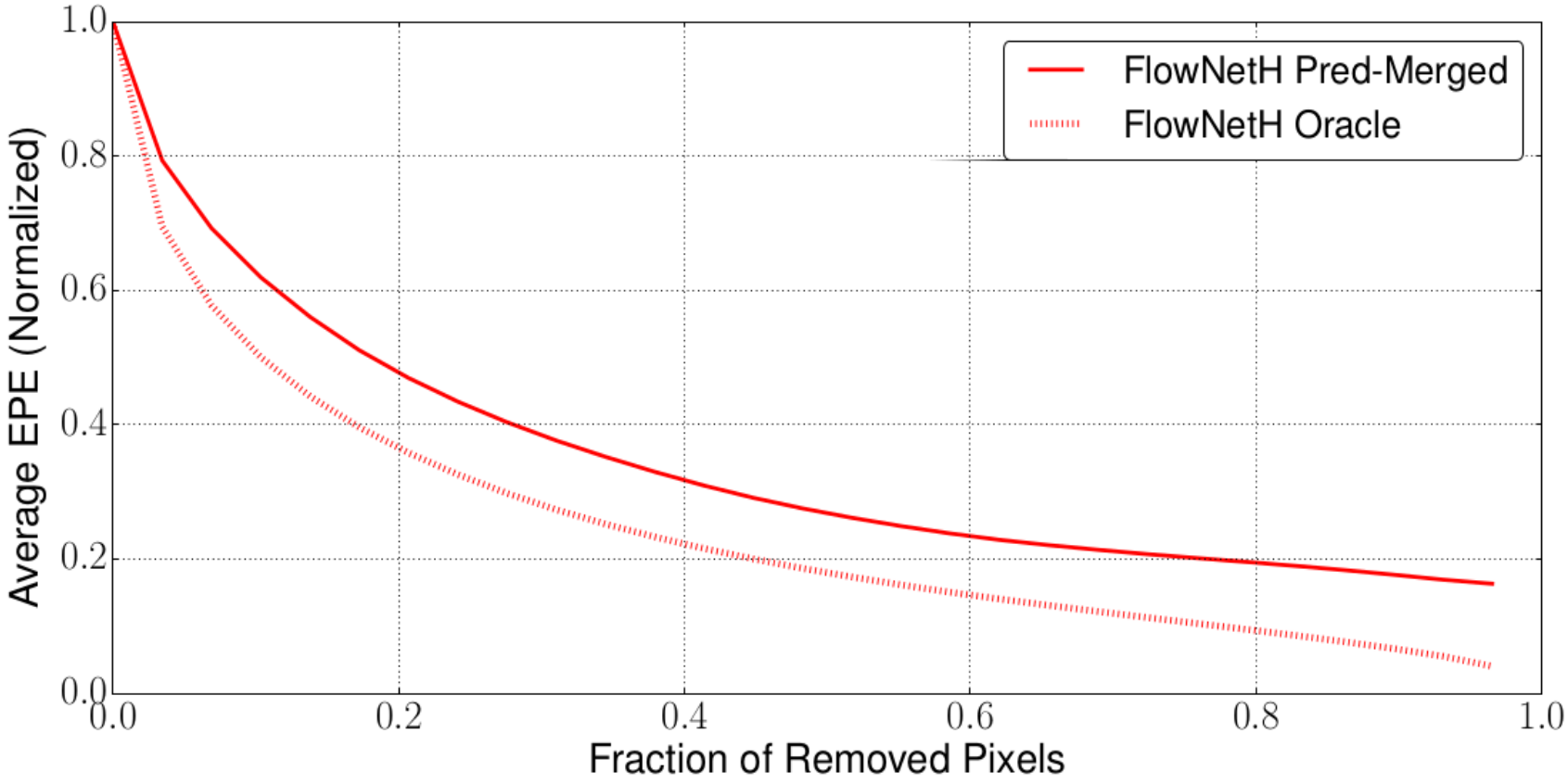


MC Dropout
Bootstrapped Ensembles
Snapshot Ensembles

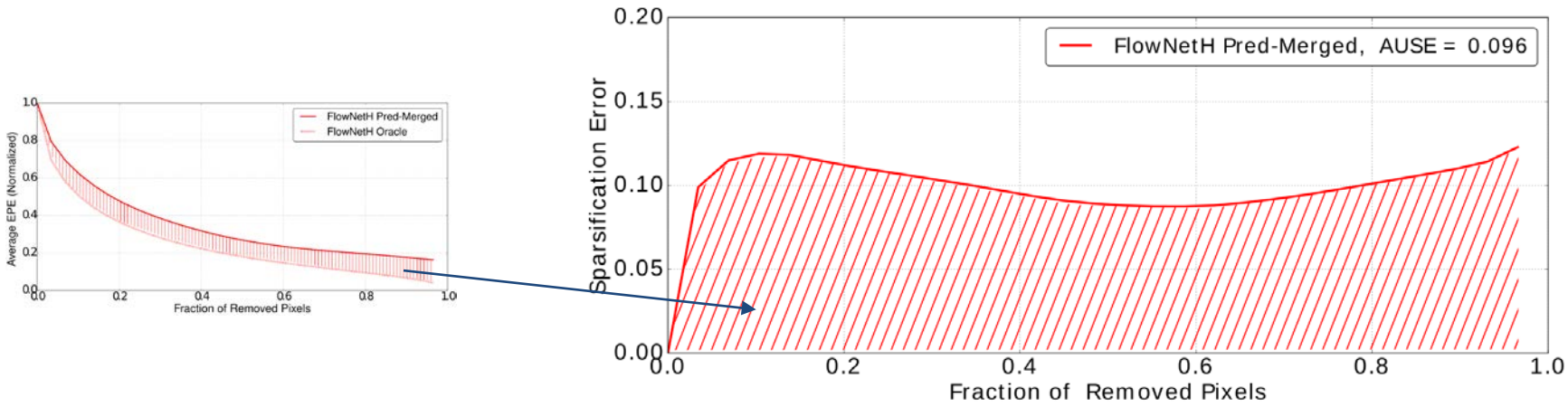
FlowNetH



Sparsification plot



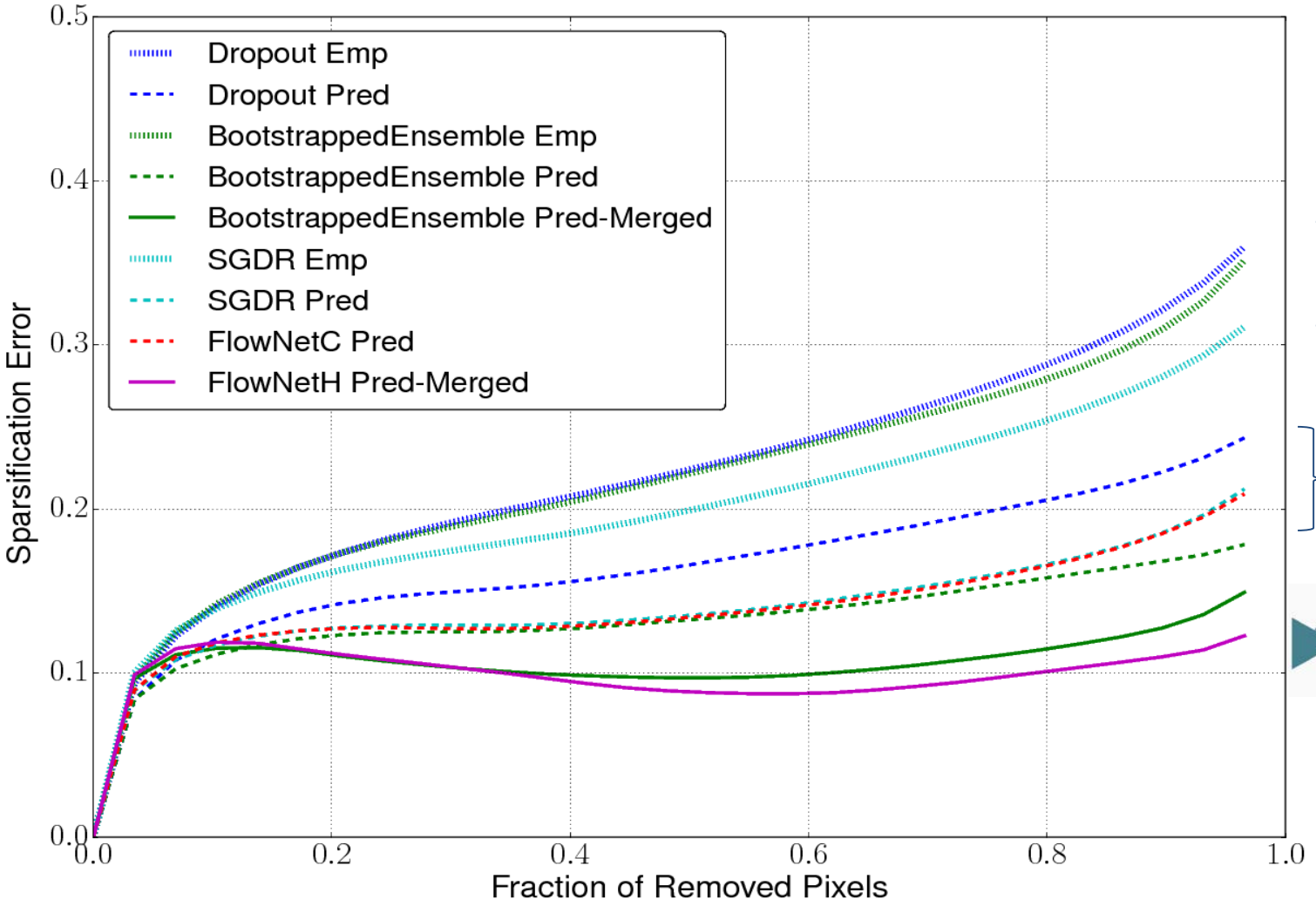
Area under sparsification error (AUSE)

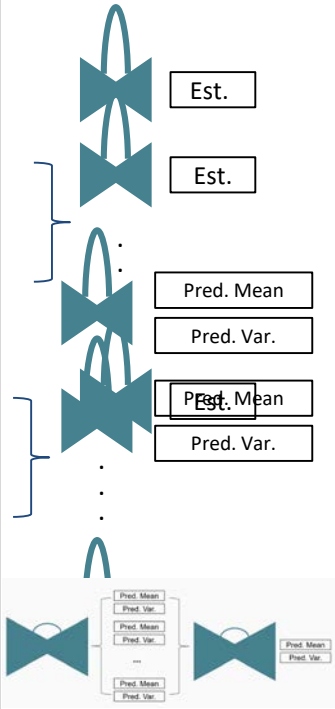


→ Model independent

→ Single benchmark number

Results on Sintel-train-clean

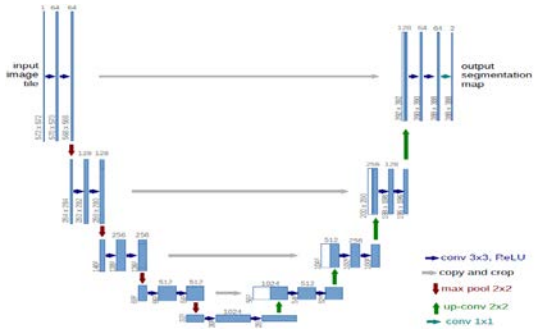




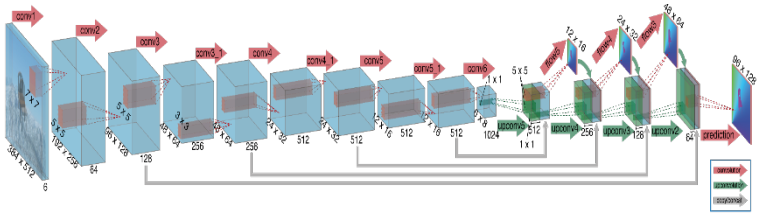
Results on real videos

End of Part II

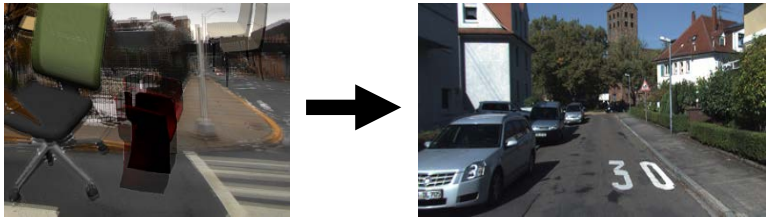
Overview



- Part I: Encoder-decoder networks



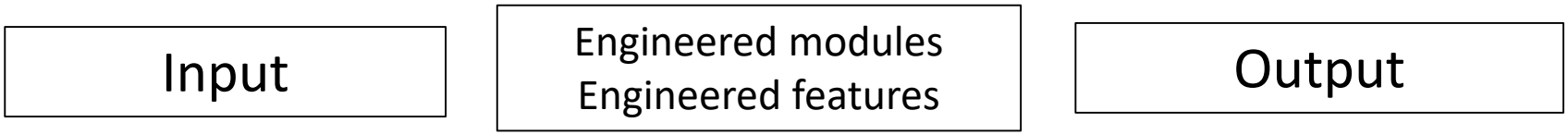
- Part II: Correspondence estimation with FlowNet



- Part III: Cross-dataset generalization

Learning shifts emphasis to the data

Engineering:



Deep Learning:

Training data

The ideal deep learning scenario

1. Training and test data are sampled from the same distribution

   → Both datasets show the same types of variation
   → There is no additional variation in the test set

2. The sampling density is sufficiently dense (big data)

   → The data variation can be picked up during training

This scenario is true for most benchmarks

The common application scenario

1. Training and test data are sampled from different distributions

   → The test set shows new kinds of variation
   → Domain transfer or cross-dataset generalization needed

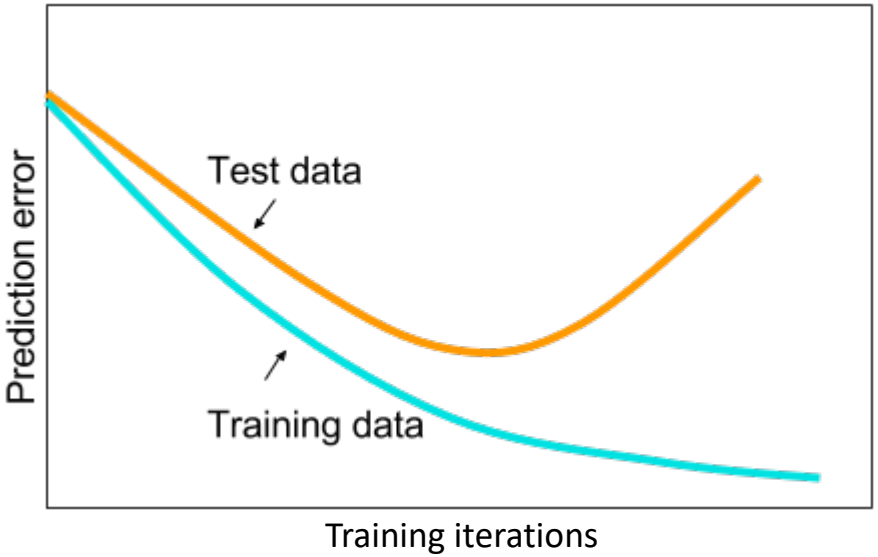
2. Training sets are small

   → Variation of the training distribution is not fully covered
   → Overfitting

Overfitting

Deep networks have millions of parameters and are prone to overfitting

→ Always verify that your network does not overfit

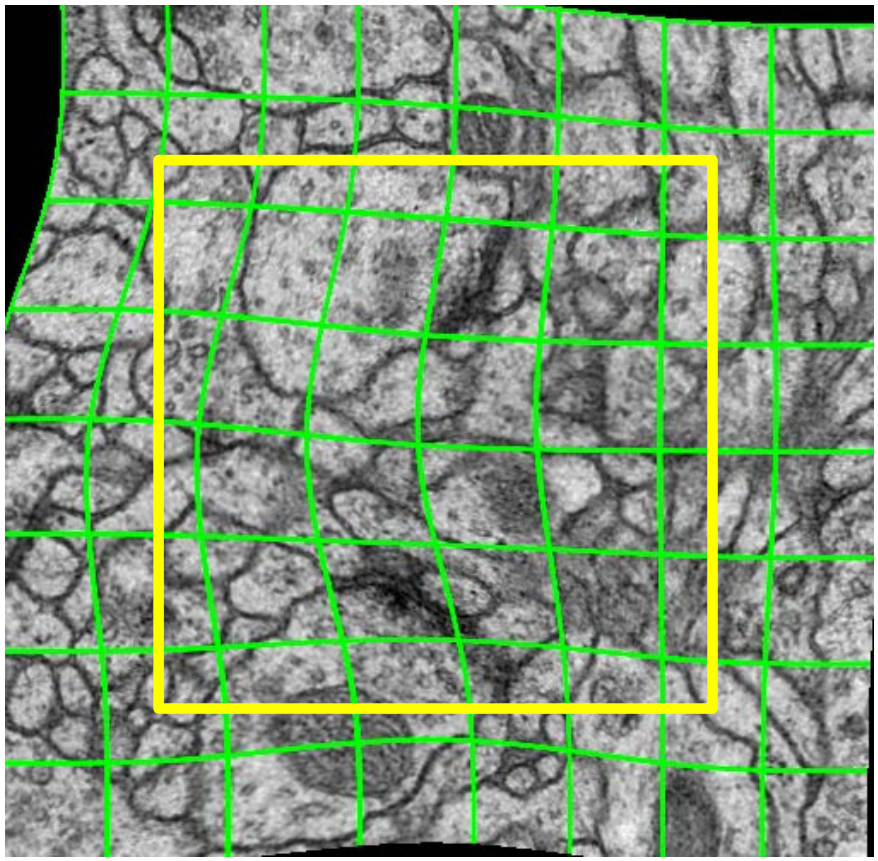


1. Check training and validation curves

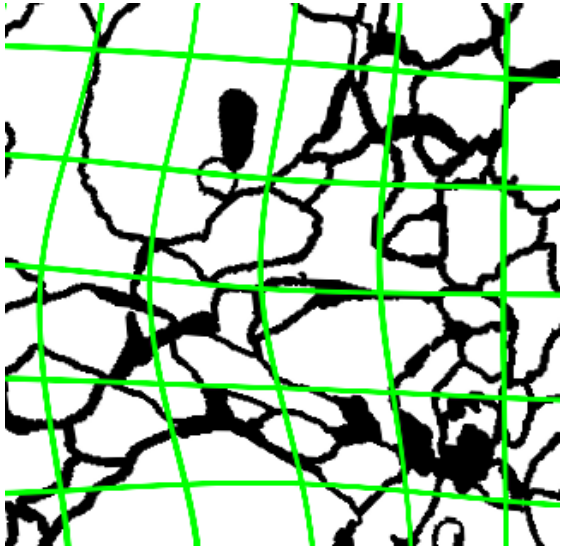
2. Evaluate your network on a different dataset (cross-dataset generalization)

3. Visualize your results and do a qualitative sanity check
   Don't trust the numbers alone

Data augmentation: synthesizing variation



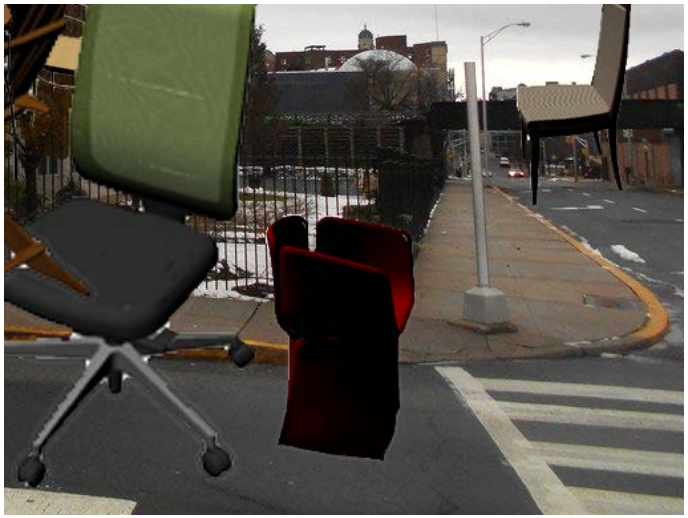
Randomly deformed image

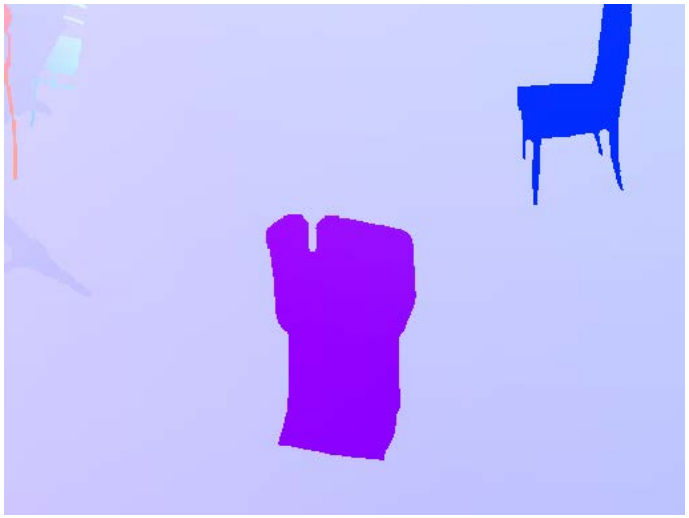


Correspondingly deformed manual labels

Very powerful for many biomedical tasks

Synthetic data for learning a concept





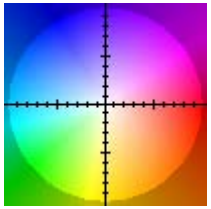
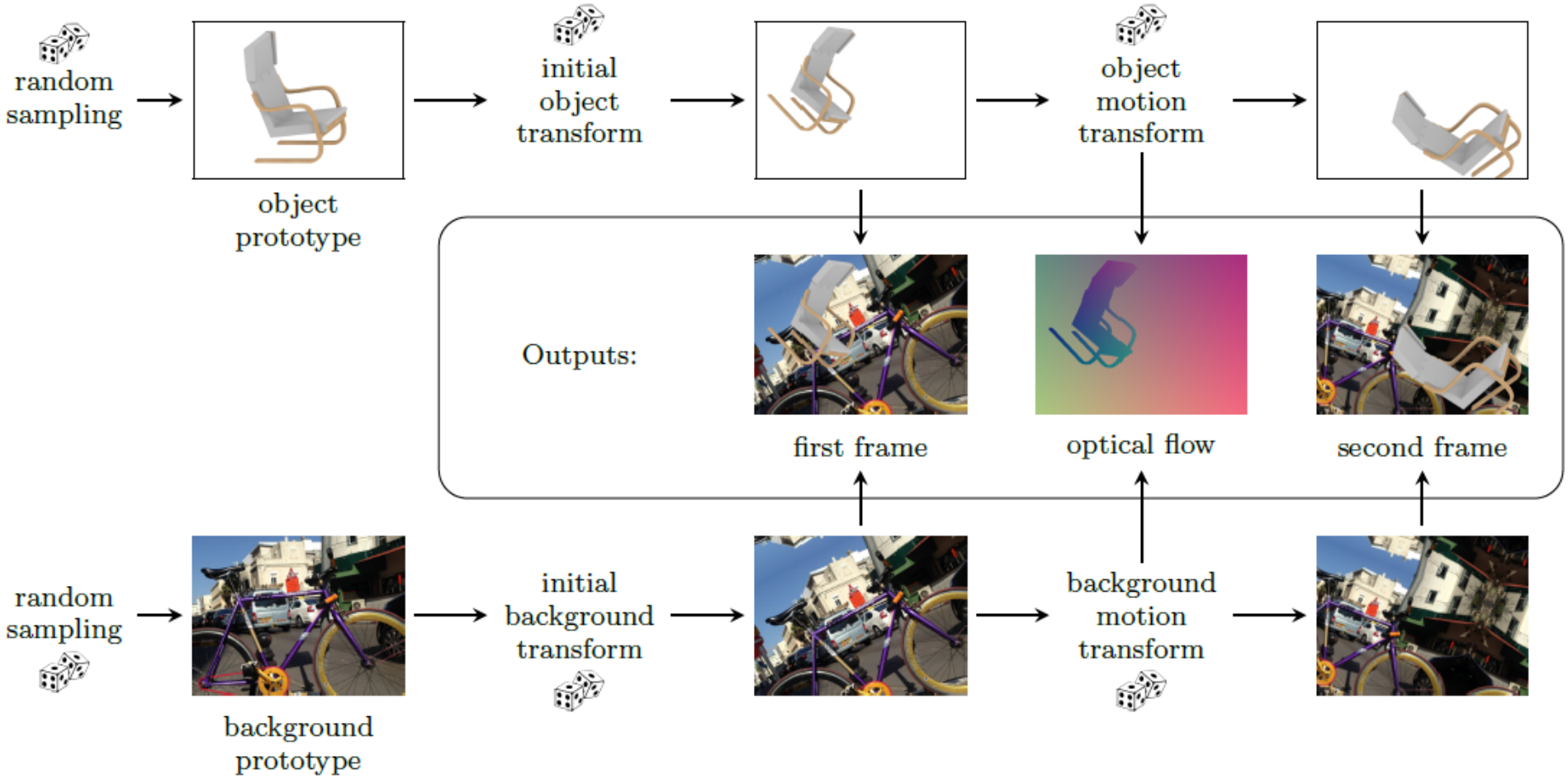


Image pair

Optical flow

Data generation process for Flying Chairs



Synthetic 3D datasets

Mayer et al. 2016

Driving, Monkaa, FlyingThings3D datasets publicly available

Quiz: Which dataset yields best results on Sintel?

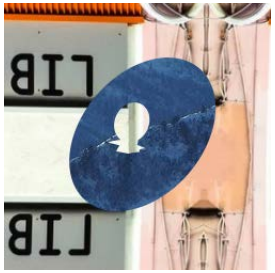
A. Flying Chairs

B. Driving

C. Monkaa

D. FlyingThings3D

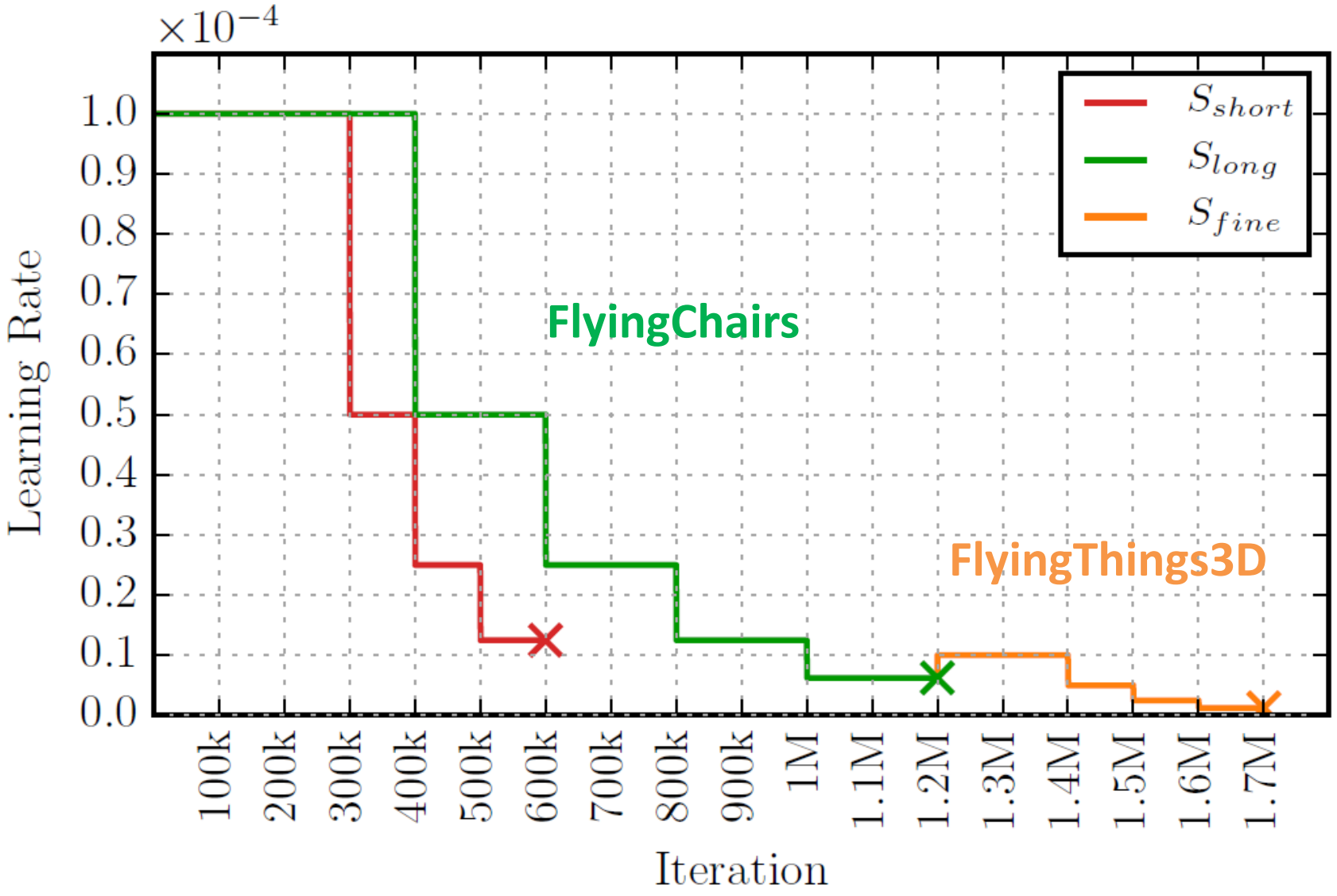
Simple objects are sufficient to learn optical flow



Training data	Polygons	Ellipses	Translation	Rotation	Scaling	Holes in objects	Thin objects	Deformations	Sintel train clean	KITTI 2015 train	FlyingChairs
Boxes	(✓)		✓						5.29	17.69	4.95
Polygons	✓		✓						4.93	17.63	4.60
Ellipses		✓	✓						4.88	17.28	4.87
Polygons+Ellipses	✓	✓	✓						4.86	17.90	4.62
Polygons+Ellipses+Rotations	✓	✓	✓	✓					4.79	18.07	4.38
Polygons+Ellipses+Scaling	✓	✓	✓	✓	✓				4.52	15.48	4.22
Polygons+Ellipses+Holes in objects	✓	✓	✓	✓	✓	✓			4.71	16.36	4.20
Polygons+Ellipses+Thin objects	✓	✓	✓	✓	✓	✓	✓		4.60	16.45	4.13
Polygons+Ellipses+Deformations	✓	✓	✓	✓	✓	✓	✓	✓	**4.50**	**14.97**	4.23
FlyingChairs	(✓)	(✓)	✓	✓	✓	✓	✓		4.67	16.23	(**3.32**)

Mayer et al. 2018

Learning schedule



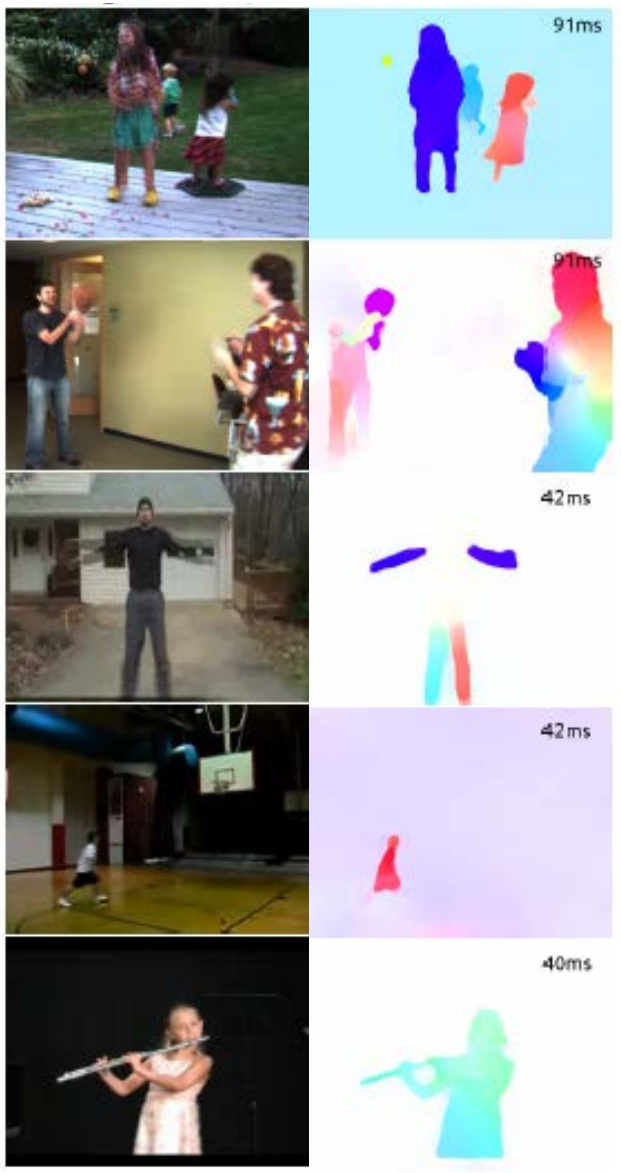
FlyingChairs vs FlyingThings3D

	EPE Sintel train clean
FlyingChairs (1.7M iterations)	4.21
FlyingThings3D (1.7M iterations)	4.50
FlyingChairs+FlyingThings3D (mixed, 1.7M iterations)	4.10
FlyingChairs (1.2M iterations) + FlyingThings3D (500k)	3.79

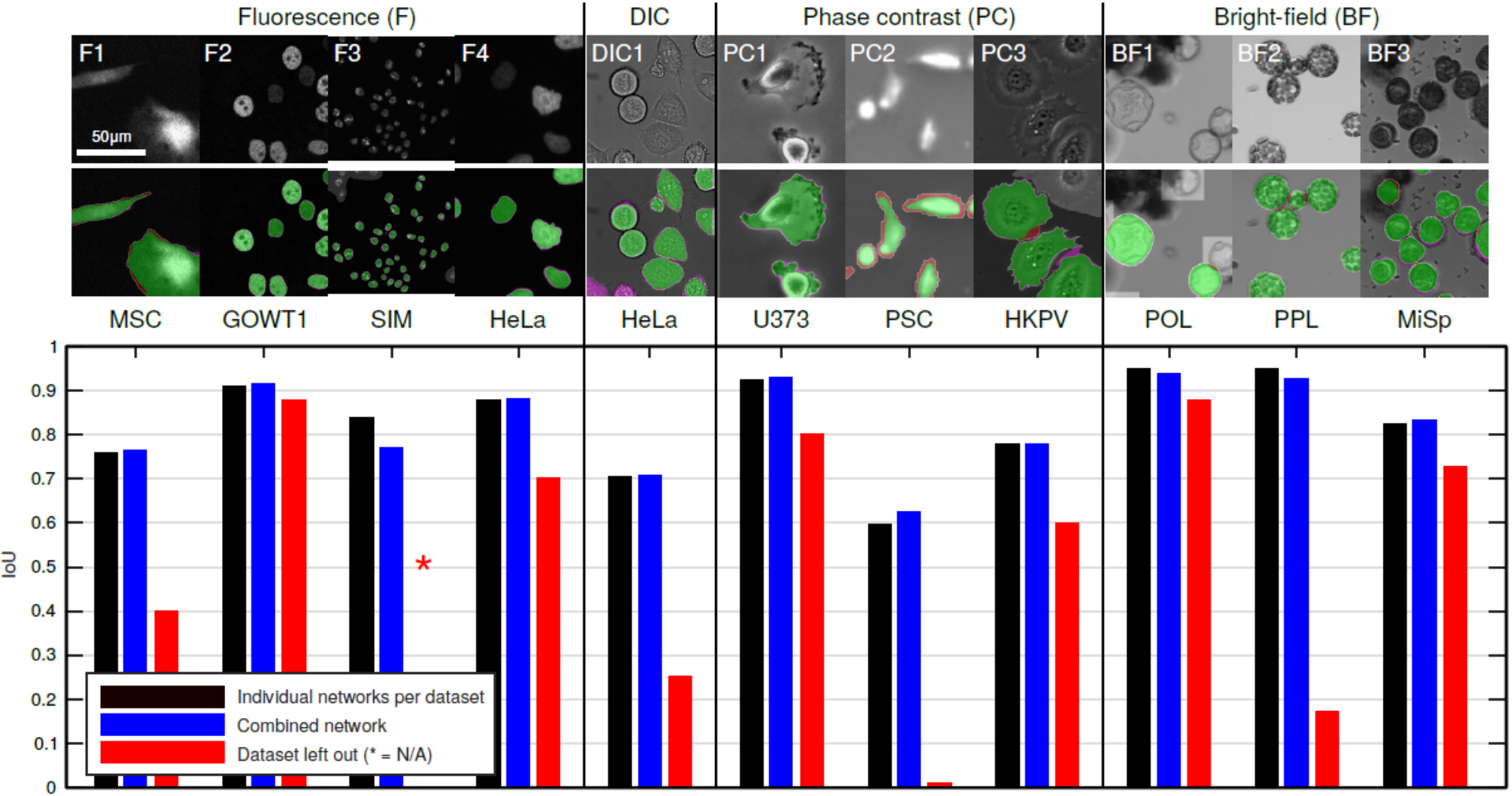
Simple dataset best to learn the basic concept
Complex dataset improves the learned prior

FlowNet generalizes across datasets

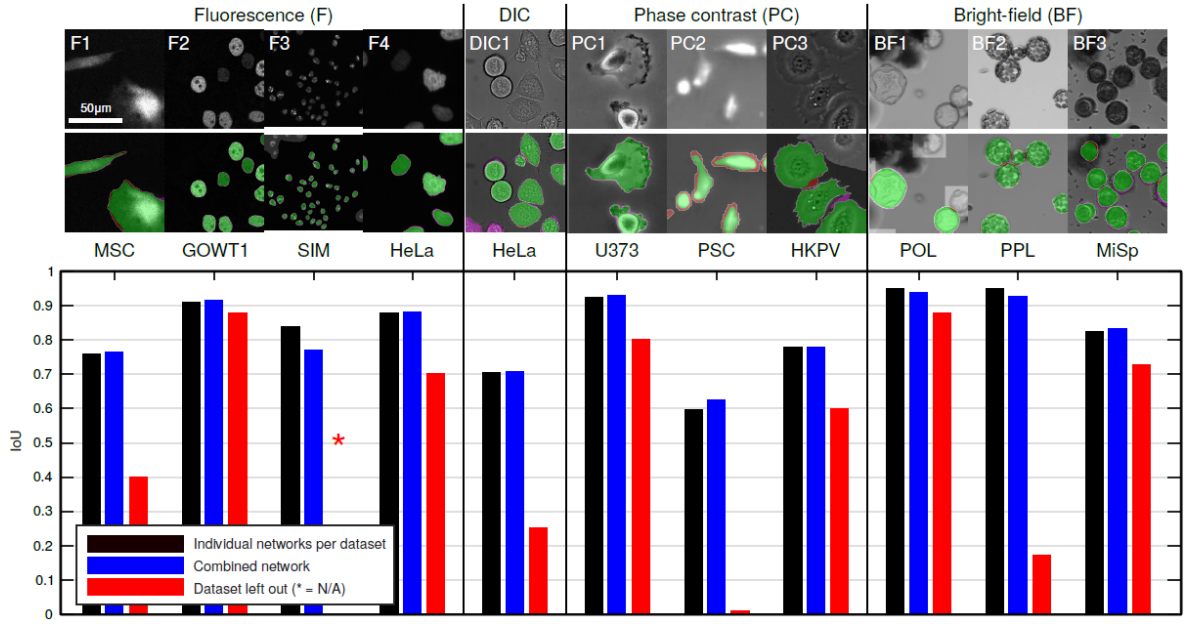
- FlowNet training just on synthetic data is sufficient to generalize to any dataset
- Different from semantic tasks
- Why is this?



U-Net generalization to unseen datasets



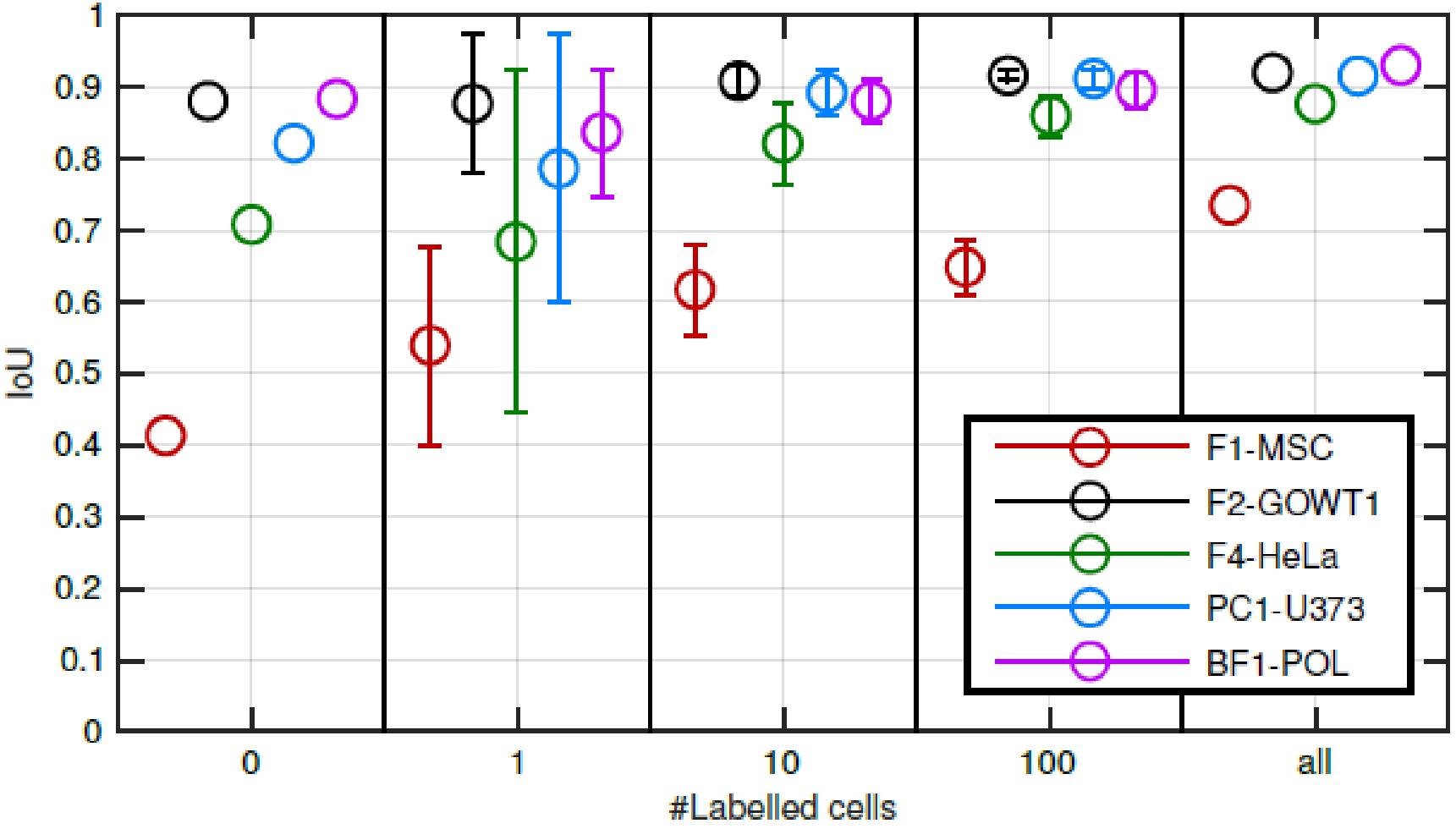
Conclusions



Conclusion 1: One combined network is not worse than many specialized networks

Conclusion 2: Generalization only within the variation bounds of the training data

Fine-tuning on few samples



Exotic data requires more samples fore fine-tuning

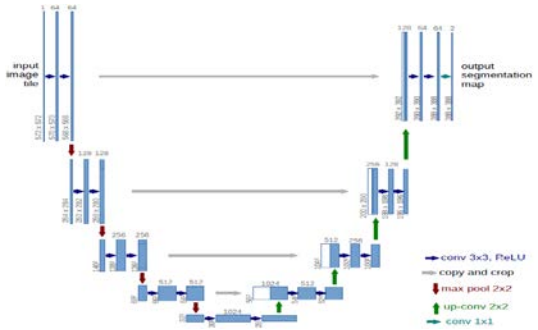
Take-home Part III

1. Make sure your training data covers all variation of your application scenario

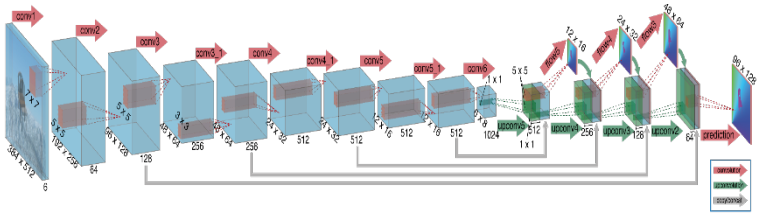
2. Check for over-fitting and prevent it by collecting or synthesizing enough data
   Ideally learn the underlying concept

3. Be careful with benchmark results if the benchmark is not representative for the application scenario

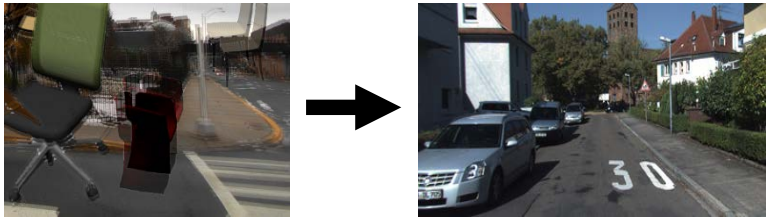
Thank you



- Part I: Encoder-decoder networks



- Part II: Correspondence estimation with FlowNet



- Part III: Cross-dataset generalization