Estimating Optic Flow with Regression

D.A. Forsyth,

University of Illinois at Urbana Champaign

Change what the decoder does…

- Regression – predict “image like thing” from image
- An “image like” thing
  - may have the same resolution as the image
  - continuous (not categorical)
  - lots of examples
  - can be predicted from an image (but how do we know?)
- Examples:
  - depth
  - normal
  - defogged image
  - superresolution
  - lots of others..

Why not predict flow?

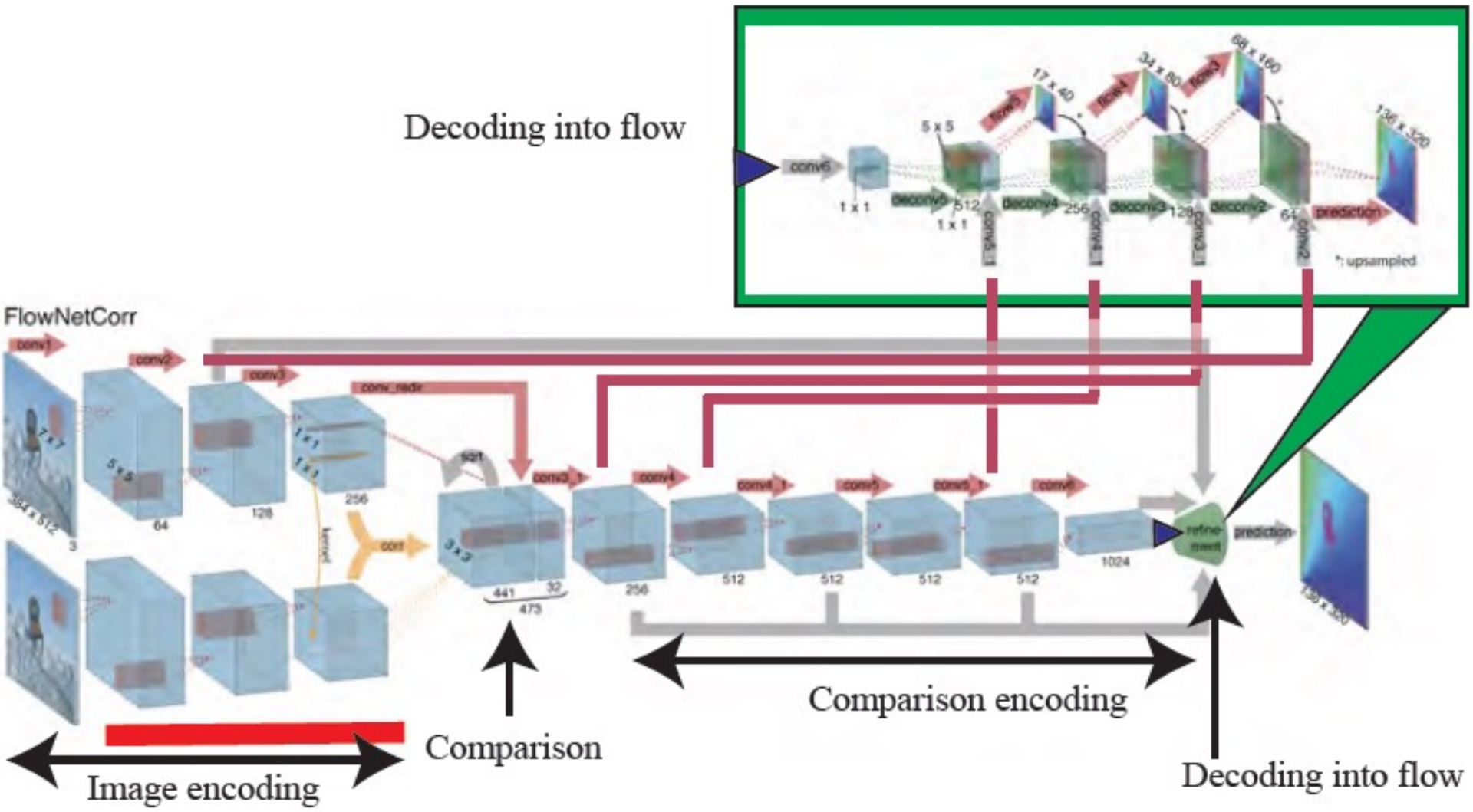
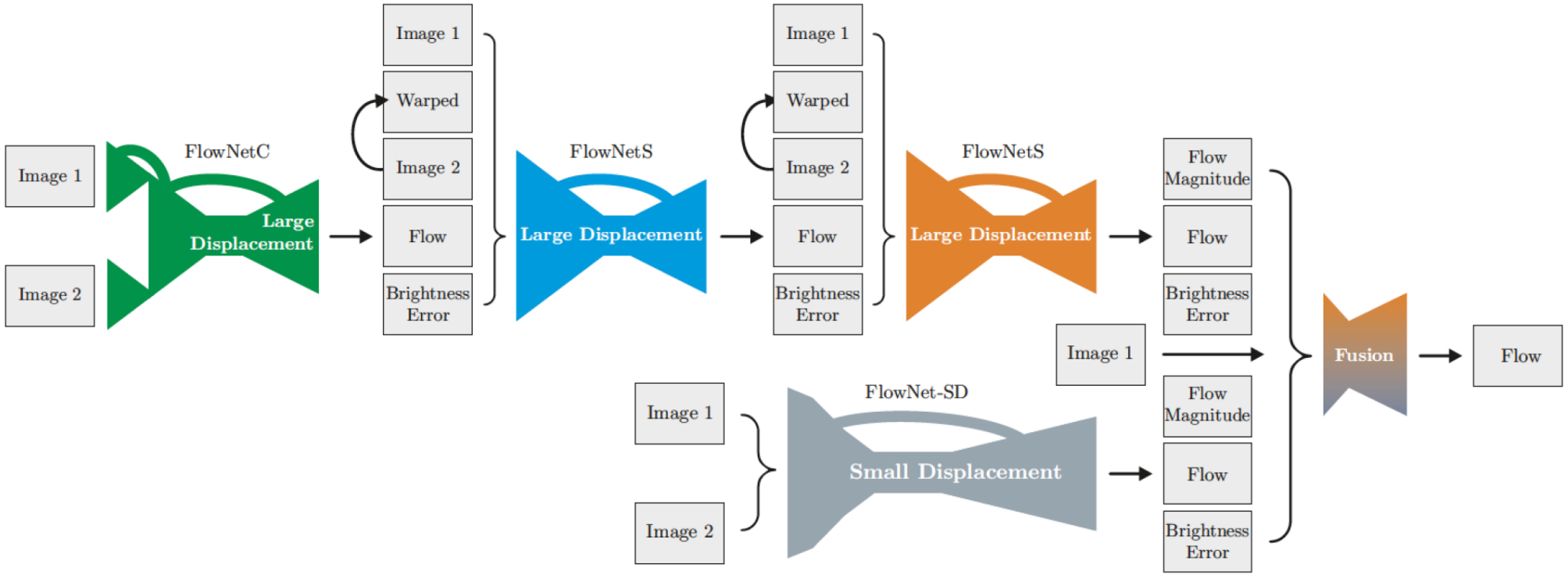


FIGURE 33.13: *The overall architecture of Flownet, a regression based optic flow predictor ([]). The frames pass through an image encoder; the results are compared (details in text), and the comparison further encoded. Finally, the result is decoded into a flow field, using skip connections from the comparison encoder and from the first frame encoder.*

More flownets => better results

- Notice analogy with incremental flow estimation
  - also, multi scale



More flownets => better results

Warp image 2 backward along flow;
should look like 1

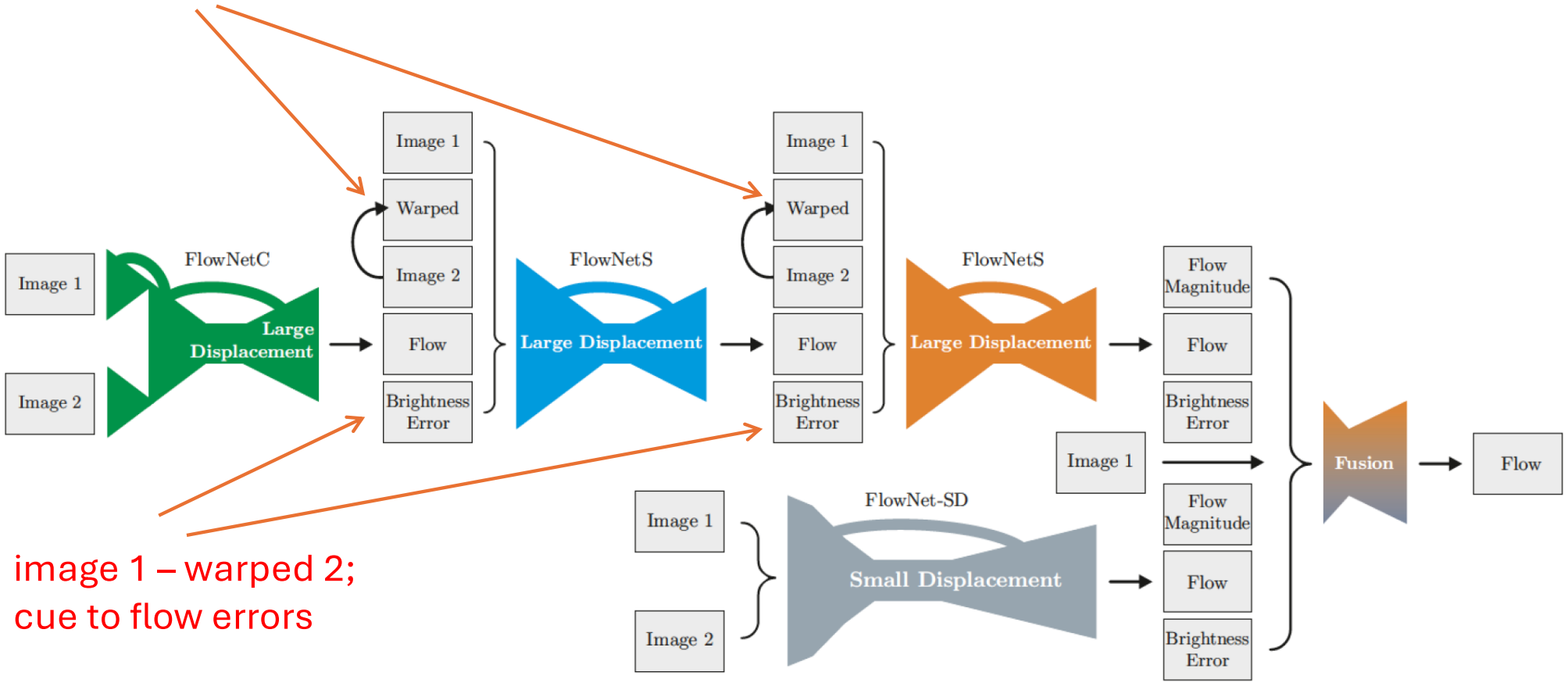
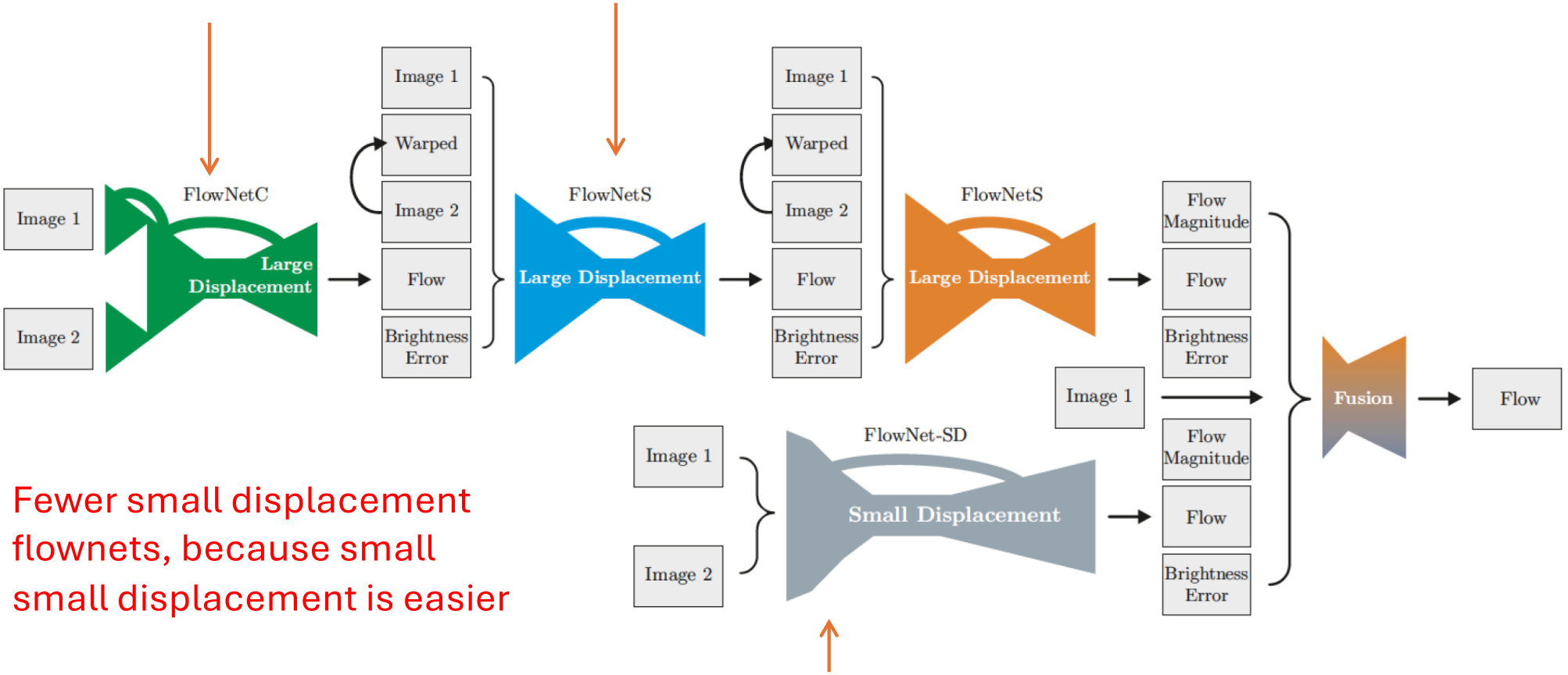


image 1 – warped 2;
cue to flow errors

More flownets => better results

Large displacement has larger stride;
better at large flows

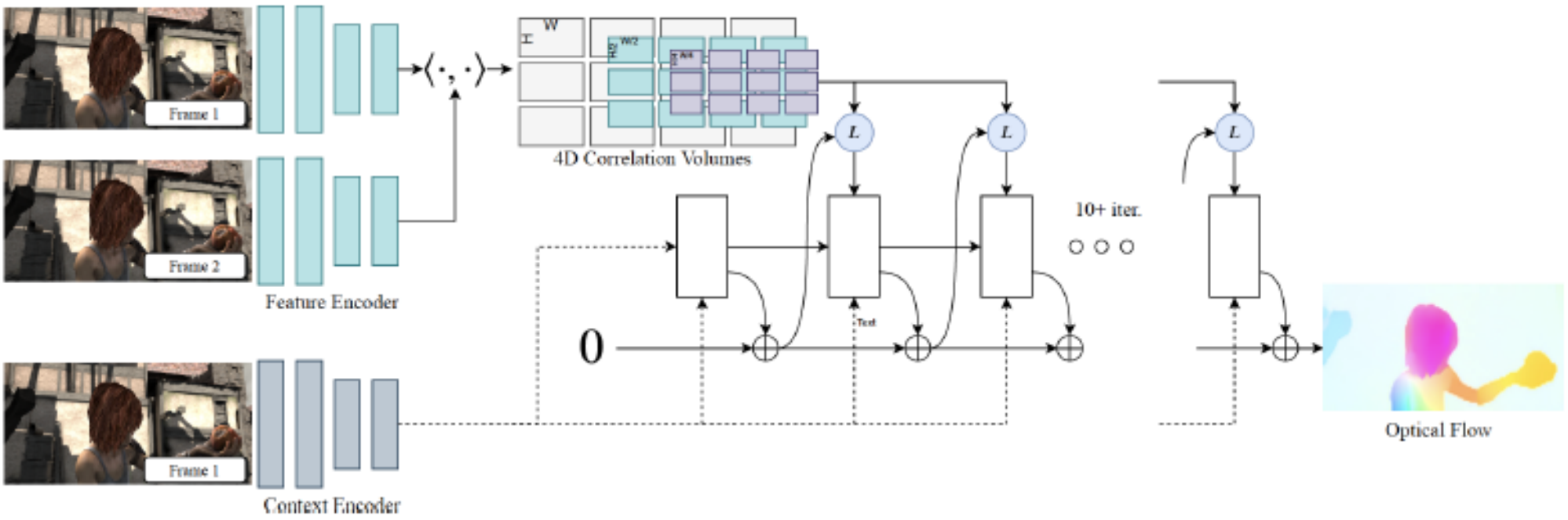


Fewer small displacement flownets, because small small displacement is easier

Small displacement has smaller stride;
better at small flows

RAFT

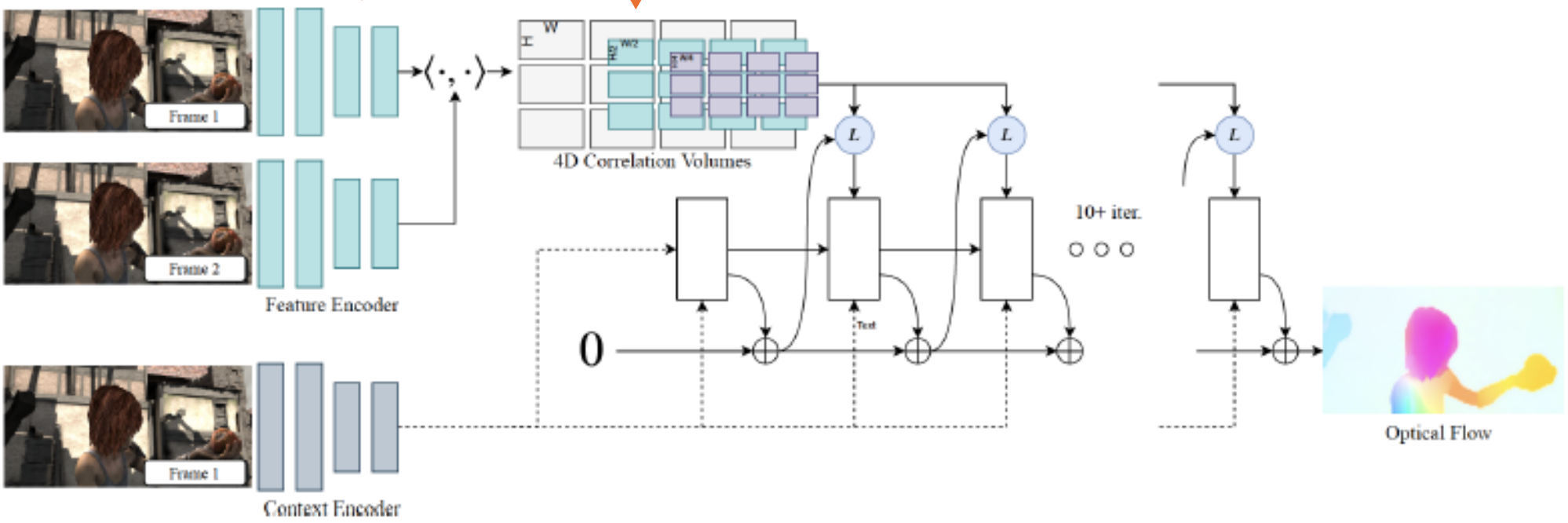
- Development
  - bigger GPUs means you can do more comparison
  - RAFT is the natural generalization



RAFT

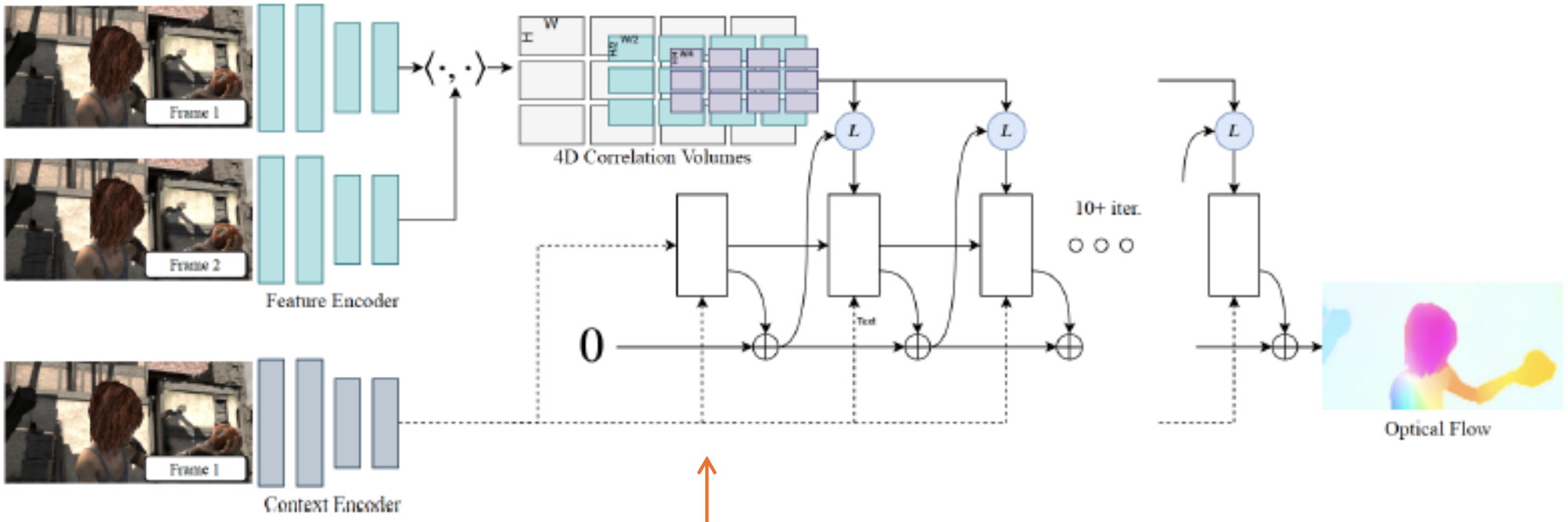
If images are 3xWxH, then this is WxHxWxH
- compare features at every location in 1 with every location in 2

Each image is encoded with the same encoder



RAFT

Essentially, a multi-scale pyramid
For each location in image 1 (at original resolution) how good is match in image 2 (at each of multiple resolutions). This is a representation of flow at multiple precisions



Repeatedly re-estimate flow using:
current estimate, info from correlations,
context from context encoder, revised context

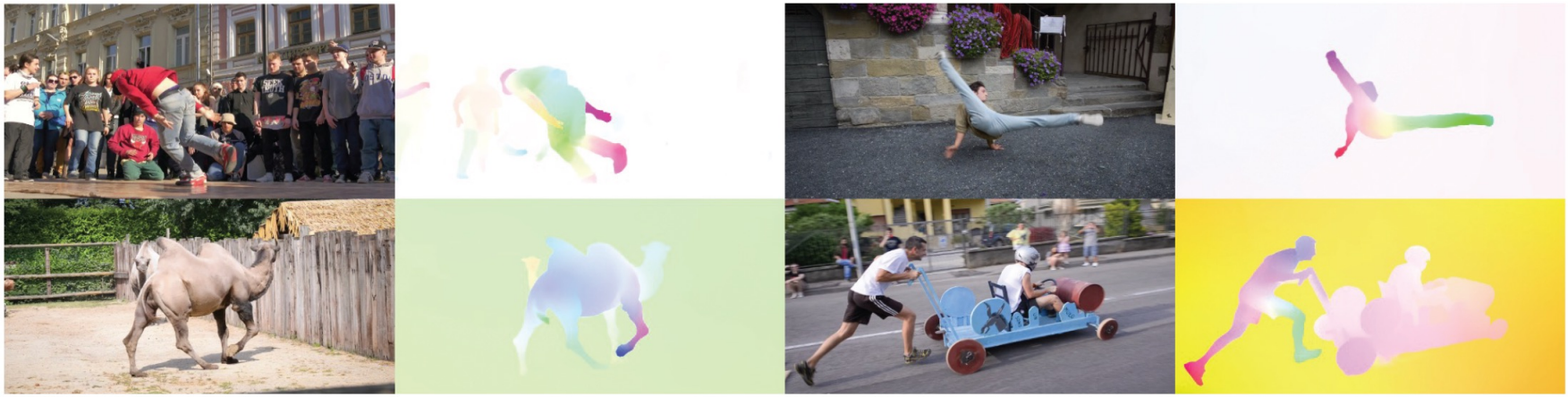
RAFT Flows

SINTEL is a rendered dataset from CGI movie, so ground truth is known (most unusual)

RAFT Flows

Kitti

Davis



KITTI – autonomous car dataset
Davis – 1088x1920 resolution

Significant datasets available

- Learning is usually on rendered/synthetic data

- Notice there are opportunities for self-learning
  - Just minimize photometric loss
  - But…