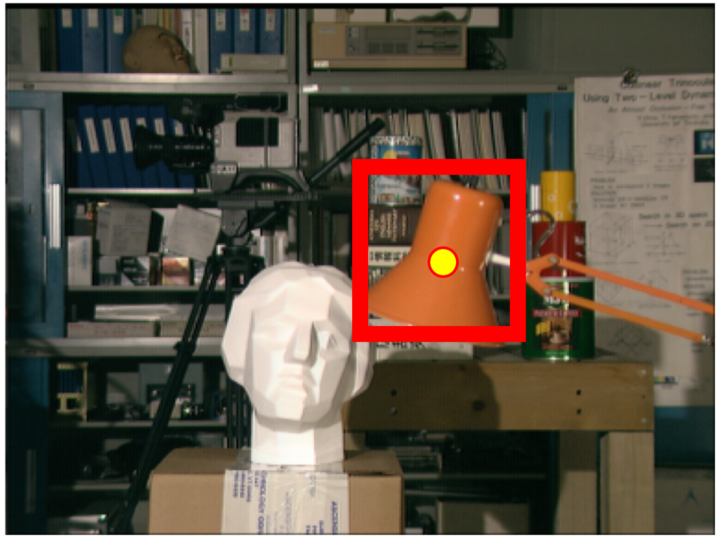
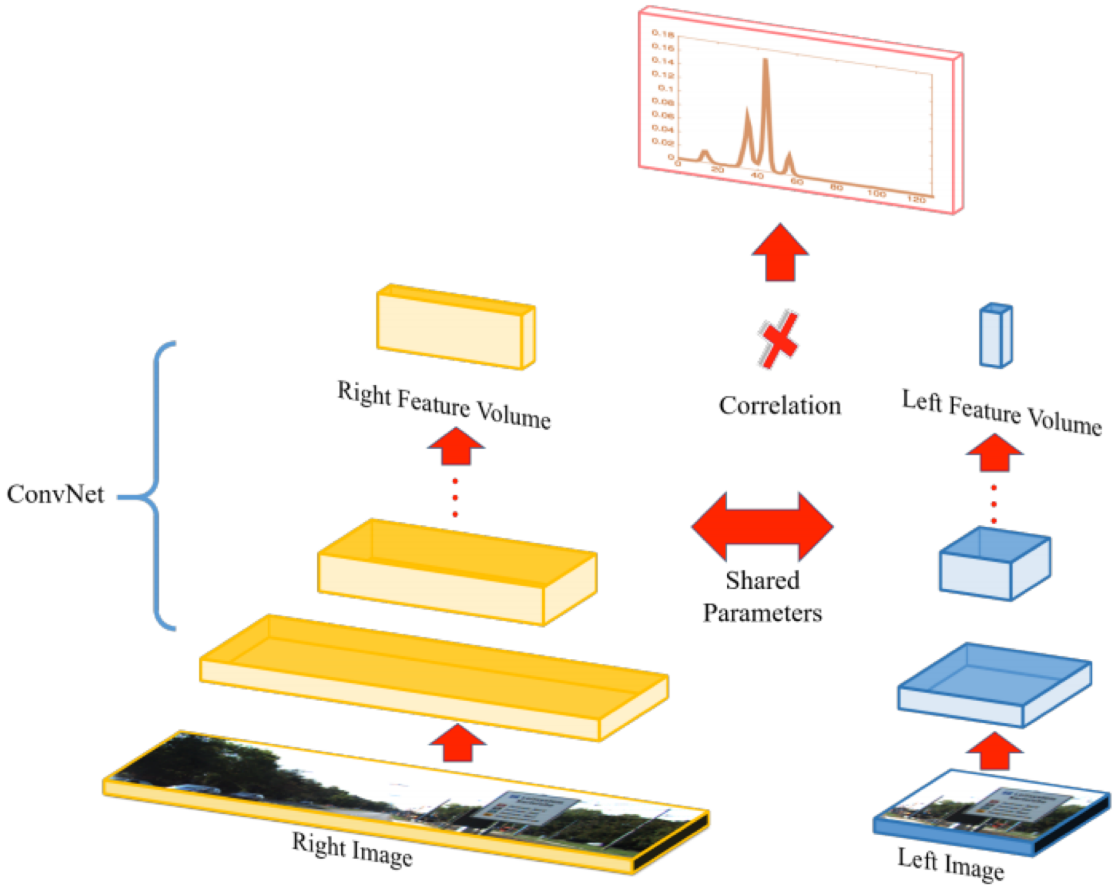
The correspondence problem

Learning patch similarity for disparity estimation



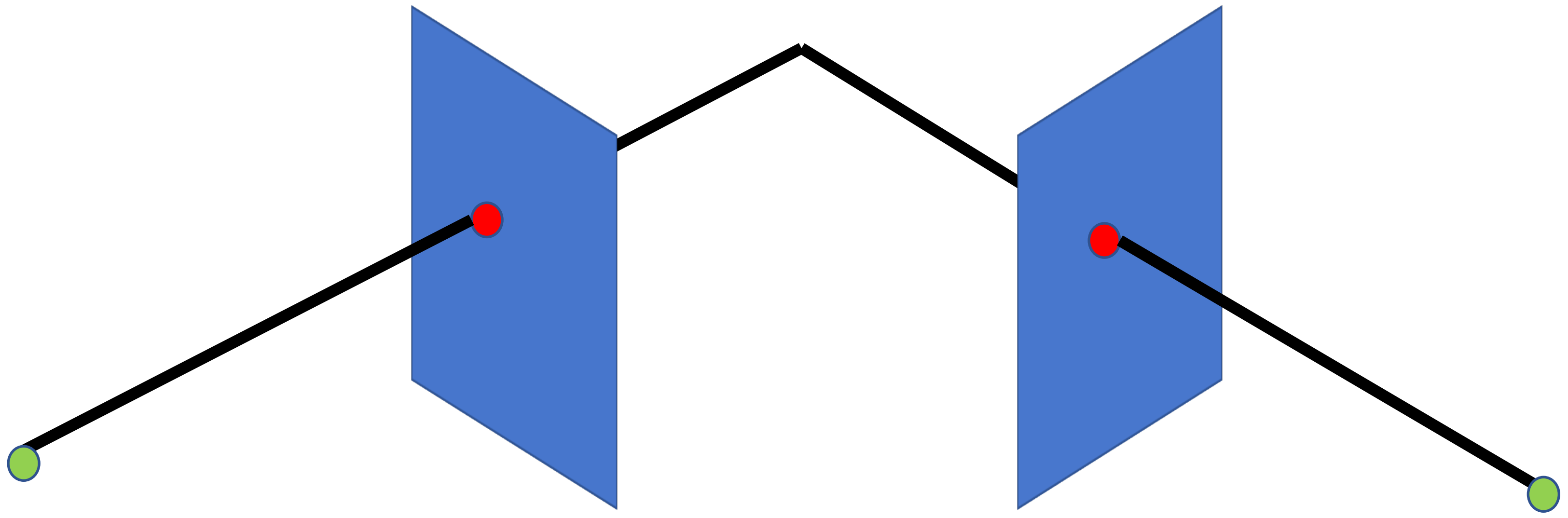
Learning patch similarity for disparity estimation



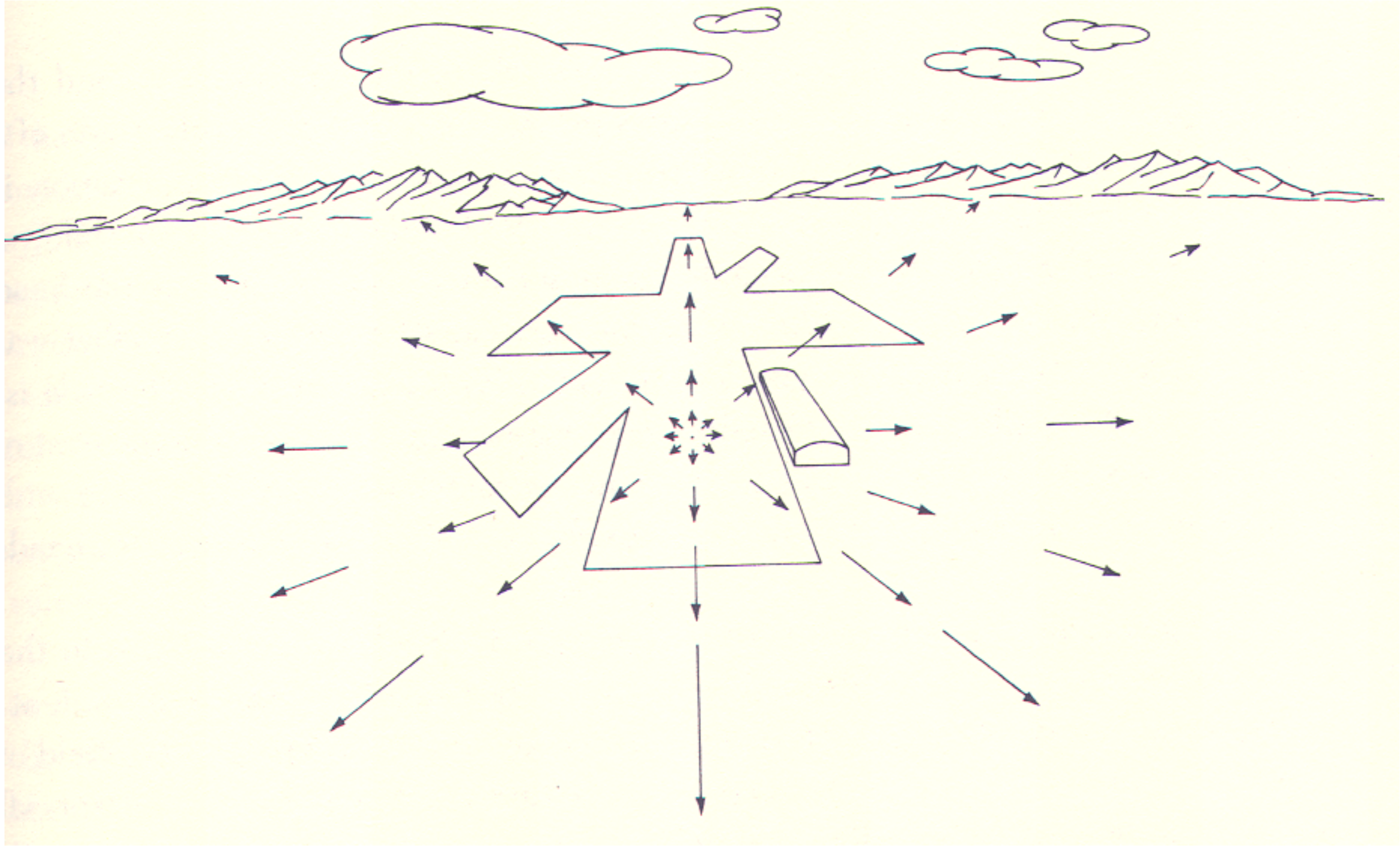
Efficient Deep Learning for Stereo Matching. Wenjie Luo, Alexander G. Schwing, Raquel Urtasun. In *CVPR,* 2016

Throwback: SIFT

Stereo

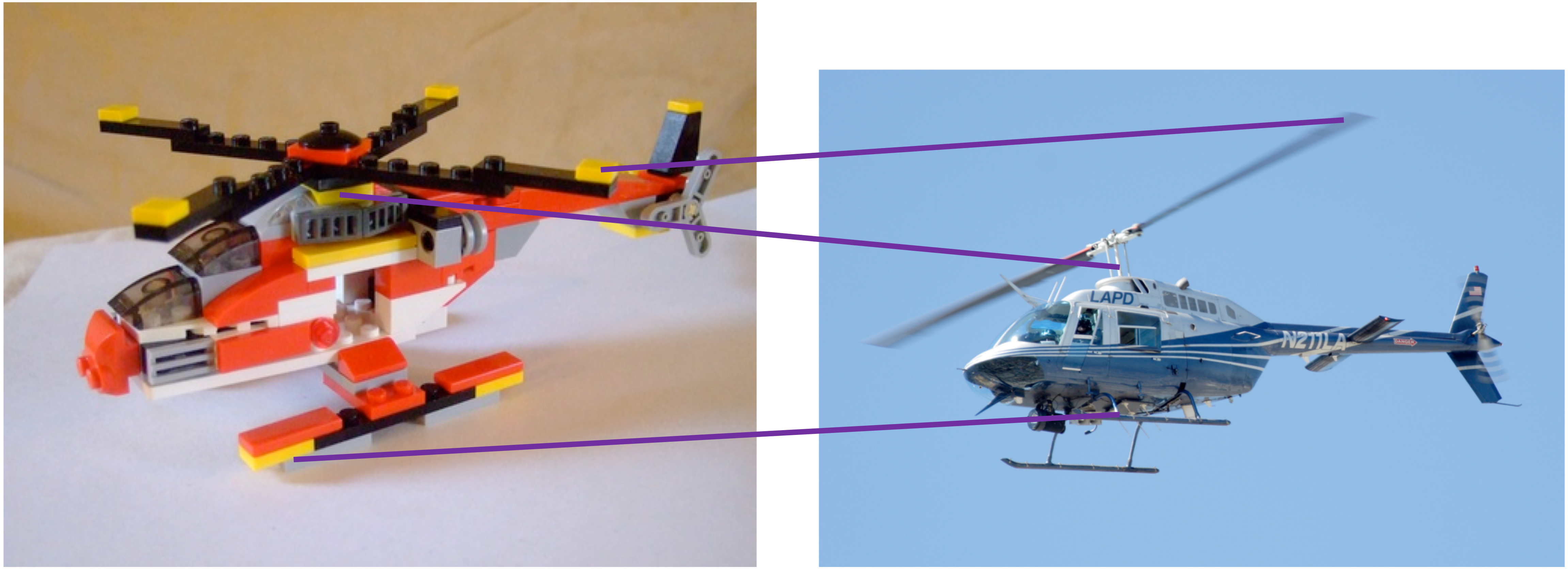


Optical flow

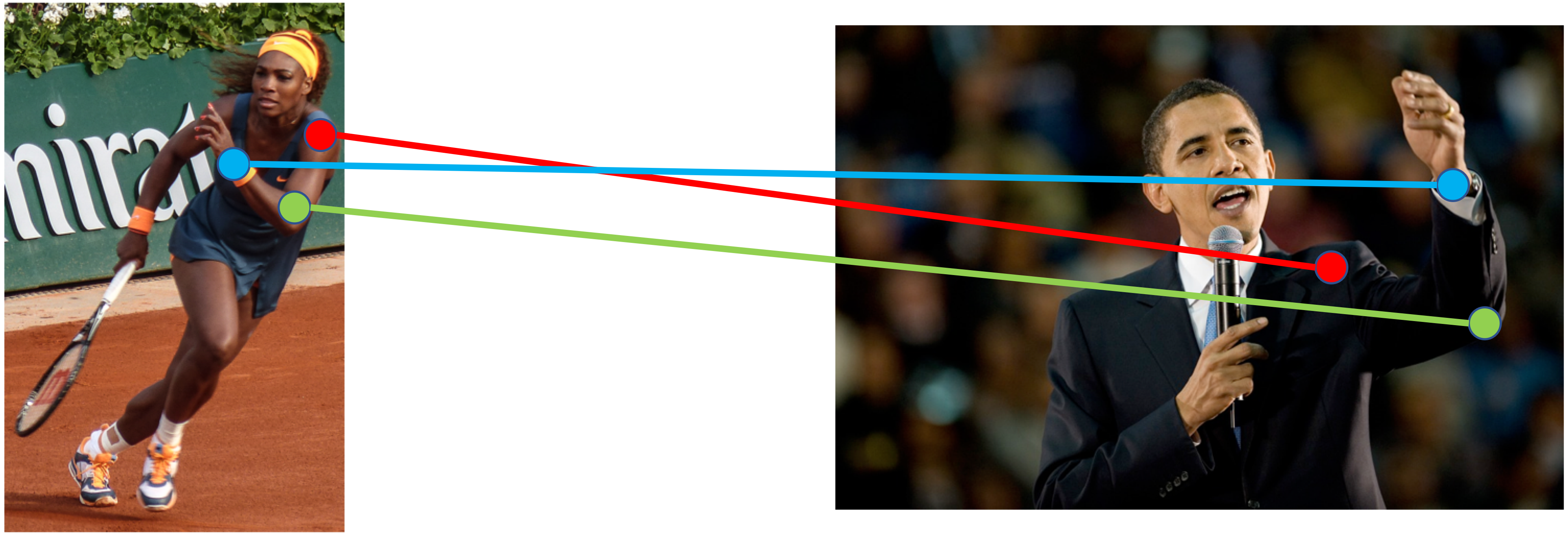


$$I(x+dx, y + dy, t+dt) = I(x,y,t)$$

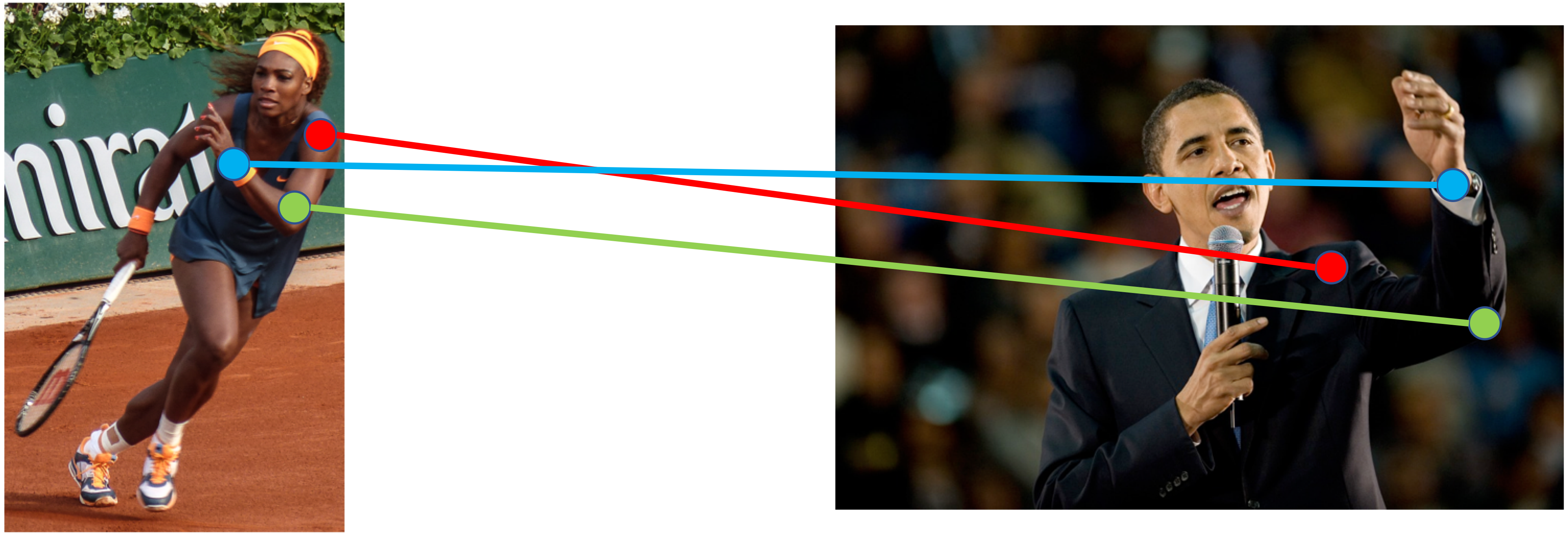
Cross-instance



Cross-instance



Cross-instance = Pose estimation



Correspondence as a general problem

- Sparse vs dense
- Same instance vs different instance

Correspondence as a general problem

- Sparse vs dense
- Same instance vs different instance
- Nearby cameras (small baseline) vs far away cameras (large baseline)

Correspondence as a general problem

- Sparse vs dense
- Same instance vs different instance
- Nearby cameras (small baseline) vs far away cameras (large baseline)
- Rigid scene/objects vs moving scene/objects

Correspondence as a general problem

- Sparse vs dense
- Same instance vs different instance
- Nearby cameras (small baseline) vs far away cameras (large baseline)
- Rigid scene/objects vs moving scene/objects
- Category-specific vs category-agnostic

Disparity estimation/ Depth estimation

- Sparse vs dense
- Same instance vs different instance
- Nearby cameras (small baseline) vs far away cameras (large baseline)
- Rigid scene/objects vs moving scene/objects
- Category-specific vs category-agnostic

Optical flow

- Sparse vs dense
- Same instance vs different instance
- Nearby cameras (small baseline) vs far away cameras (large baseline)
- Rigid scene/objects vs moving scene/objects
- Category-specific vs category-agnostic

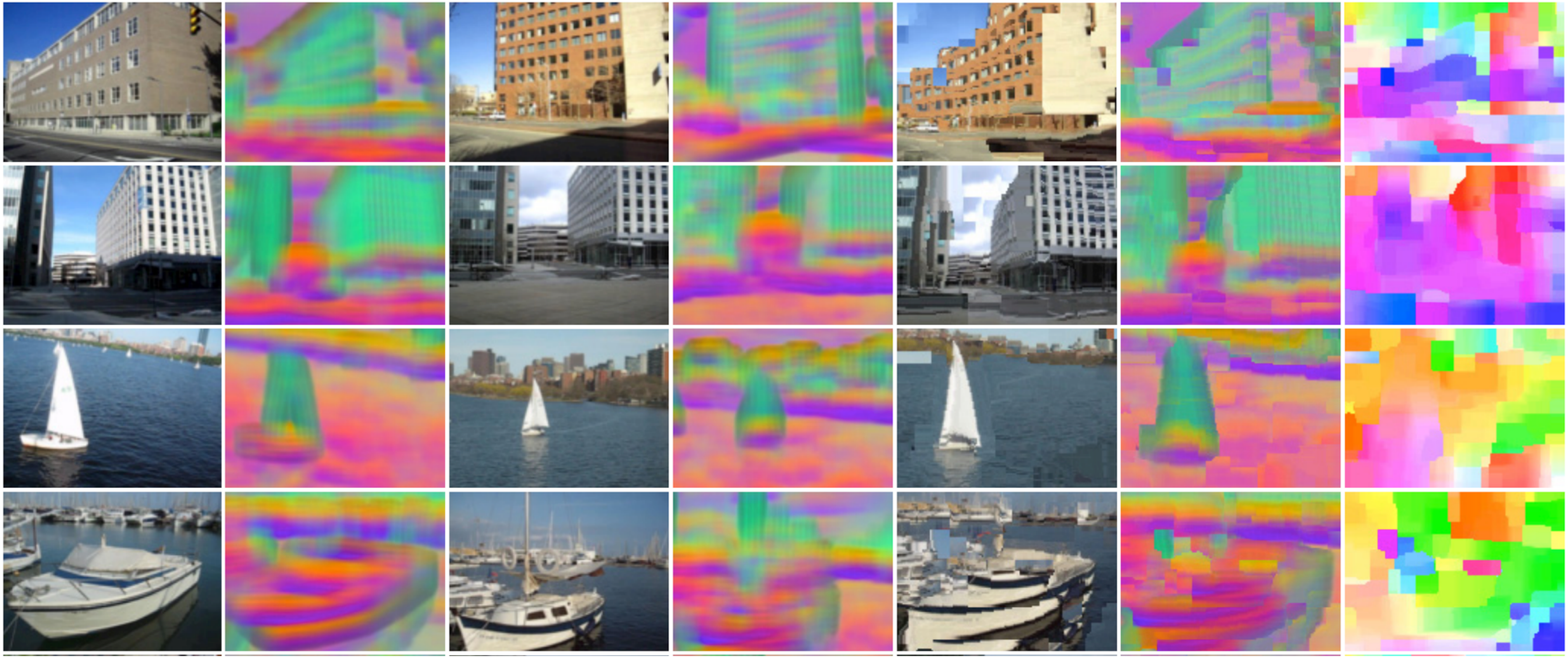
Sparse reconstruction from stereo / Estimating camera pose

- Sparse vs dense
- Same instance vs different instance
- Nearby cameras (small baseline) vs far away cameras (large baseline)
- Rigid scene/objects vs moving scene/objects
- Category-specific vs category-agnostic

Keypoint detection

- Sparse vs dense
- Same instance vs different instance
- Nearby cameras (small baseline) vs far away cameras (large baseline)
- Rigid scene/objects vs moving scene/objects
- Category-specific vs category-agnostic

SIFT Flow



Liu, Ce, Jenny Yuen, and Antonio Torralba. "Sift flow: Dense correspondence across scenes and its applications." *IEEE transactions on pattern analysis and machine intelligence* 33.5 (2011): 978-994.

Learning correspondence

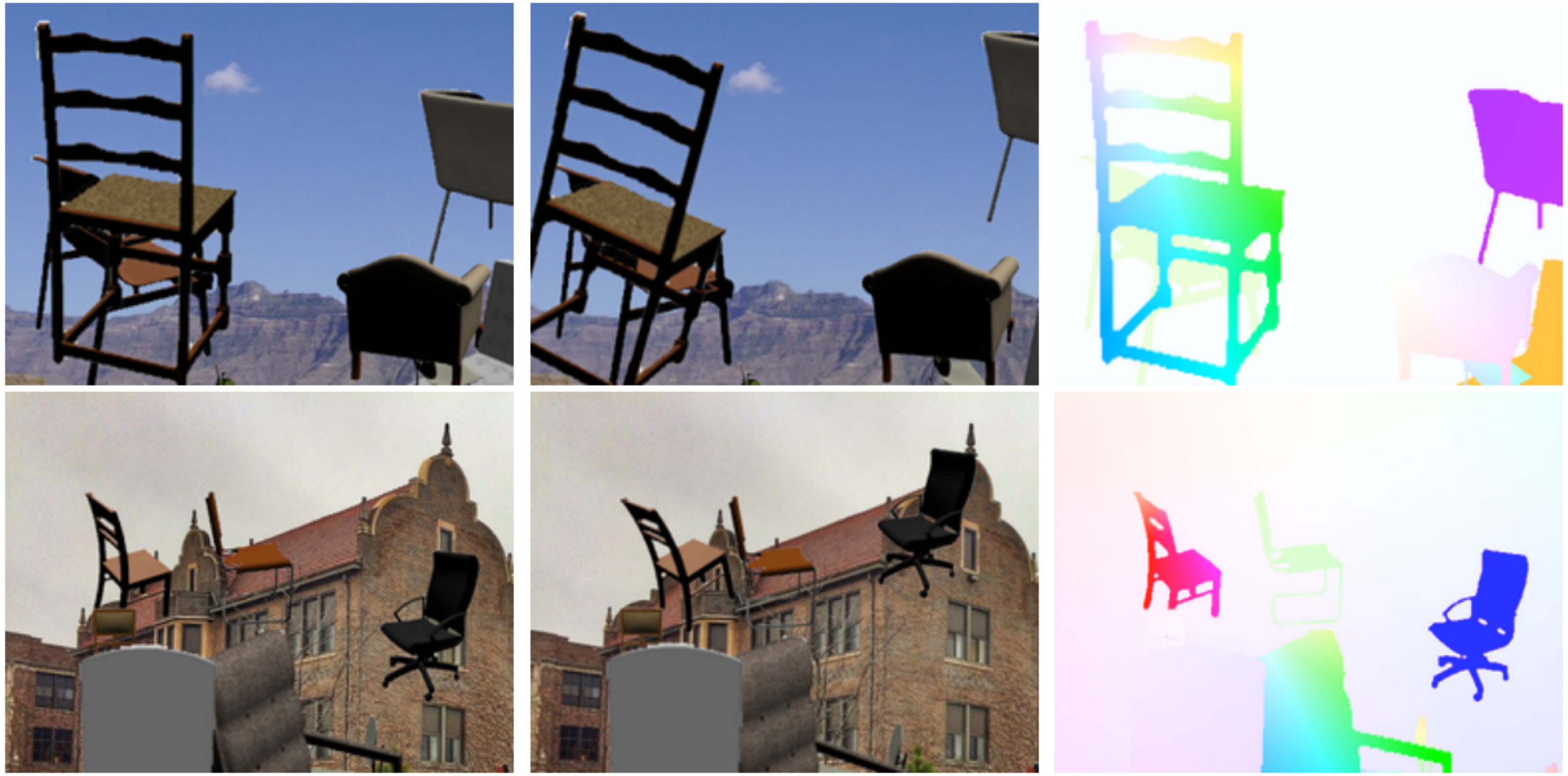
- Two main questions:
- Training data?
  - Need pairs of corresponding points for supervised
- Model architecture?
  - Takes two images and outputs a correspondence map
  - Just compute patch similarity and do post processing, or…
  - Directly compute finished product?

Simulated data for optical flow

Sintel
1628 frames

A Naturalistic Open Source Movie for Optical Flow Evaluation. Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. In *ECCV,* 2012.

Simulated data for optical flow



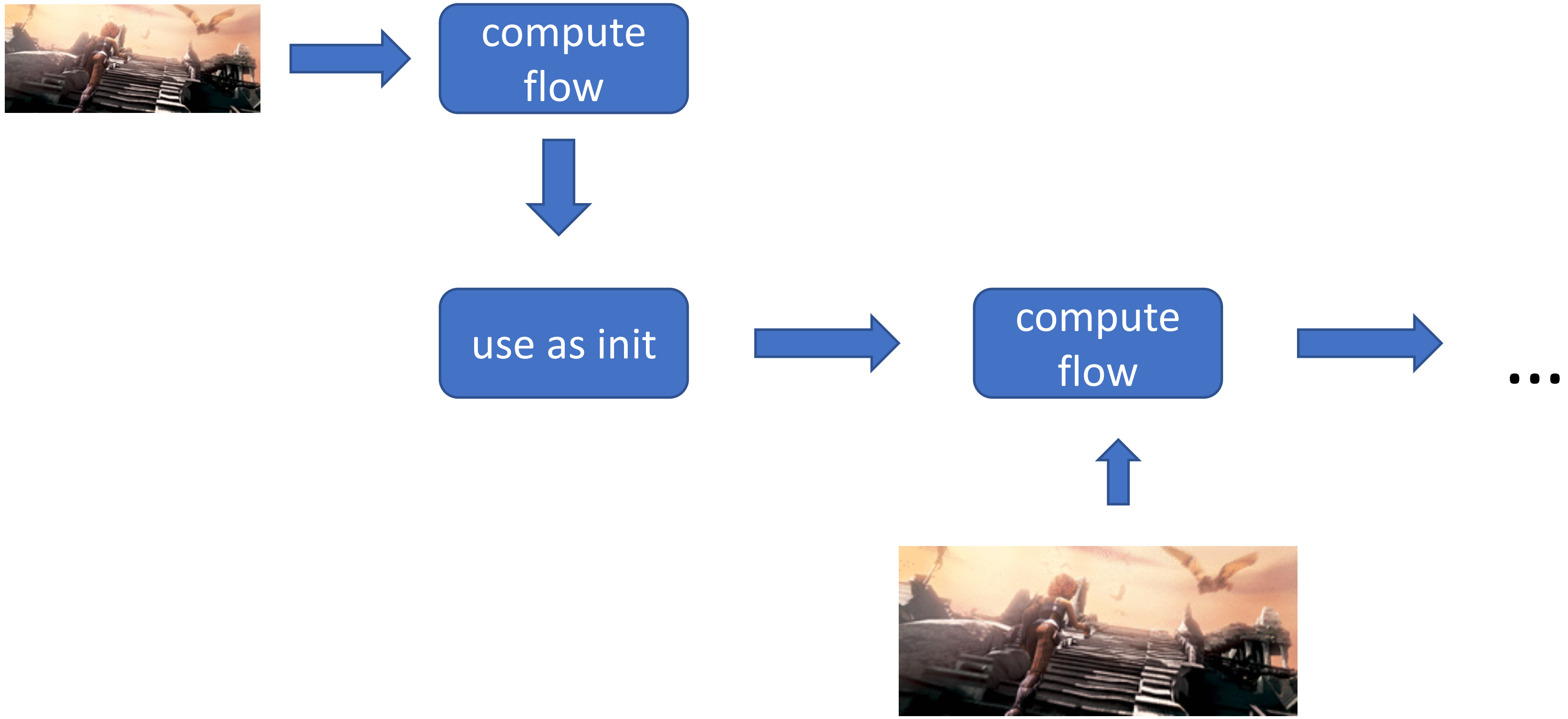
FlowNet: Learning Optical Flow with Convolutional Networks. Philipp Fischer, Alexey Dostovitskiy, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, Thomas Brox. In *ICCV* 2015.

Optical flow with large displacements

- Optical flow constraint equation assumes differential optical flow
- “Large displacement”?
- Key idea: reducing resolution reduces displacement
- Reduce resolution, then upsample?
  - will lose fine details

Optical flow with large displacements

- Key idea 2: Use upsampled flow as *initialization*
- *Changes to initialization will be infinitesimal*

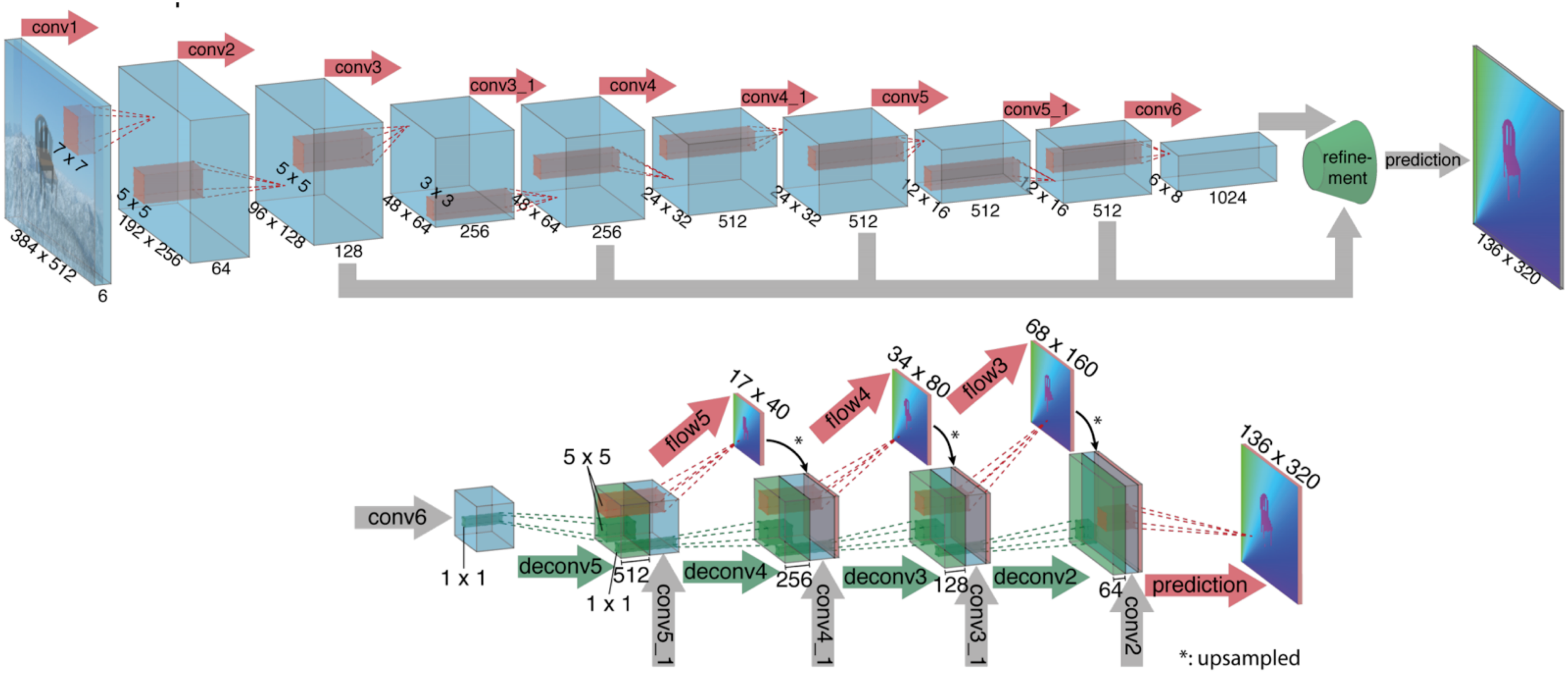


Brox, Thomas, et al. "High accuracy optical flow estimation based on a theory for warping." *Computer Vision-ECCV 2004* (2004)

Coarse-to-fine processing

- A specific instance of a general idea
- Coarse scales:
  - Global / large structures
  - Long-range relationships
  - But: imprecise localization
- Fine scales:
  - Precise localization
  - But: aperture problem
- Idea: start from coarse scales, add fine scale detail

Learning convnets for optical flow



FlowNet: Learning Optical Flow with Convolutional Networks. Philipp Fischer, Alexey Dostovitskiy, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, Thomas Brox. In *ICCV* 2015.

Learning convnets for correspondence

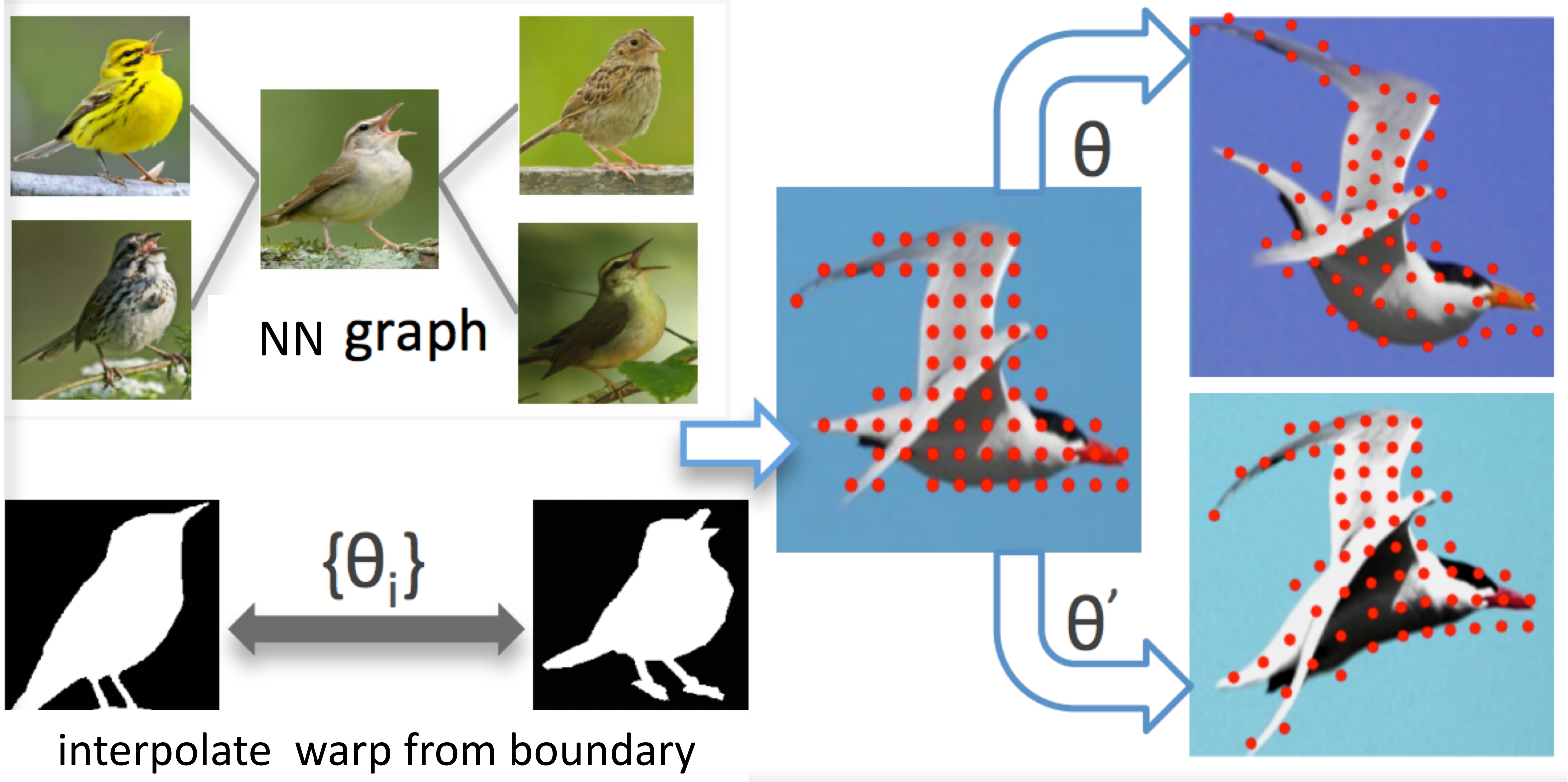
- Luo et al (Disparity)
- Real data from KITTI
- Predict matching costs and post-process

- Fischer et al (Optical flow)
- Simulated synthetic data
- Directly predict smooth correspondence

Generalizing across instances

- Can we learn from deformations of single instances and correspond *across* instances / categories?

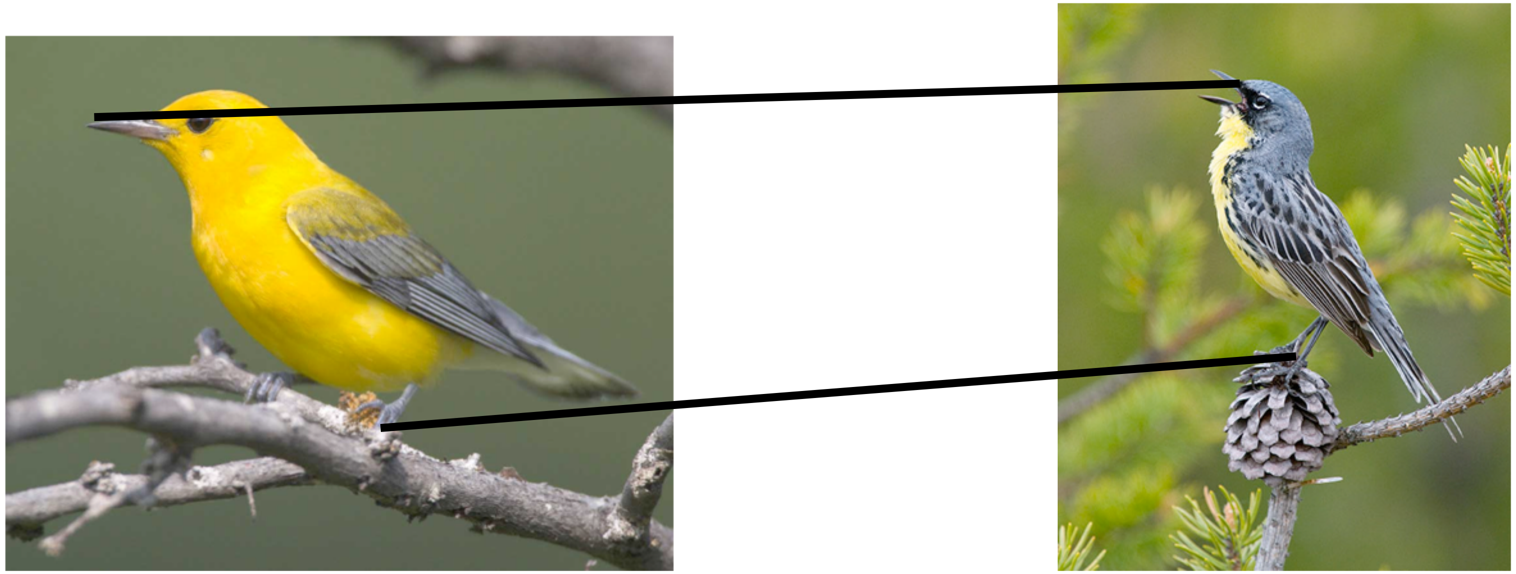
Simulating non-rigid objects



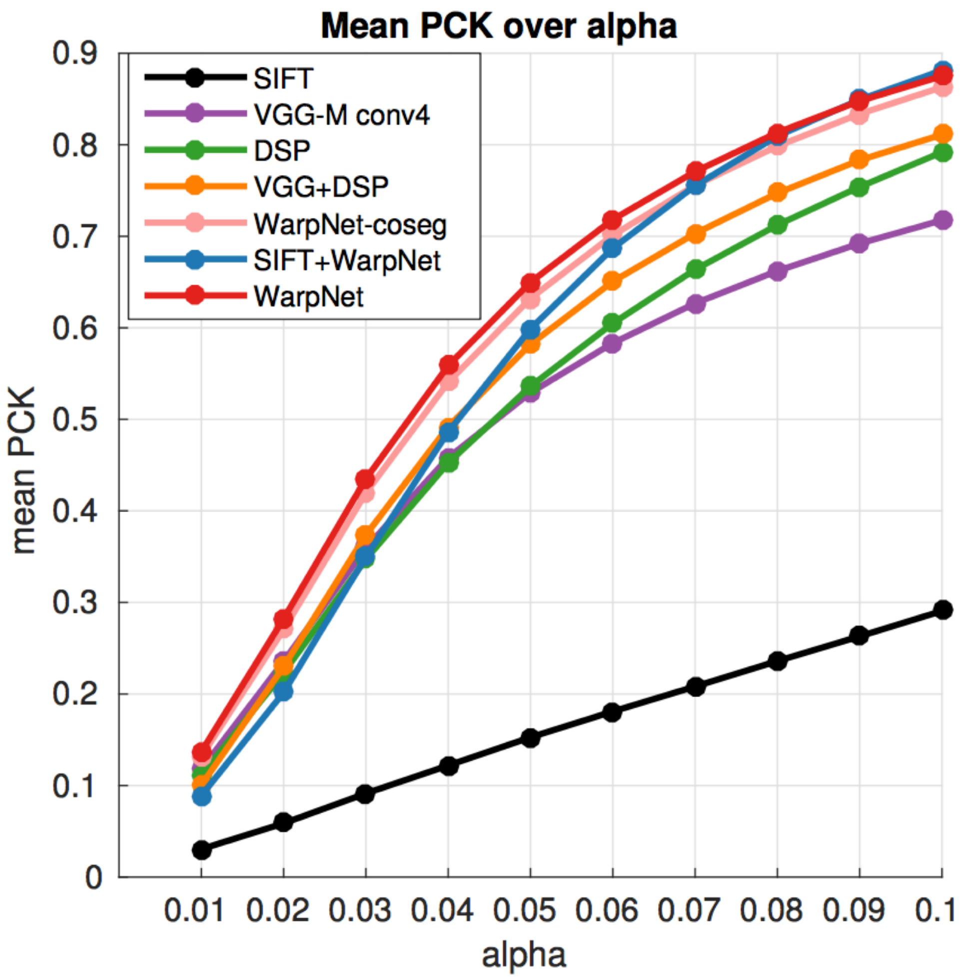
WarpNet: Weakly Supervised Matching for Single-view Reconstruction. Angjoo Kanazawa, David W. Jacobs, Manmohan Chandraker. In *CVPR,* 2016

Evaluating cross-instance correspondence

- Idea: Use keypoint annotations
- Problem: can only match nearby poses



Evaluating cross-instance correspondence



From small to large baselines

- Given 3D models, can construct ground truth

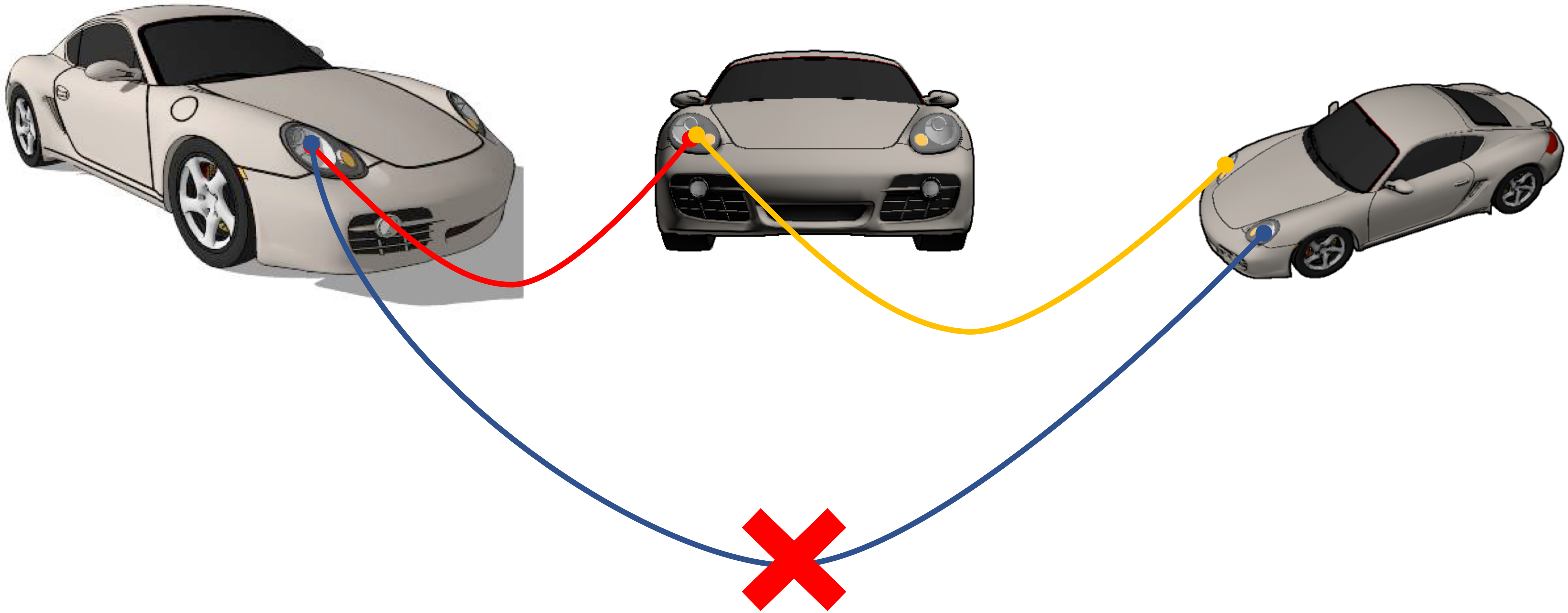
$$\vec{\mathbf{p}}^{(1)} = \mathbf{K}_1[\mathbf{R}_1|\mathbf{t}_1]\vec{\mathbf{P}} \qquad \vec{\mathbf{p}}^{(2)} = \mathbf{K}_2[\mathbf{R}_2|\mathbf{t}_2]\vec{\mathbf{P}}$$

From small to large baselines

- Additional output required: “matchability”

The cycle consistency constraint

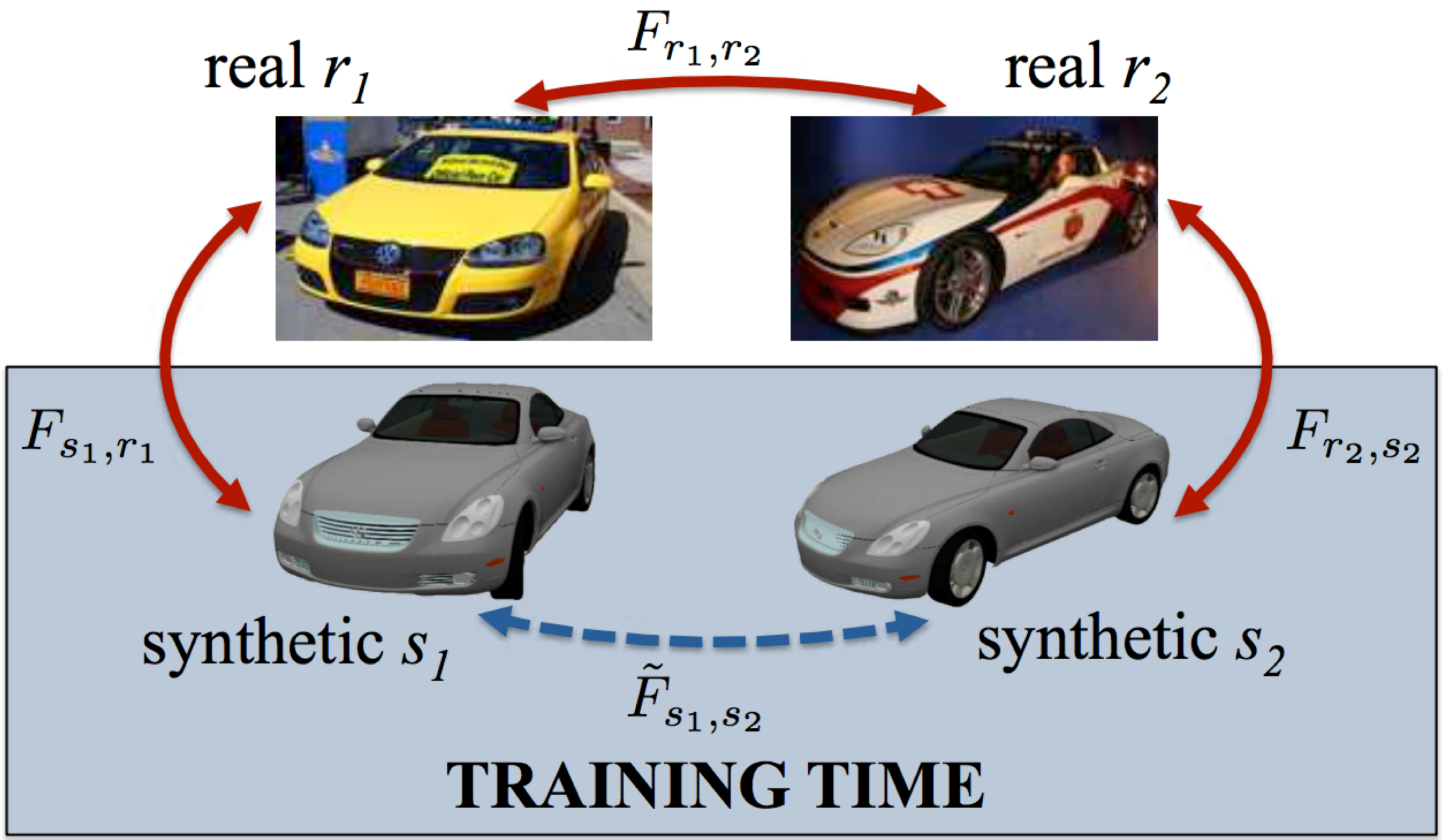
- Can we use unannotated images?



The cycle consistency constraint

The cycle consistency constraint

$$L(\tilde{F}_{s_1,s_2} - F_{s_1,r_1} \circ F_{r_1,r_2} \circ F_{r_2,s_2})$$



Learning Dense Correspondence via 3D-guided Cycle Consistency. Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, Alexei A. Efros. In *CVPR,* 2016.

Vision and Language

Image captioning - The task

A group of young men playing soccer.

Image captioning - why?

- Alt-text for visually impaired
- Test for true understanding?

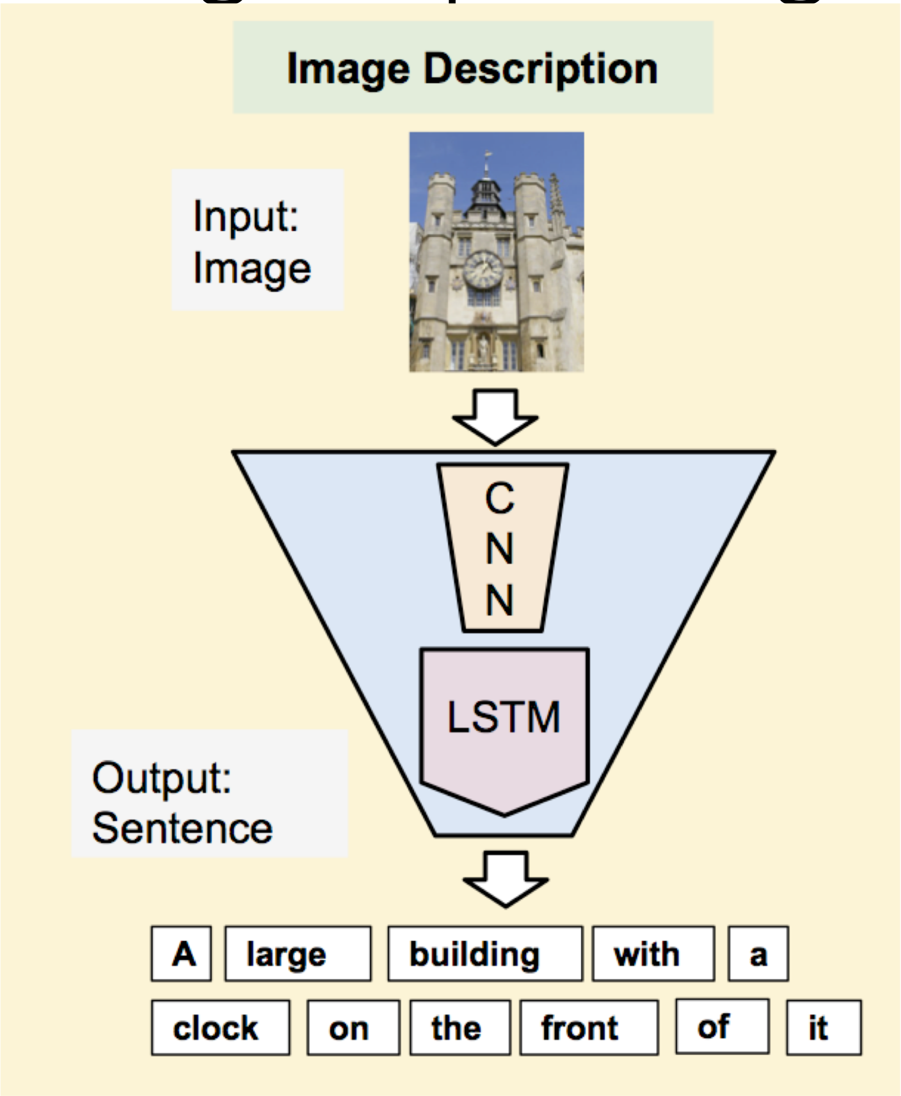
Image captioning - evaluation

- Given computer-generated caption and human caption, compute match
- BLEU from machine translation community
- Computes (modified) n-gram precision

Reference: A group of people playing soccer

Candidate: People playing baseball.
BLEU: 1/3

Image captioning

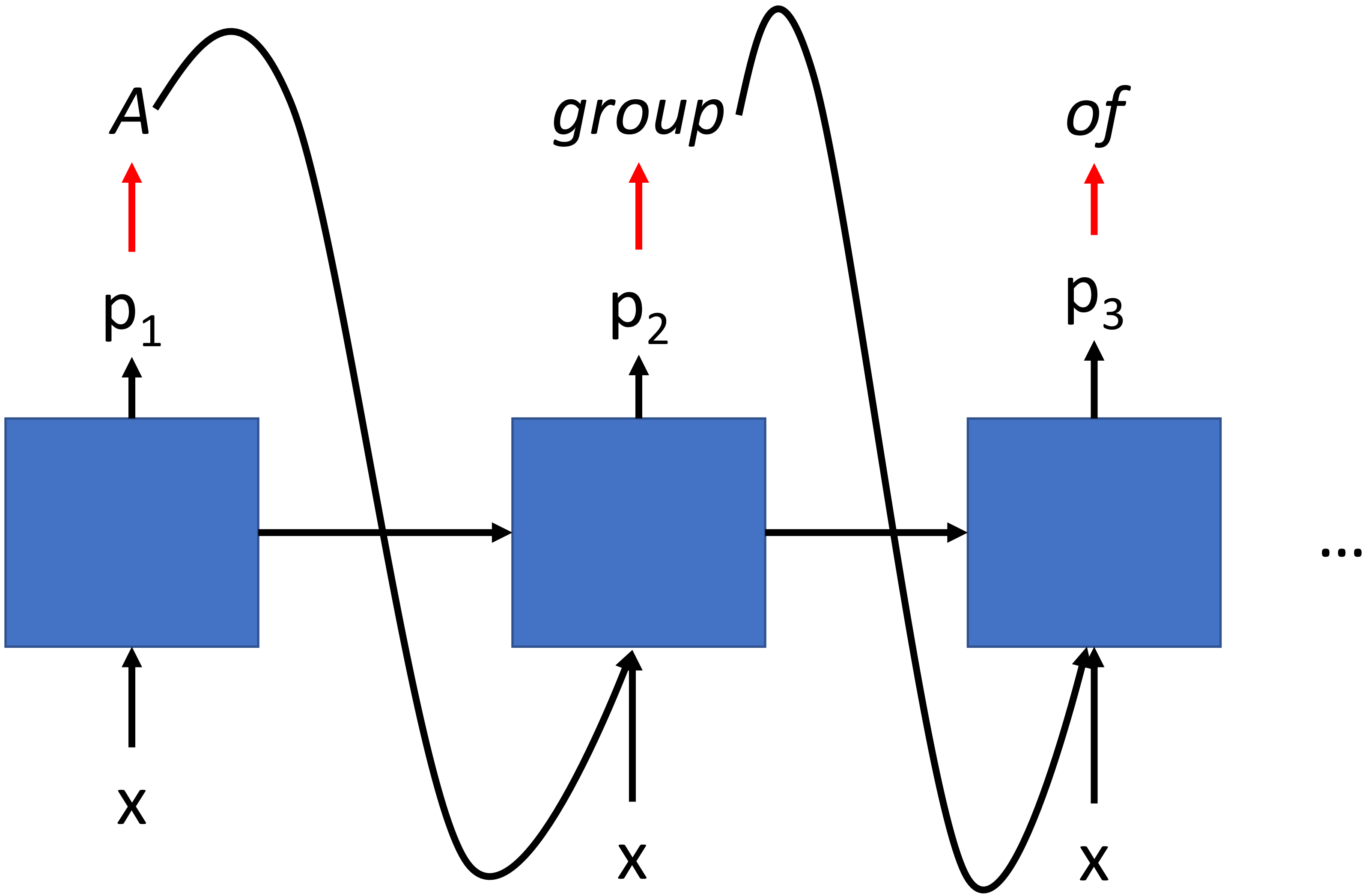


Long-term Recurrent Convolutional Networks. J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell. In *CVPR,* 2015.

Deep Visual-Semantic Alignments for Generating Image Descriptions. Andrej Karpathy and Li Fei-Fei. In *CVPR,* 2015

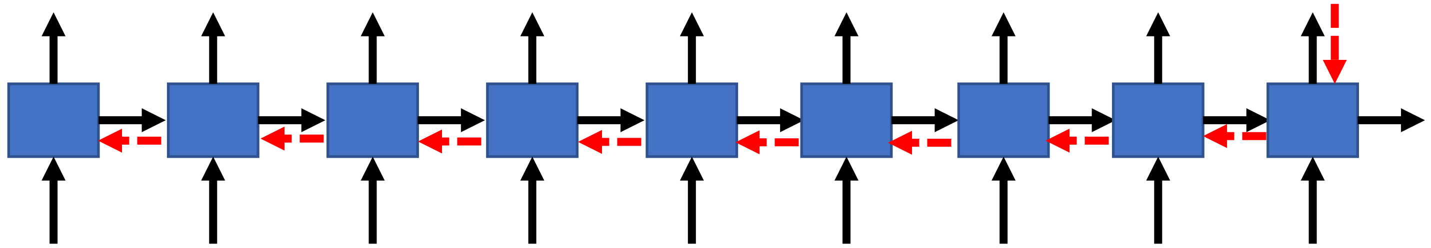
Show and tell: A neural image caption generator
Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan. In *CVPR,* 2015.

Generating sequences with Recurrent nets



Challenge with RNNs

- Long backpropagation paths
- Vanishing / exploding gradients



Long short term memory

- Key idea: maintain a register and simply add things at each time step

- $c_t = c_{t-1} + x_t$

Long short term memory

- Key idea: maintain a register and simply add things at each time step
- Maybe don’t add input directly, do some processing first

- $c_t = c_{t-1} + g(x_t, c_{t-1})$

Long short term memory

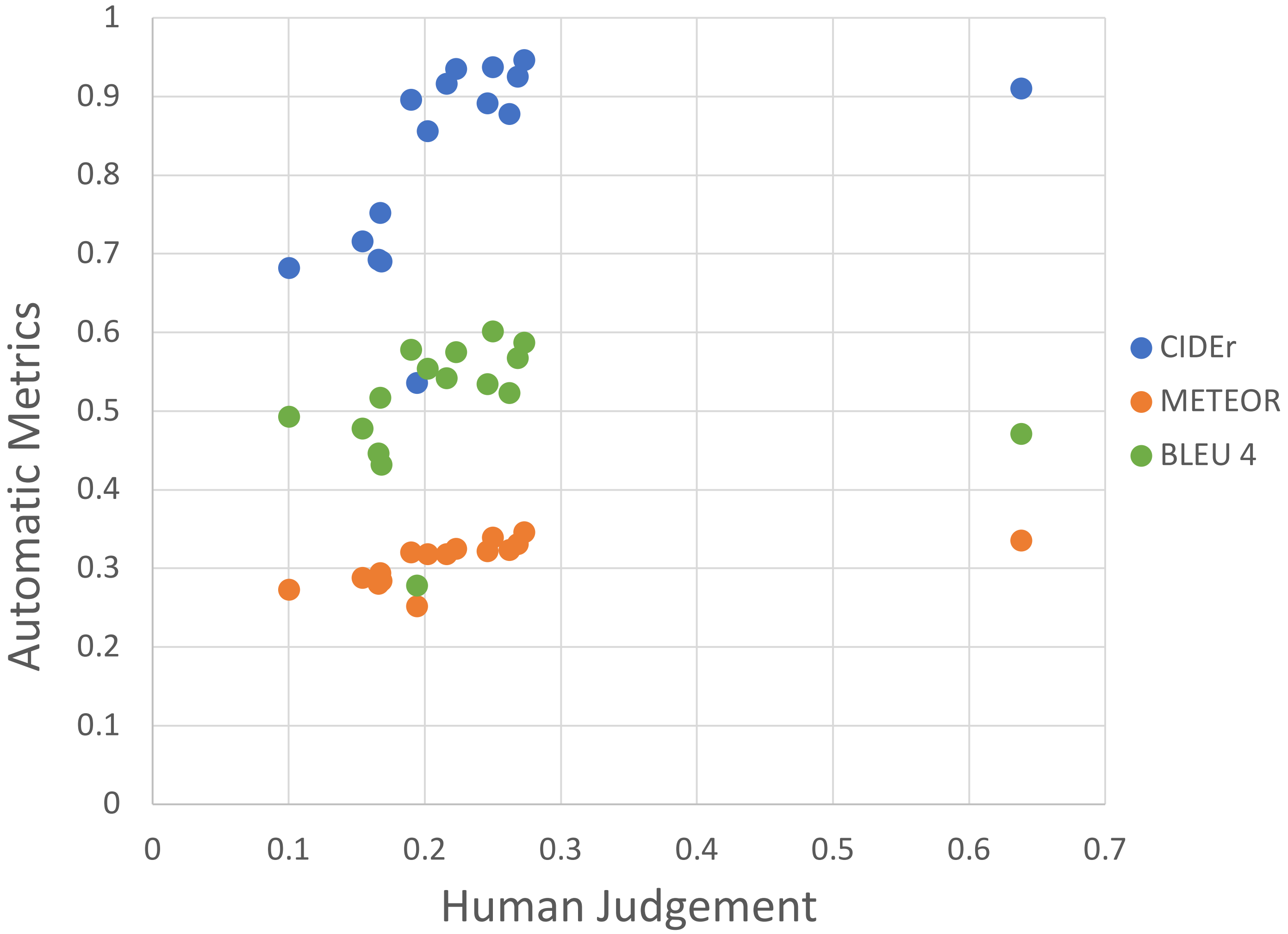
- Key idea: maintain a register and simply add things at each time step
- Maybe don’t add input directly, do some processing
- Maybe add option to ignore some input

- $c_t = c_{t-1} + i(x_t, c_{t-1}) \odot g(x_t, c_{t-1})$

Long short term memory

- Key idea: maintain a register and simply add things at each time step
- Maybe don't add input directly, do some processing
- Maybe add option to ignore some input
- Maybe add option to forget previous register state
- $c_t = f(c_{t-1}, x_t) \odot c_{t-1} + i(x_t, c_{t-1}) \odot g(x_t, c_{t-1})$
- Use register to produce output

Evaluation Metrics



Slide credit: Larry Zitnick

Evaluation Metrics

Human captions

Slide credit: Larry Zitnick

A man riding a wave on a surfboard in the water.

Slide credit: Larry Zitnick

A man riding a wave on a surfboard in the water.

“surfboard”

Slide credit: Larry Zitnick

vemödalen - n. the frustration of photographing something amazing when thousands of identical photos already exist

vemödalen - n. the fear that our datasets will never have the diversity we desire

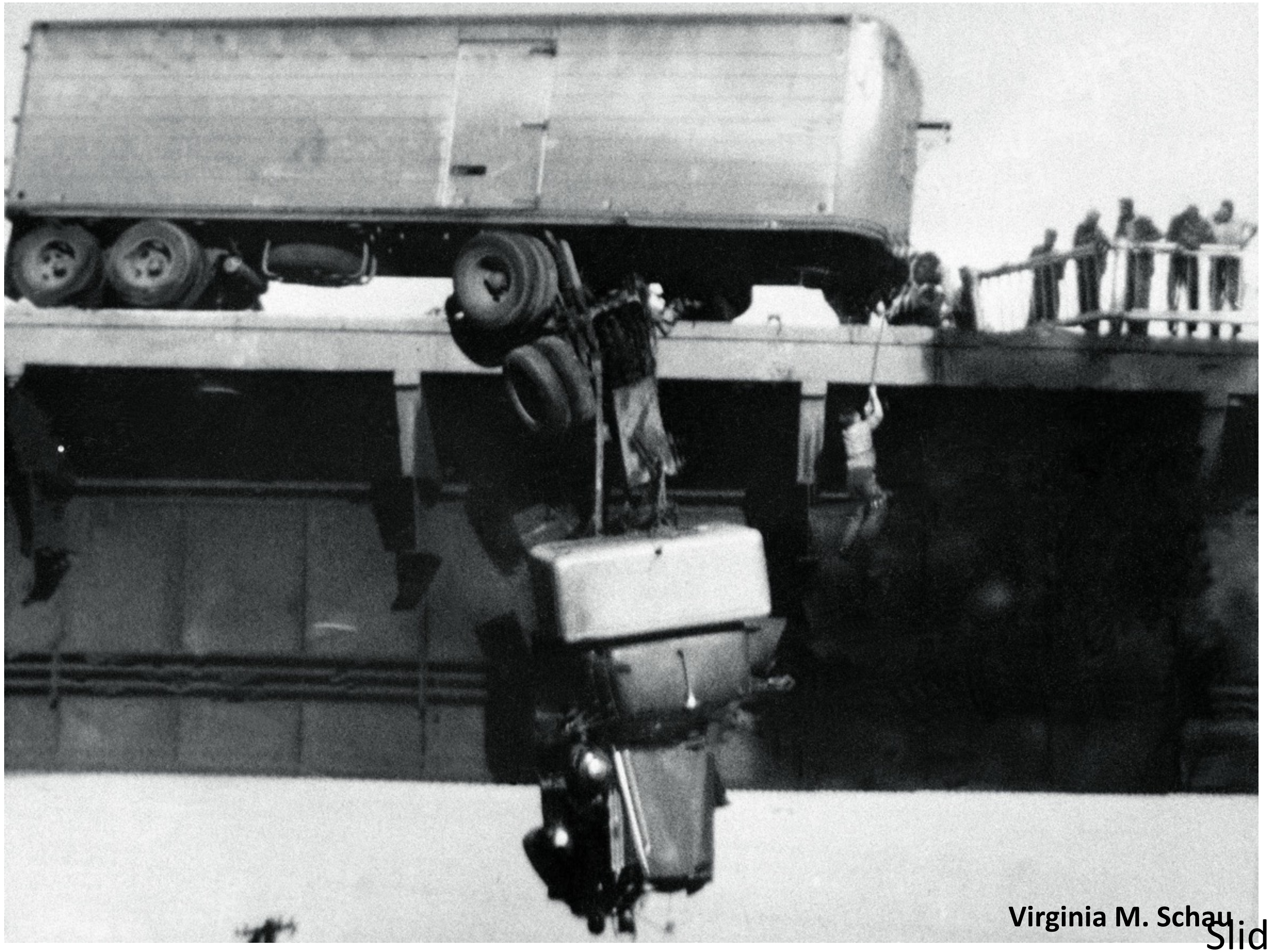
Vemödalen: The Fear That Everything Has Already Been Done
https://www.youtube.com/watch?v=8ftDjebw8aA

Slide credit: Larry Zitnick

The post-captioning world

- Captioning is hard to evaluate!
  - Frame task so that it is easy to evaluate objectively
- Datasets are biased!
  - Control dataset bias

A man is rescued from his truck that is hanging dangerously from a bridge.



Virginia M. Schau

Slide credit: Larry Zitnick

Stephanie Melnick
@unicornsteph96

Follow

I'm going to crush the rebellion... but first, let me take a selfie. #captionbot

I am not really confident, but I think it's a man taking a selfie in front of a building.

Reasoning

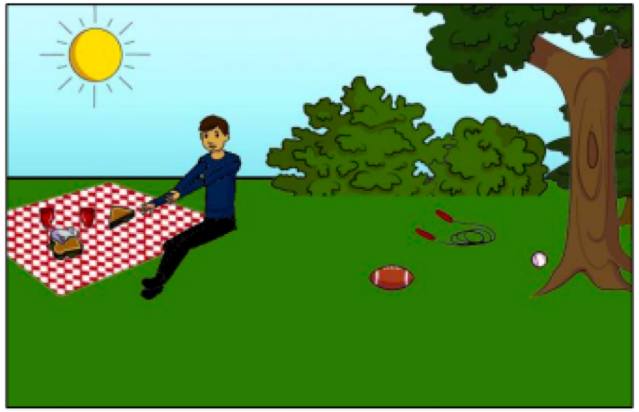
- Want vision systems to reason about what is going on
    - Identify objects and scenes
    - Identify relationships between objects
    - Understand physics of the world
    - Understand social interactions, intent etc.
    - Incorporate knowledge: common sense, pop culture, …

Visual Question Answering

- Direct motivation: assistive technology
- Indirect motivation: sandbox for reasoning

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

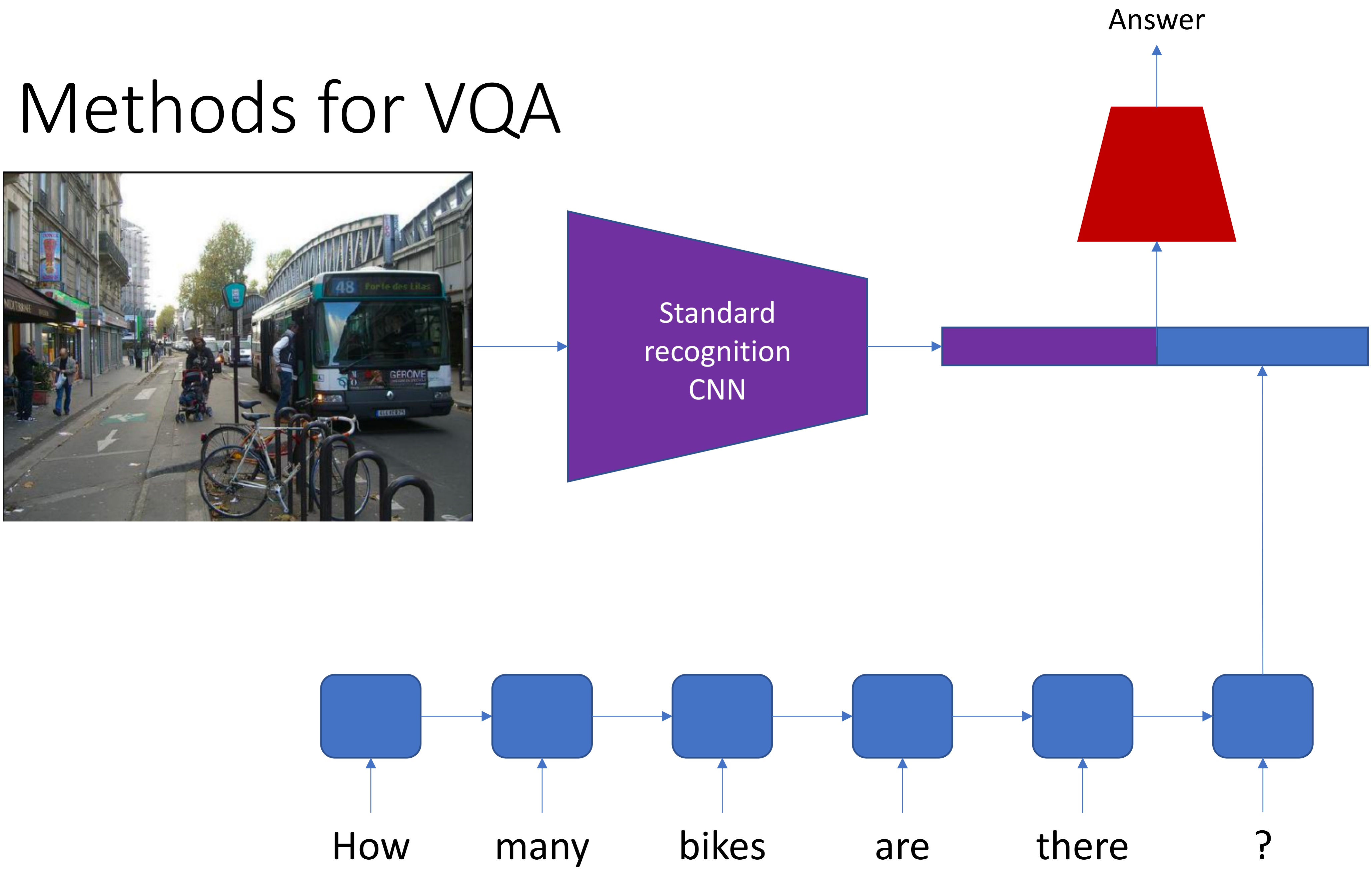


Is this person expecting company?
What is just under the tree?

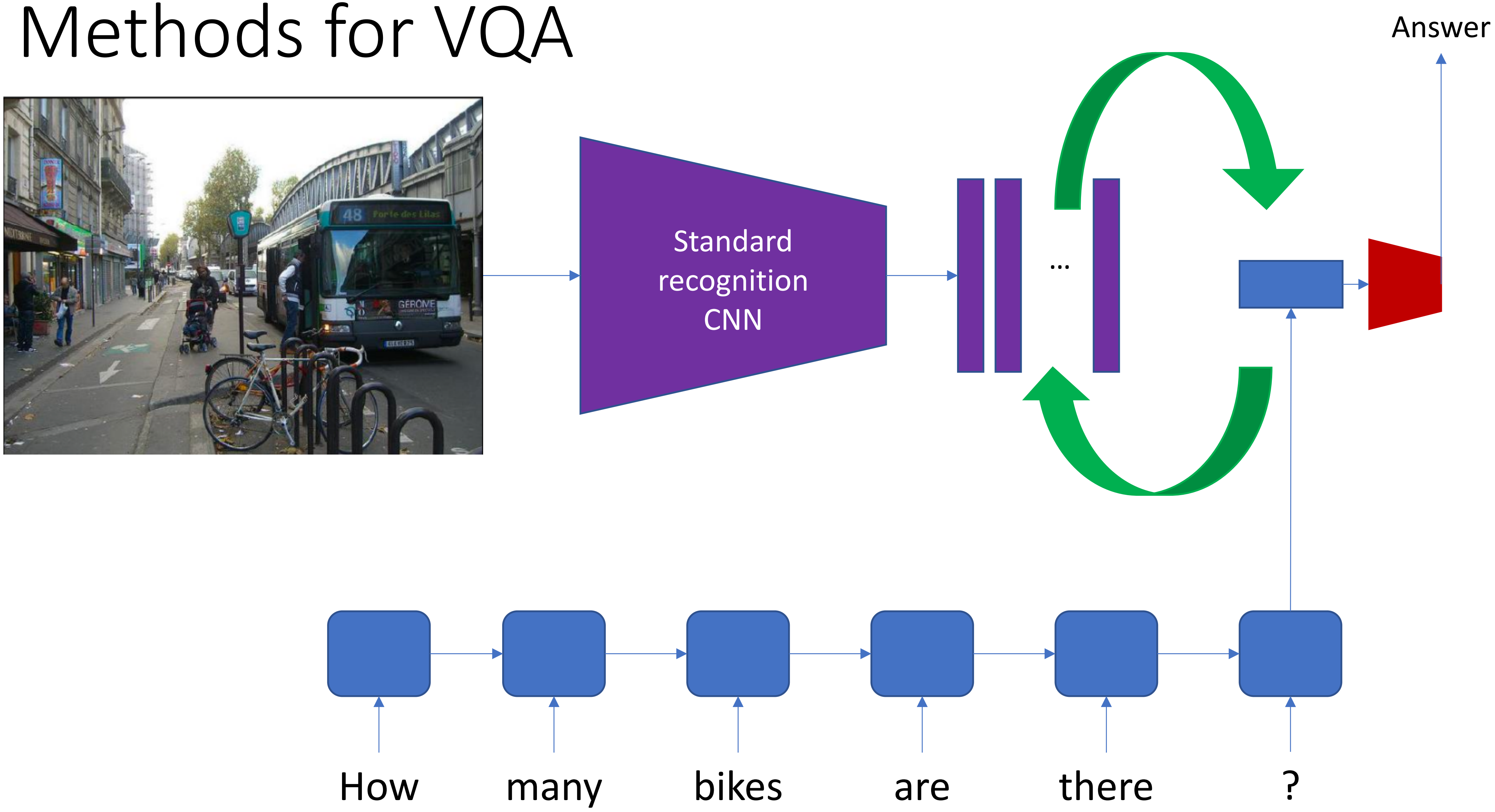
Does it appear to be rainy?
Does this person have 20/20 vision?

VQA: Visual Question Answering. Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh. In *ICCV,* 2015.

Methods for VQA

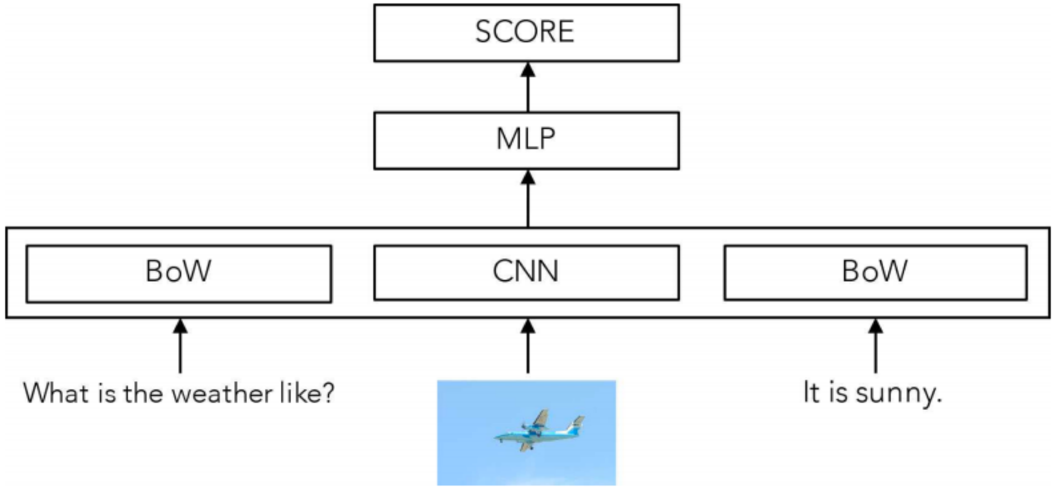


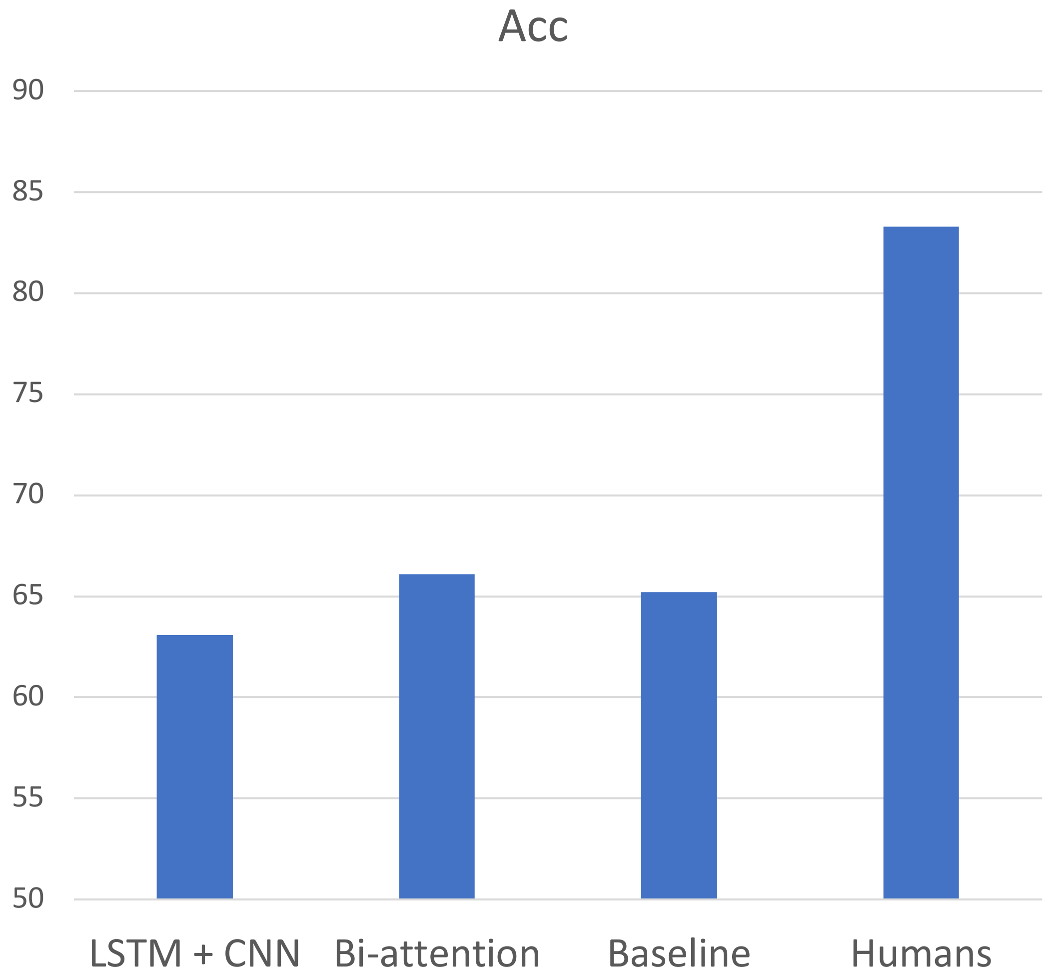
Methods for VQA



Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In CVPR, 2016

The Unreasonable Effectiveness of Baselines





A. Jabri, A. Joulin, L. van der Maaten. Revisiting visual question answering baselines. In *ECCV,* 2016.

The problem with VQA

- Dataset biases allow cheating
  - Only-question Bag-of-Words: 53.7% (vs ~65% for state-of-the-art)
- Require common sense to answer
  - “What is the moustache made of?”
- Hard to diagnose error
  - Is the problem understanding the question?
  - Or understanding the image?

What color are her eyes?
What is the mustache made of?

Clever Hans