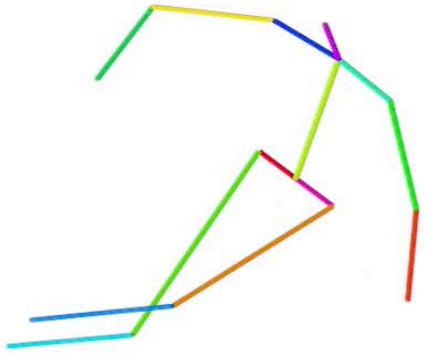
Athletic Pose Estimation

Pascal Fua
EPFL, CVLab

Activity Specific Training Data Required

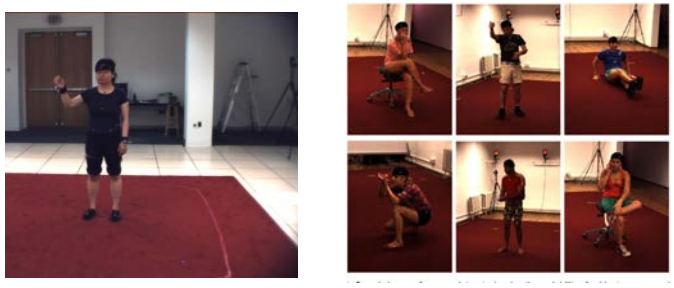
Cropped input

Training on Human 3.6M is
insufficient to reconstruct ski motion



Baseline
(trained on H3.6M)

Finding more Data?!



HumanEva & Human3.6M
[Sigal 2010, Ionescu 2014]

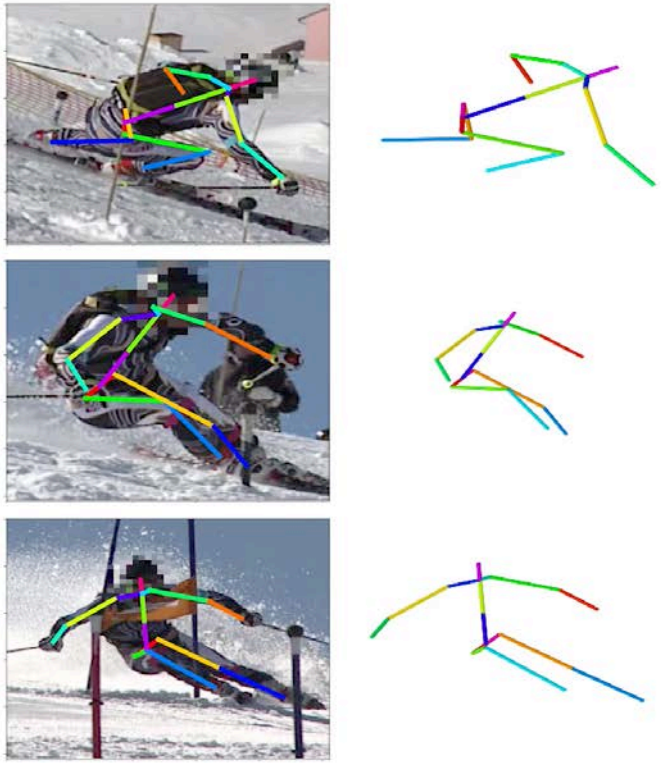
EgoCap Dataset
[Chen 2016]

SURREAL Dataset
[Varol 2017]

Image stitching
[Rogez 2016]

MPI-INF-3DHP
[Me**Rh**CaFuSoXuTh, 3DV 2017]

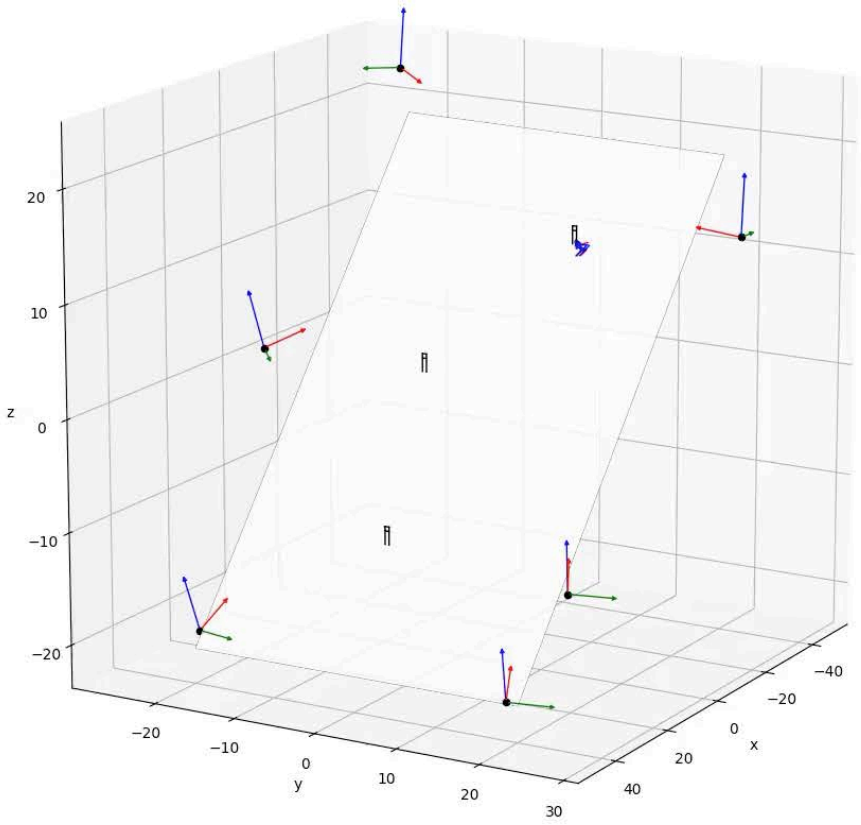
Ski-Pose PTZ-Camera Dataset

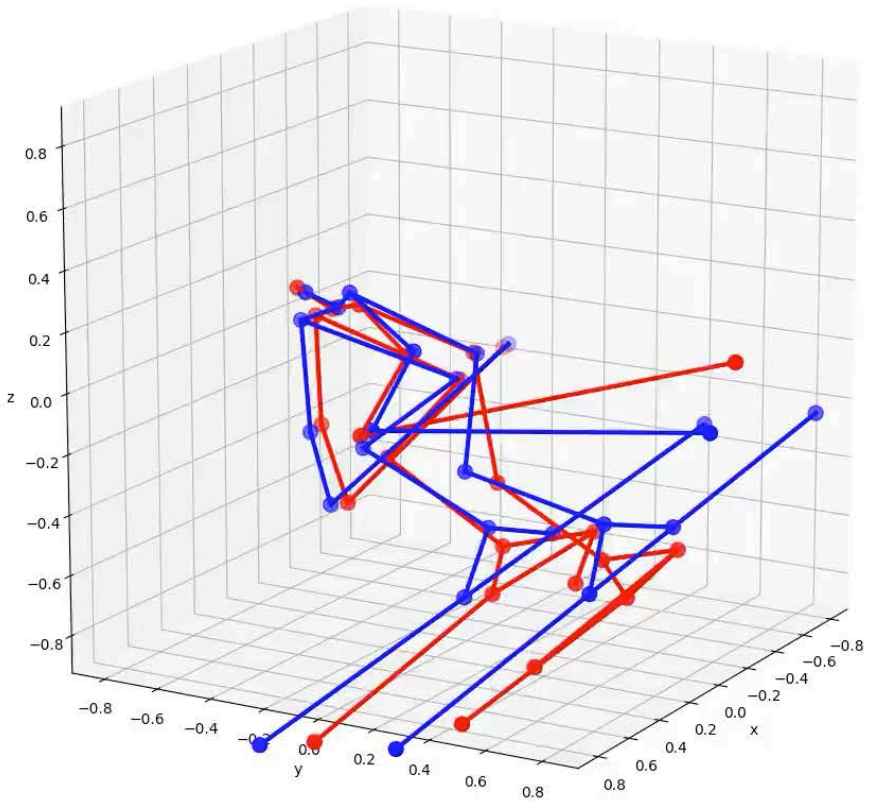


Available at

https://cvlab.epfl.ch/Ski-PosePTZ-Dataset

Full Supervision using Ski Data

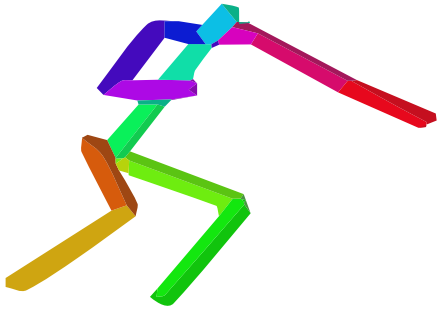


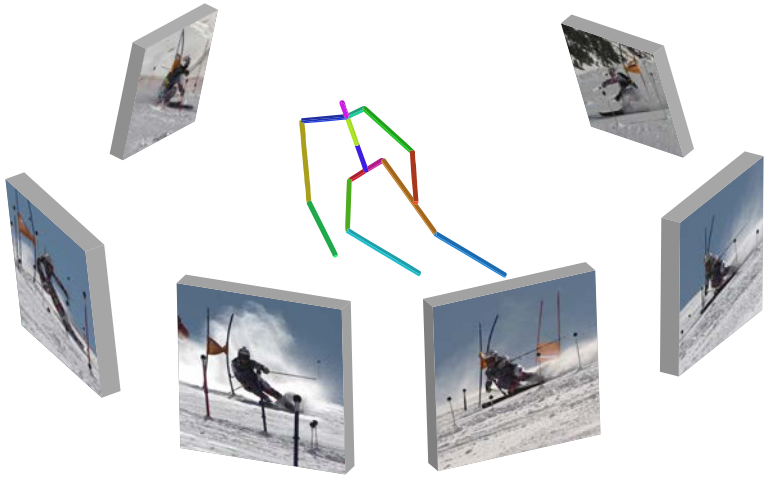


… but an unfortunate PhD student had to spend a loooong time annotating!

Multiview Semi-Supervised Training
CVPR'18

Semi-Supervised Training

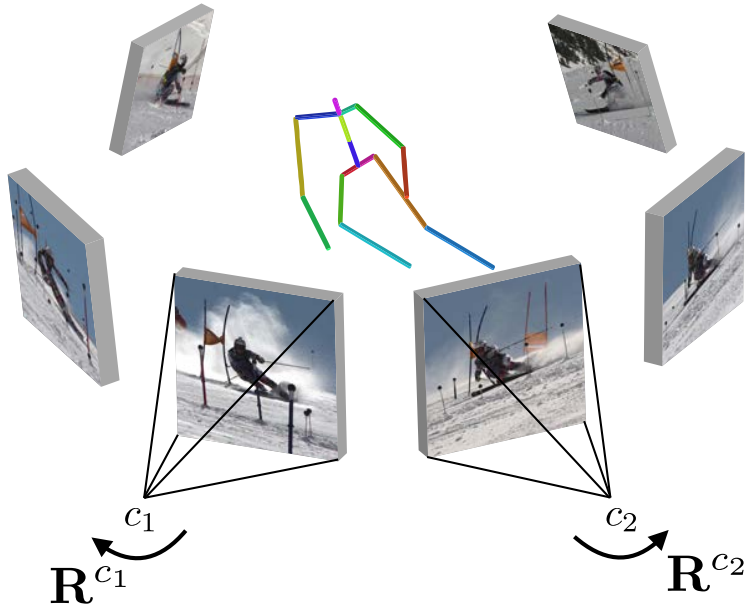


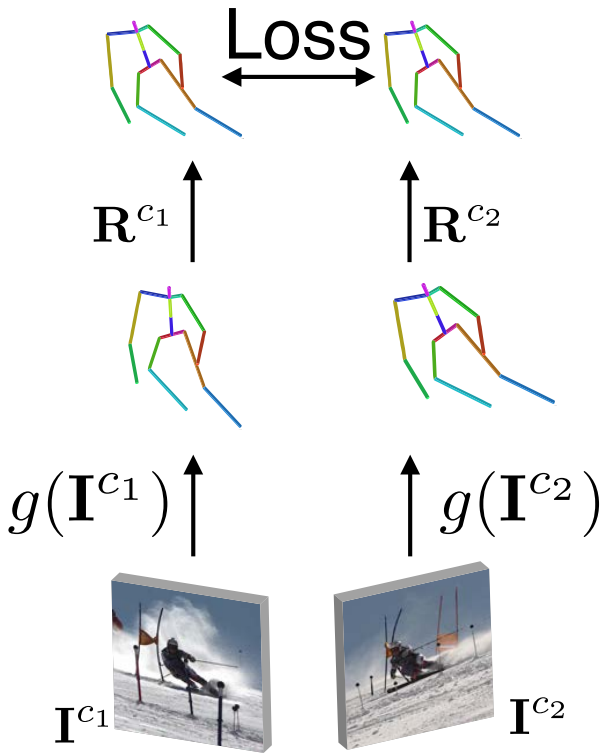


Labeled

Unlabeled

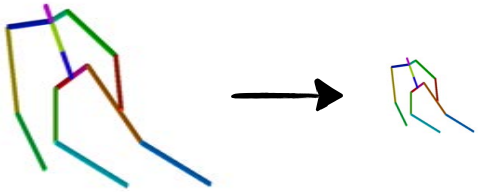
Multiview Consistency Constraint



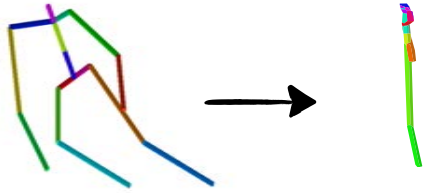


Multiview loss: $\|\mathbf{R}^{c_1} g(\mathbf{I}^{c_1}) - \mathbf{R}^{c_2} g(\mathbf{I}^{c_2})\|$

Eliminating Trivial Solutions



Towards zero



Rotational symmetry

Normalized loss

$$d(\mathbf{p}_1, \mathbf{p}_2) = \left\| \frac{\mathbf{p}_1}{\|\mathbf{p}_1\|} - \frac{\mathbf{p}_2}{\|\mathbf{p}_2\|} \right\|^2$$

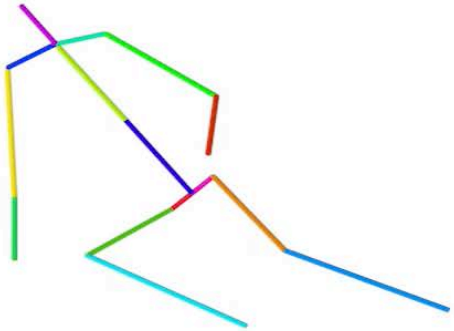
Regularization

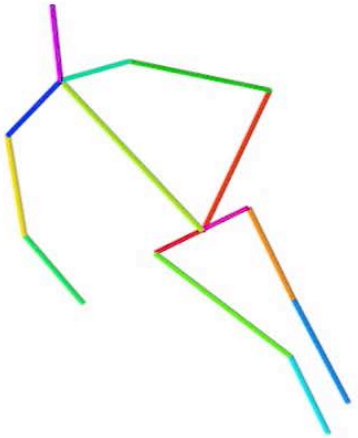
$$\|g(\mathbf{I}^i) - \hat{g}(\mathbf{I}^i)\|$$

—> Effective but makes the training more difficult.

Qualitative Improvement

Training on the ski dataset with weak multi-view supervision improves accuracy.





Our method* (trained on new ski dataset)

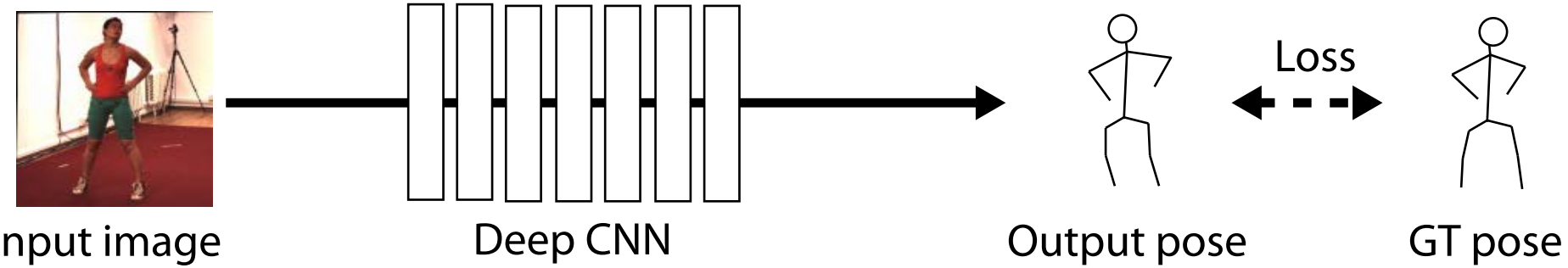
Baseline* (trained on H3.6M)

*smoothed temporally with a Gaussian window of std=1

Geometry-Aware 3D Body Representation

ECCV'19

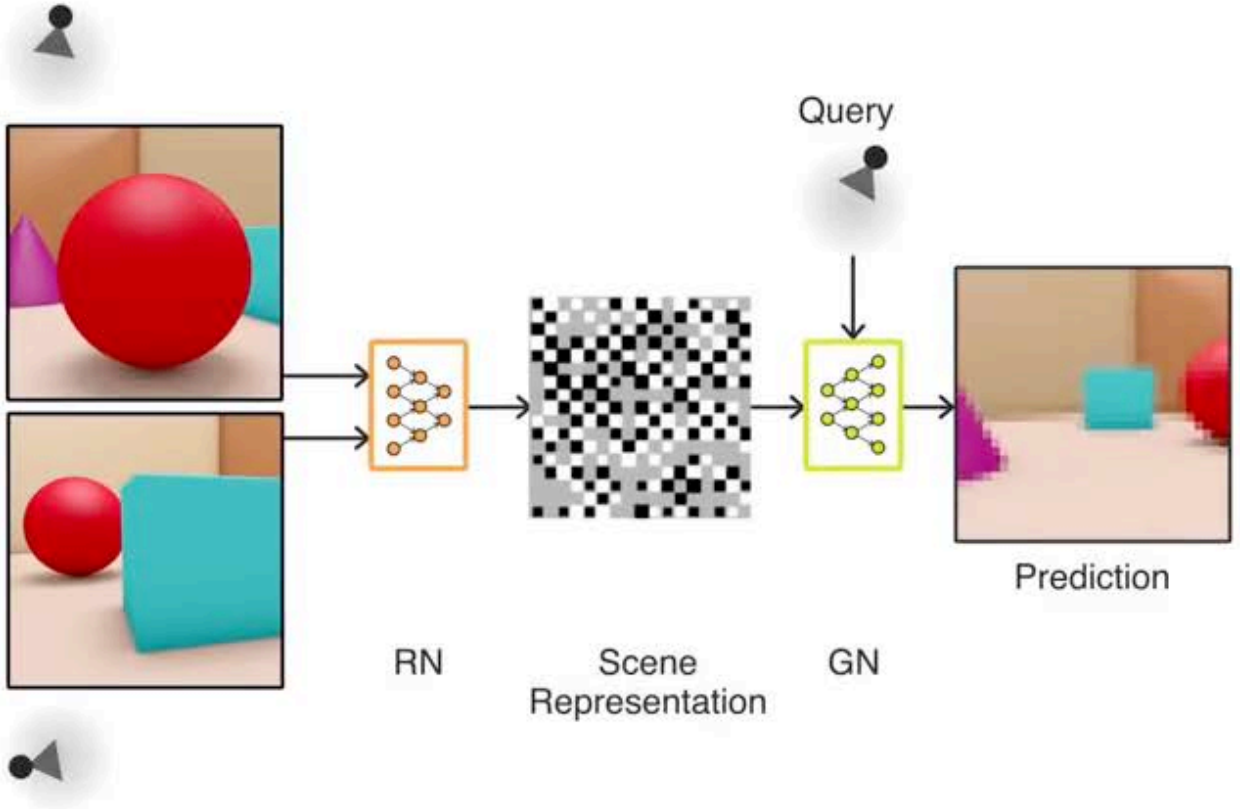
The Problem with Direct Regression



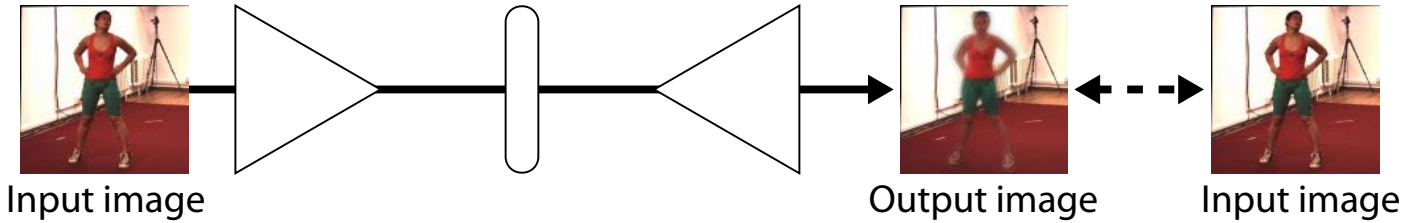
- The human body has many degrees of freedom.
- Going directly from image to 3D pose requires a very deep net.
- Training such a deep net requires a lot of training data.

Can we learn a representation that has fewer degrees of freedom for specific activities?

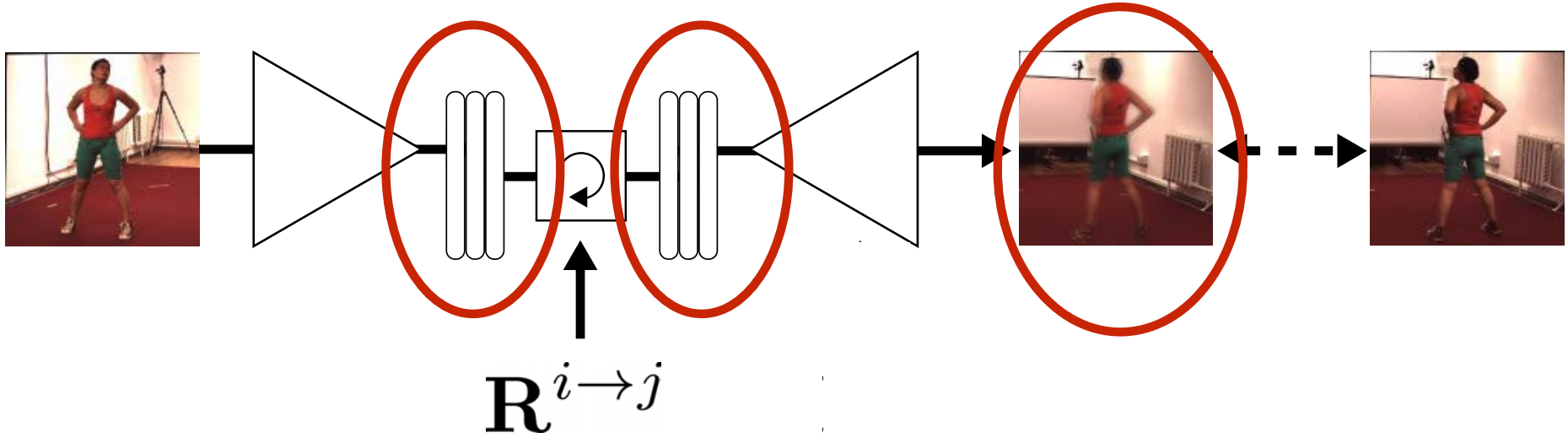
Novel View Synthesis



Autoencoders for Novel View Synthesis

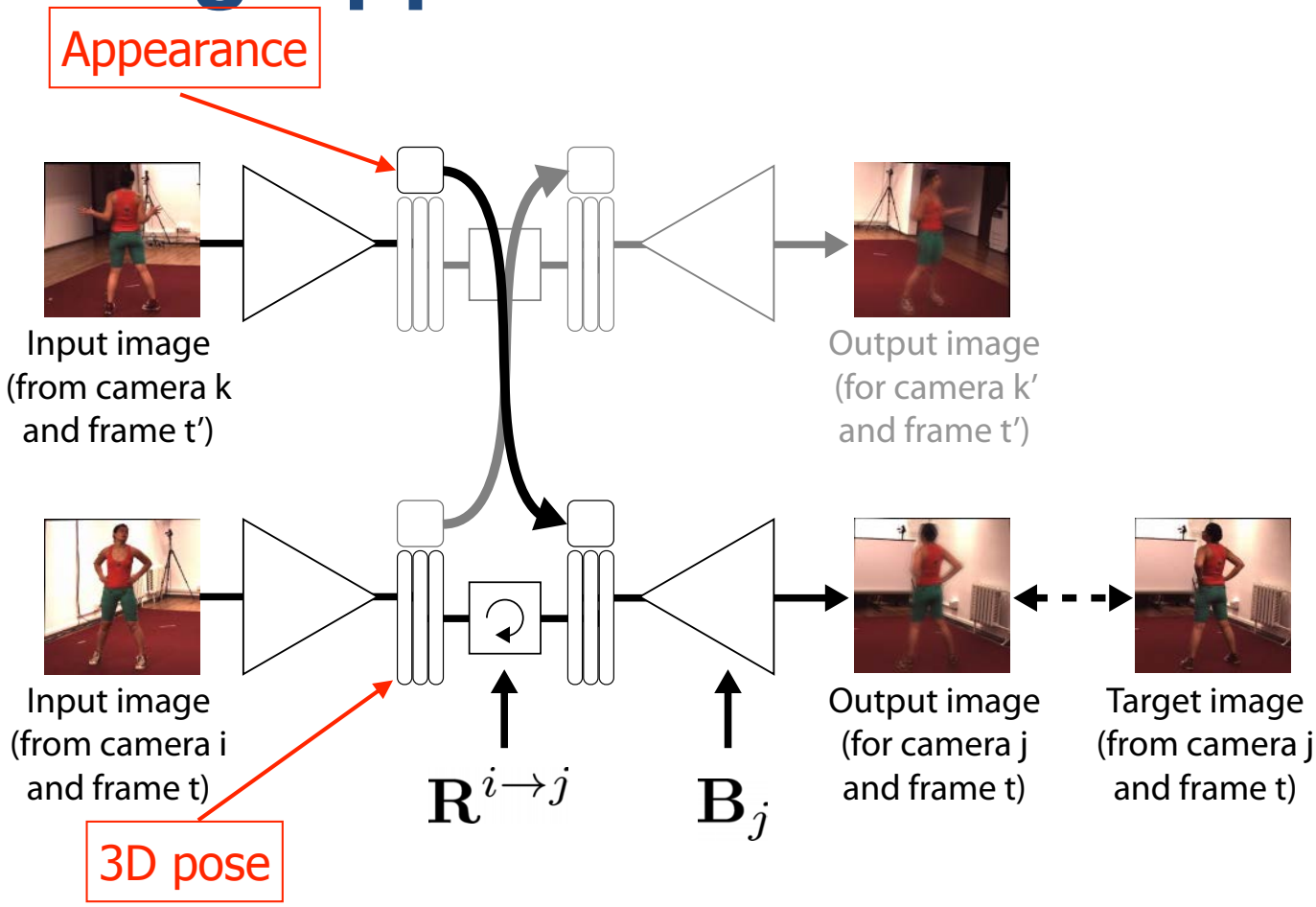


Conventional Autoencoder



Rotation Aware Autoencoder

Separating Appearance from Geometry



- The latent representation comprises a Nx3 matrix that encodes the 3D pose and a separate vector that models appearance.
- Before decoding the appearance vectors are swapped to ensure that they are similar in different images.

Rhodin et al. ECCV'19

Latent Representation

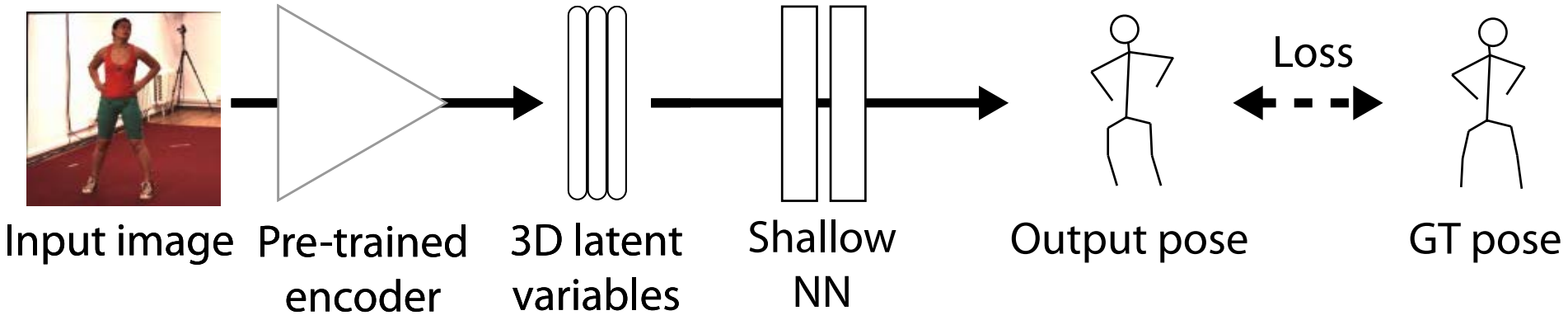
Input

Geometric encoding
(rotating point cloud)

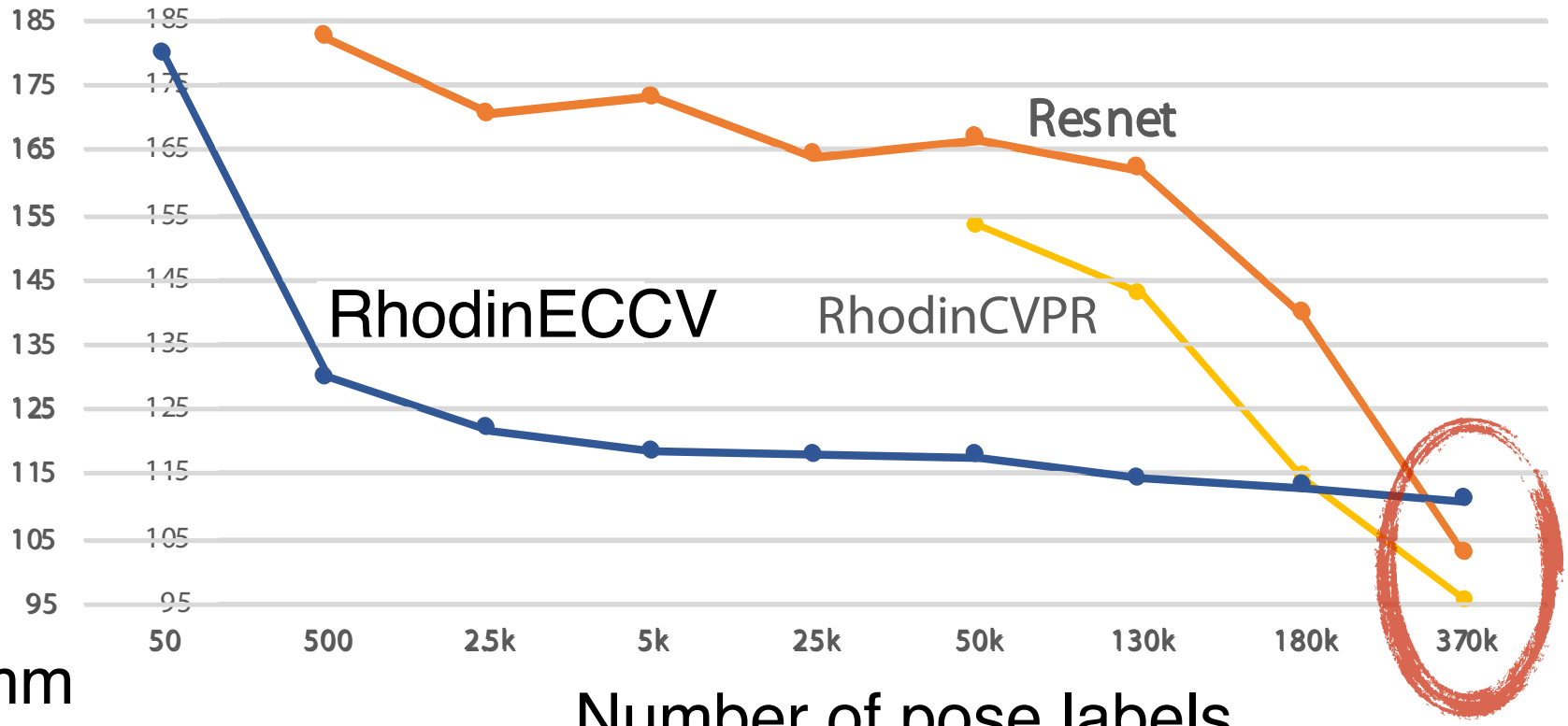
Output

The test subject is reconstructed with the right pose but an approximate appearance.

Direct Regression vs Using Latent Variables



Quantitative Comparison



NMJPE in mm

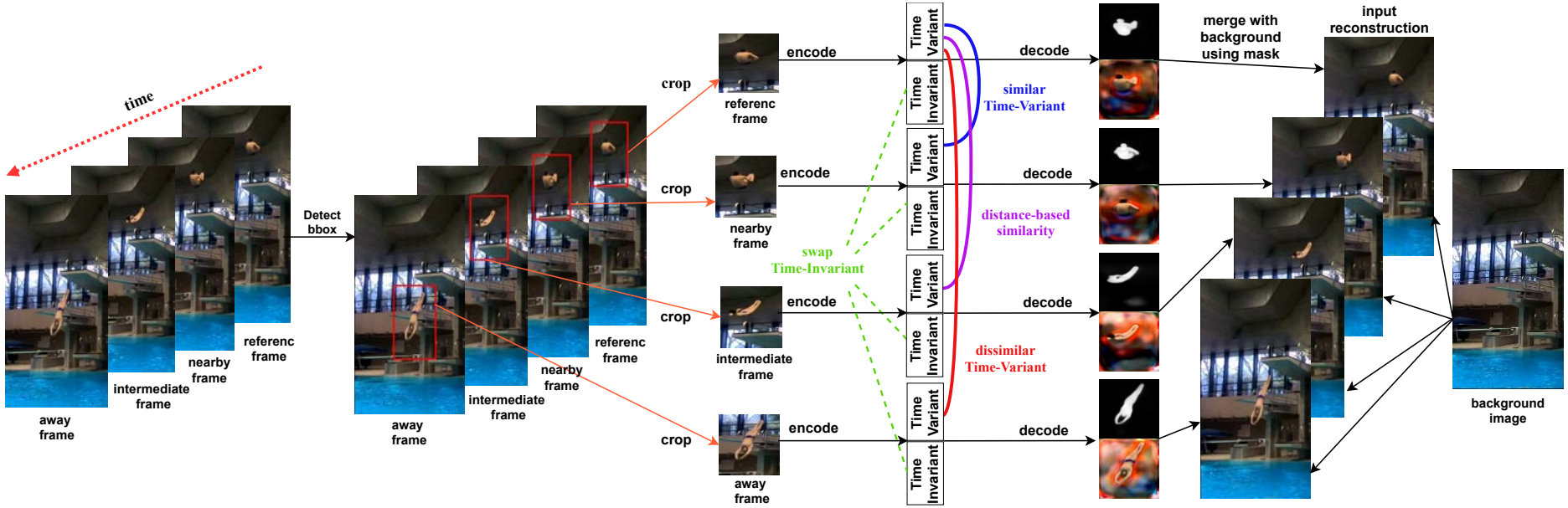
Number of pose labels

Use more people to learn the representation?

Unsupervised Learning on Monocular Videos

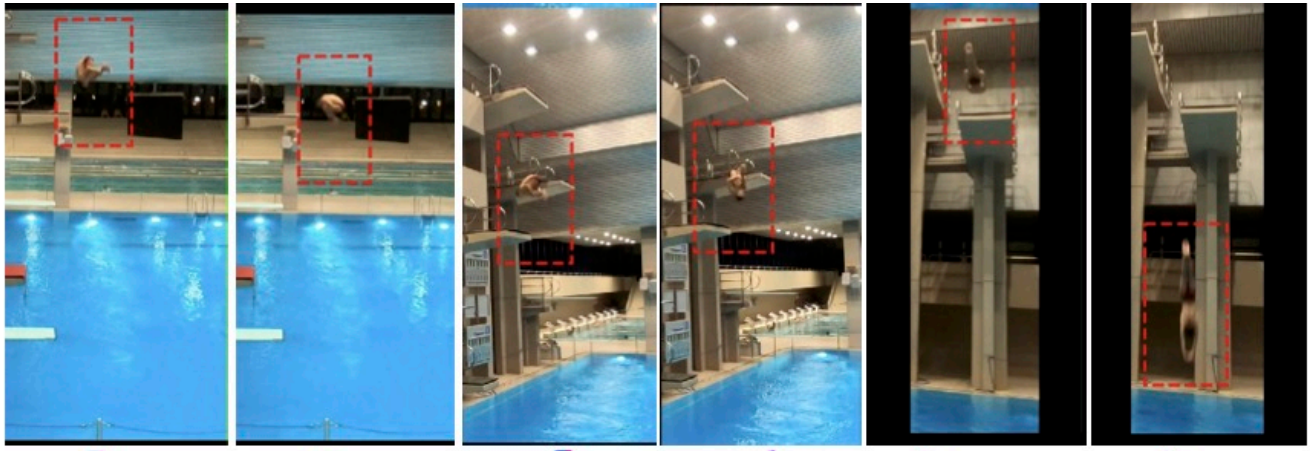
ArXiv'20

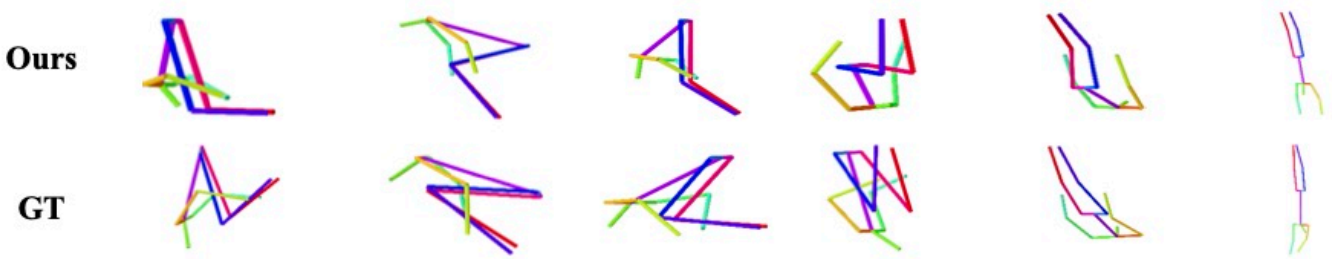
Contrastive Learning

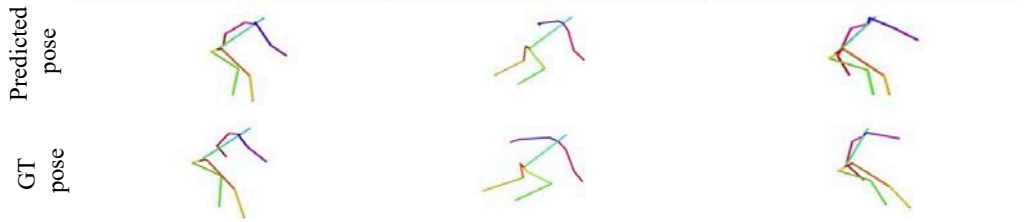


- Separate latent vectors into a time-variant component and a time-invariant one.
- Use contrastive learning to ensure that the time-variant part is usually more similar across close time intervals than long ones.
- Detect bounding boxes.
- Account for physics, in the case of divers, gravity.

Divers and Skiers







Tracking the Ball

- Ball state indicated in yellow at the top left
- Red bounding box denotes interaction between player and ball.
- Blue bounding box denotes ballistic flight.

Maksai et al. CVPR'16

Challenges

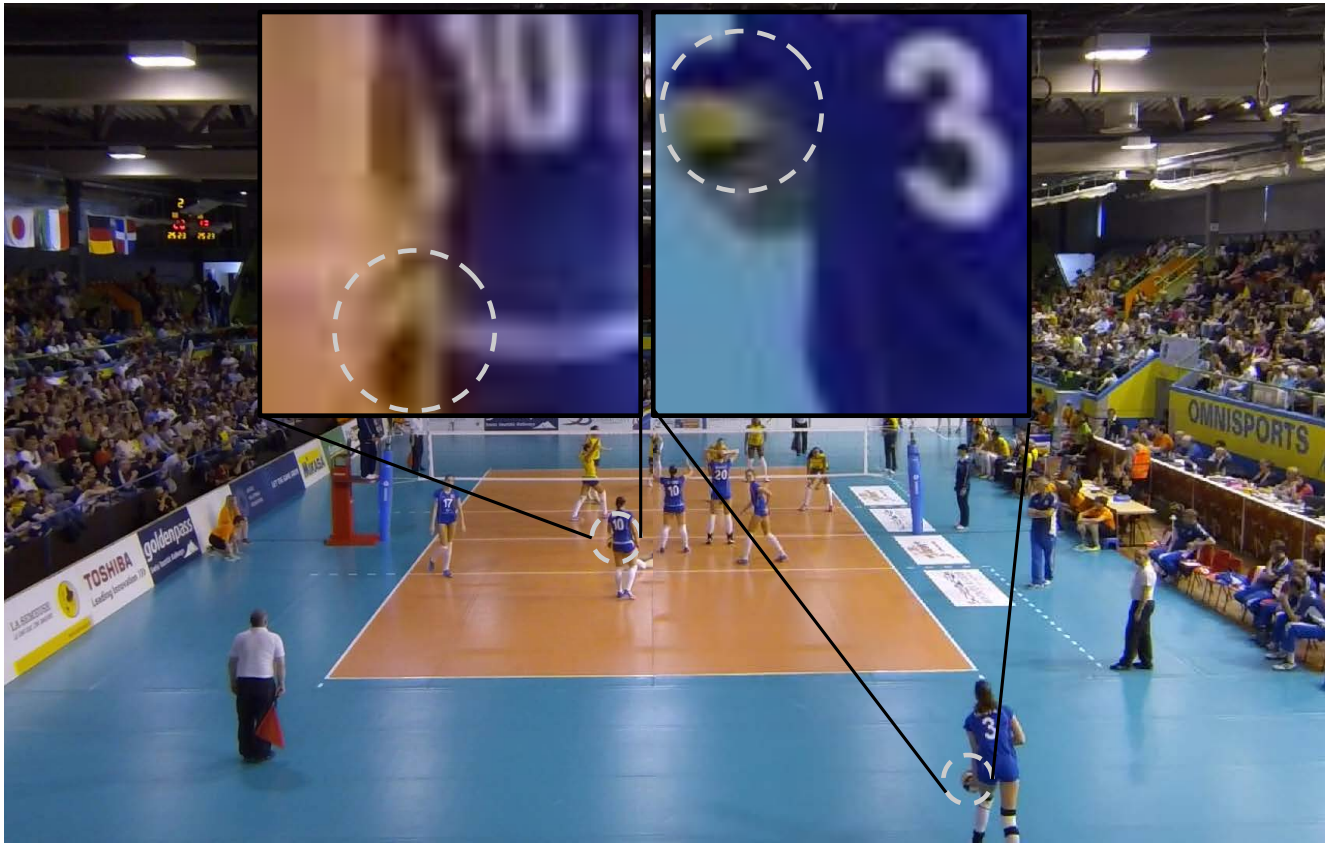
1. Fast motion and low visibility

Challenges

1. Fast motion and low visibility

2. Occlusions and hard to distinguishable appearances

Challenges



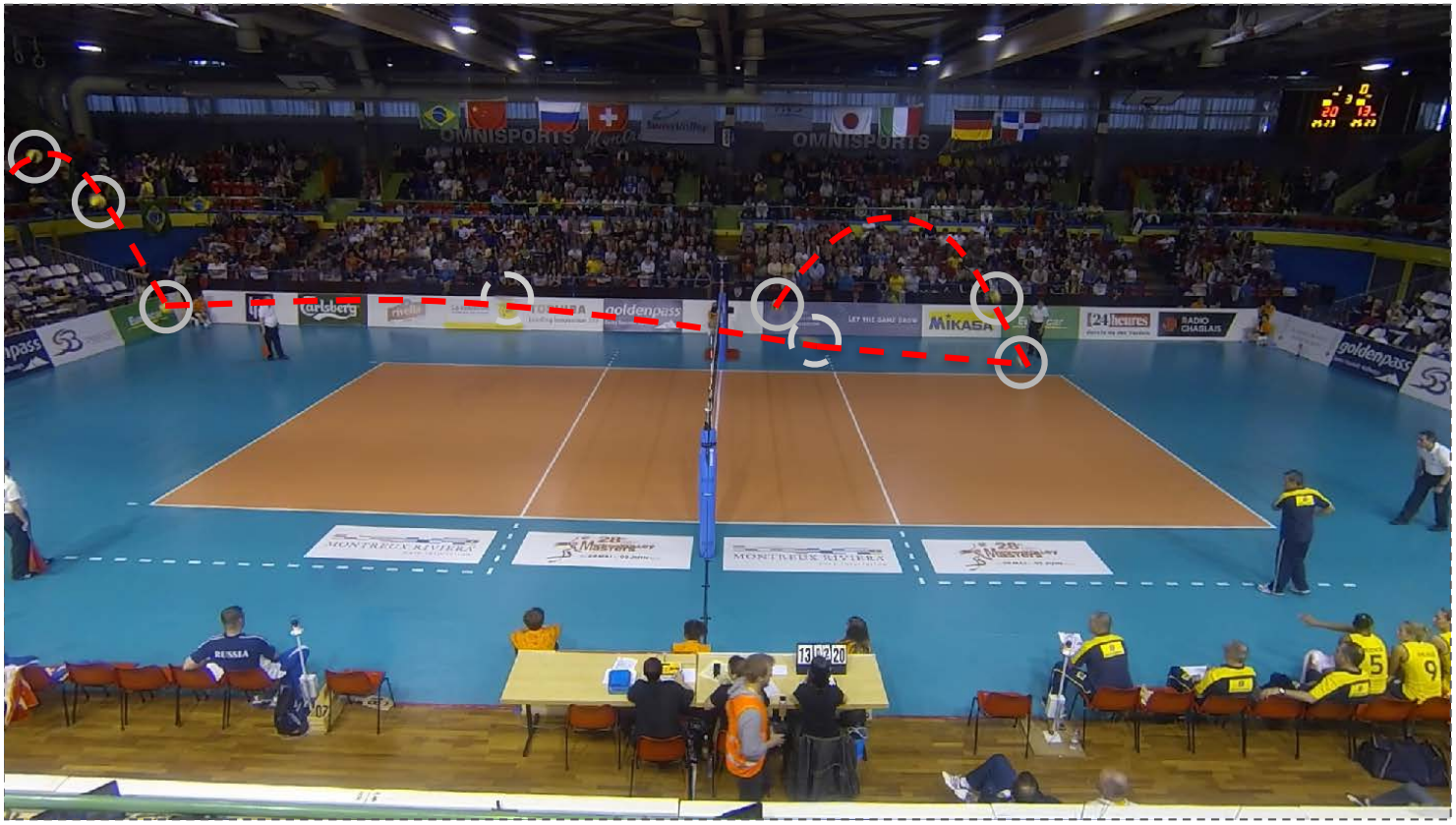
1. Fast motion and low visibility

2. Occlusions and very similar appearances

3. Prolonged occlusions

Standard Approaches

Tracking the ball with physical motion model



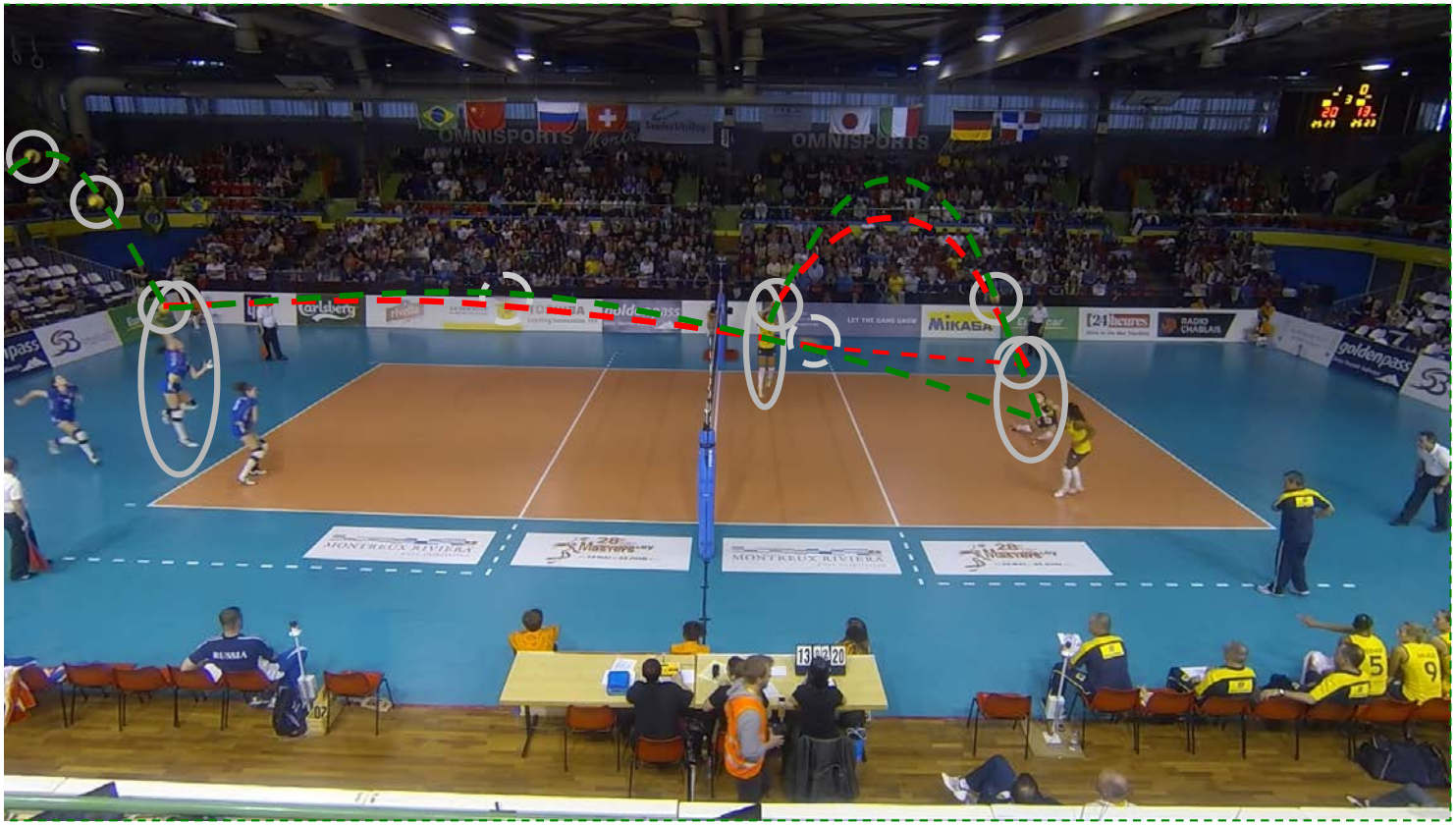
Physical models ignore interactions

Interaction models ignore physics

Combining the two is non-trivial

Our Approach

Accounting for interactions, physical model, and ball state



- **Physical model when the ball flies**
- **Interaction model when the ball is in possession**
- **Learning the state of the ball to distinguish between the two**

Weakly Supervised Methods

- Deep Networks are hard to train in sports scenarios because we rarely have large enough databases.
- However, we can exploit multi-view geometry and physics-based constraint to develop effective approaches to supervision.

Conclusion

Sidenblah et al., ECCV'00 AlexNet, ECCV'12 Urtasun et al., CVPR'06

Geometry and physics still rock!

Thanks To

- Victor Constantin
- Leonardo Citraro
- Sina Honari
- Andrii Maksai
- Erich Mueller
- Mirela Ostrek
- Helge Rhodin
- Mathieu Salzmann
- Jörg Spörri