

Tracking Emerges by Colorizing Videos

ECCV 2018

Carl Vondrick, Abhinav Shrivastava, Alireza Fathi,
Sergio Guadarrama, Kevin Murphy

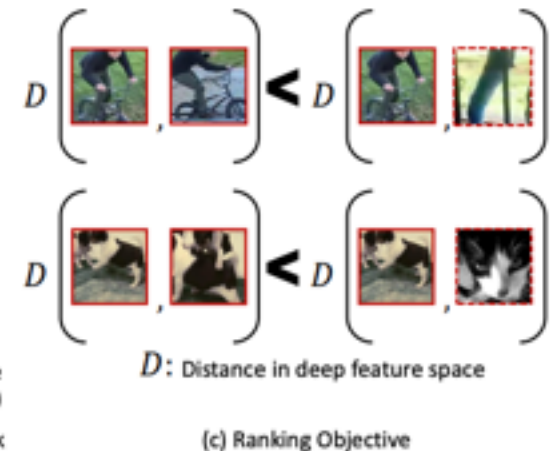
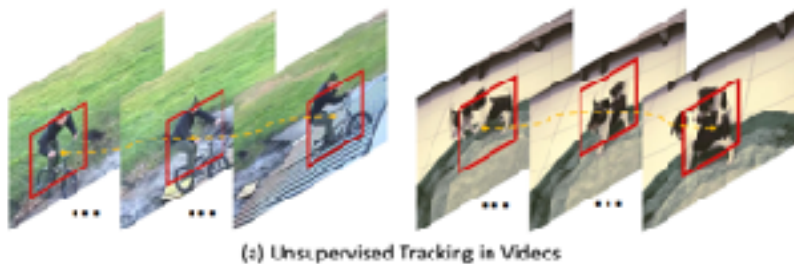
Motivation

- Collecting the large-scale tracking datasets often requires extensive effort that is impractical and expensive.



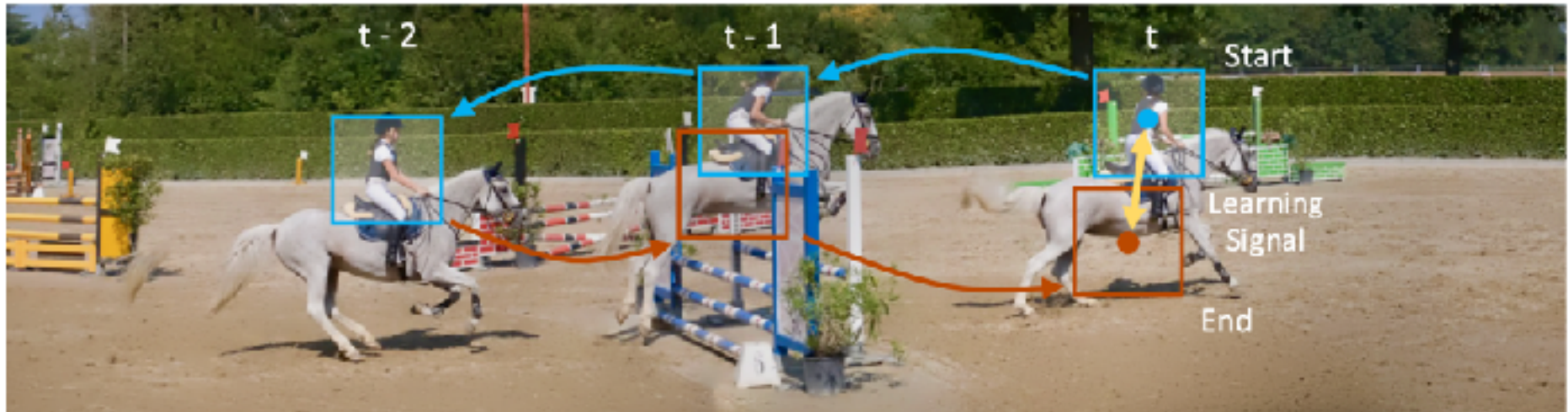
Related Work

- The key idea is to use unsupervised tracking methods to construct a good supervisory signal.



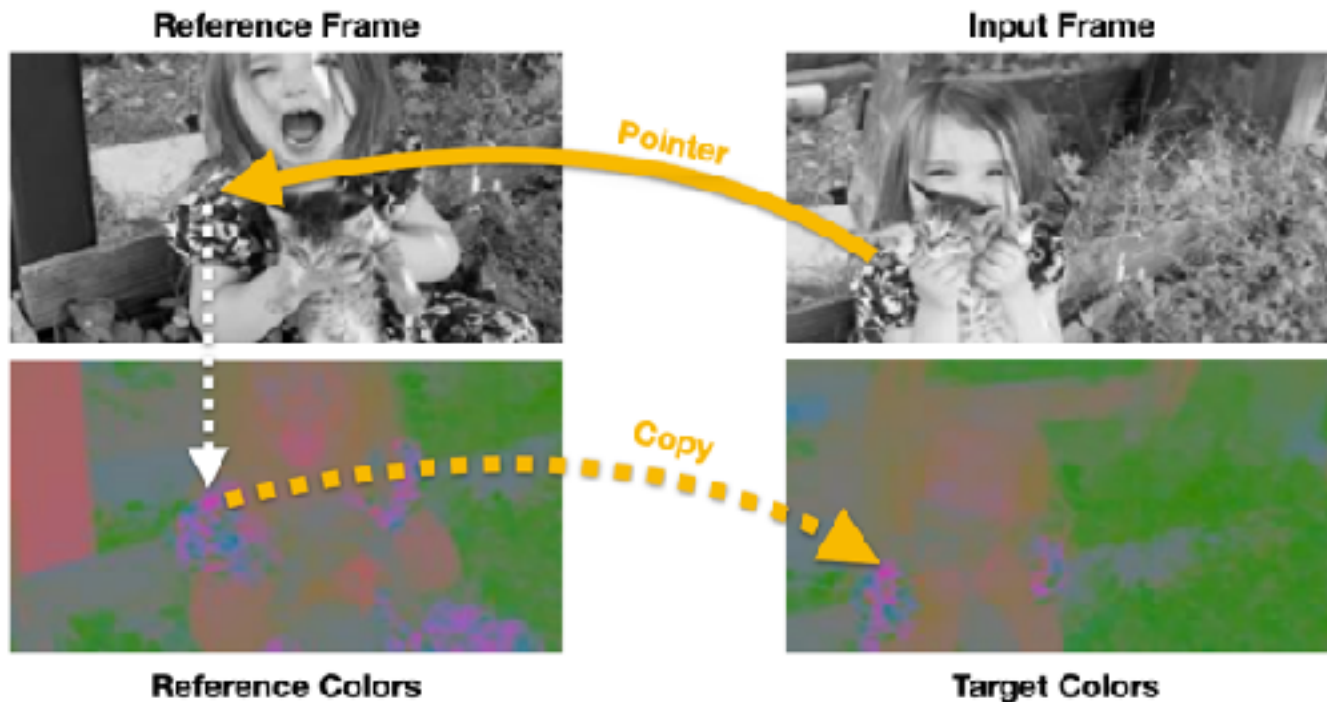
Related Work

- Cycle-consistency in time is a suitable learning objective for learning good features to track objects.



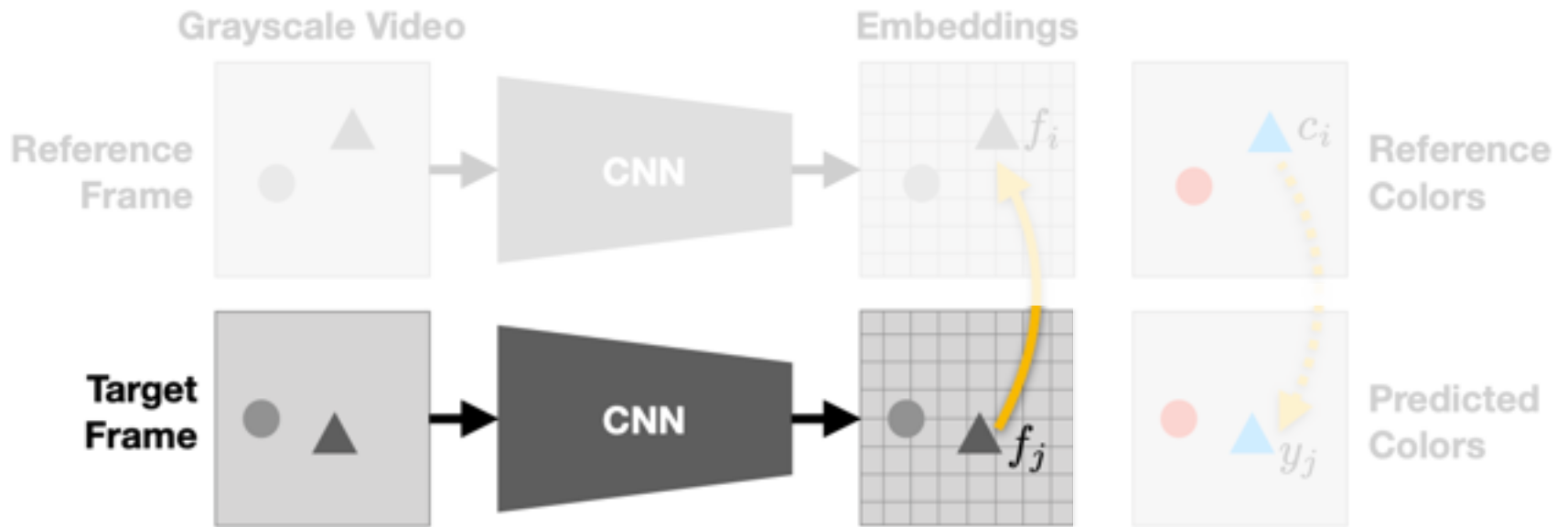
Proposed Solution

- The model is forced to learn to colorize gray-scale videos by copying colors from a reference frame to the target frame.



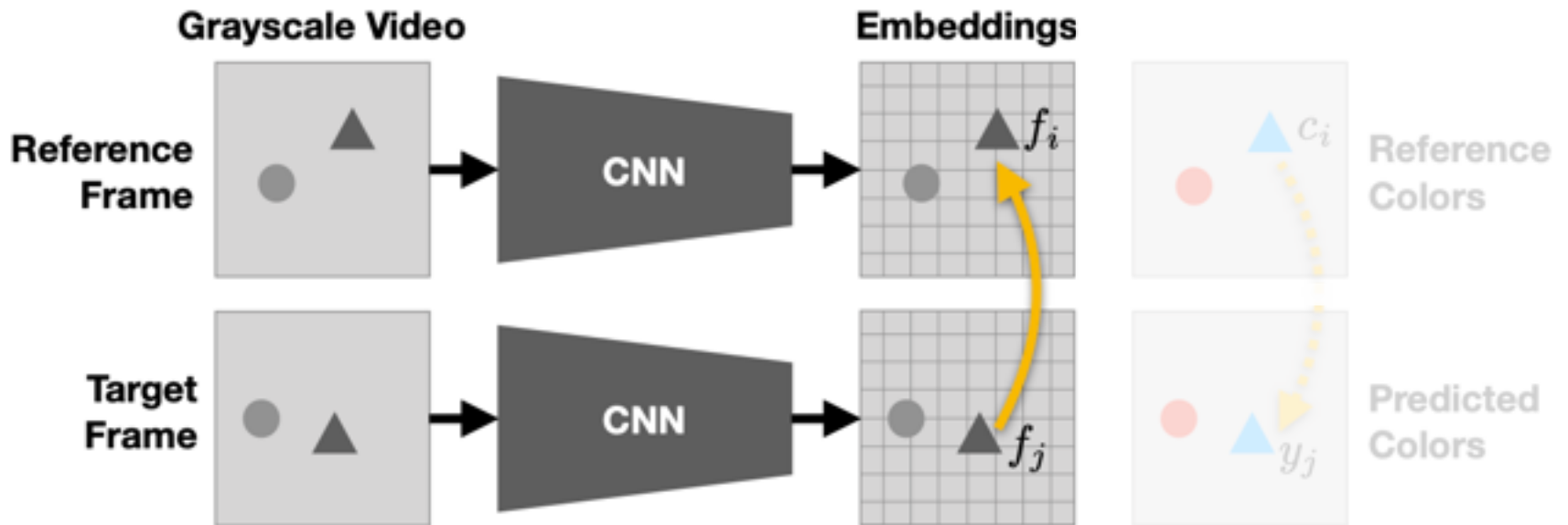
Technical Approach

- The model learns to copy the colors from the reference frame into the target frame.



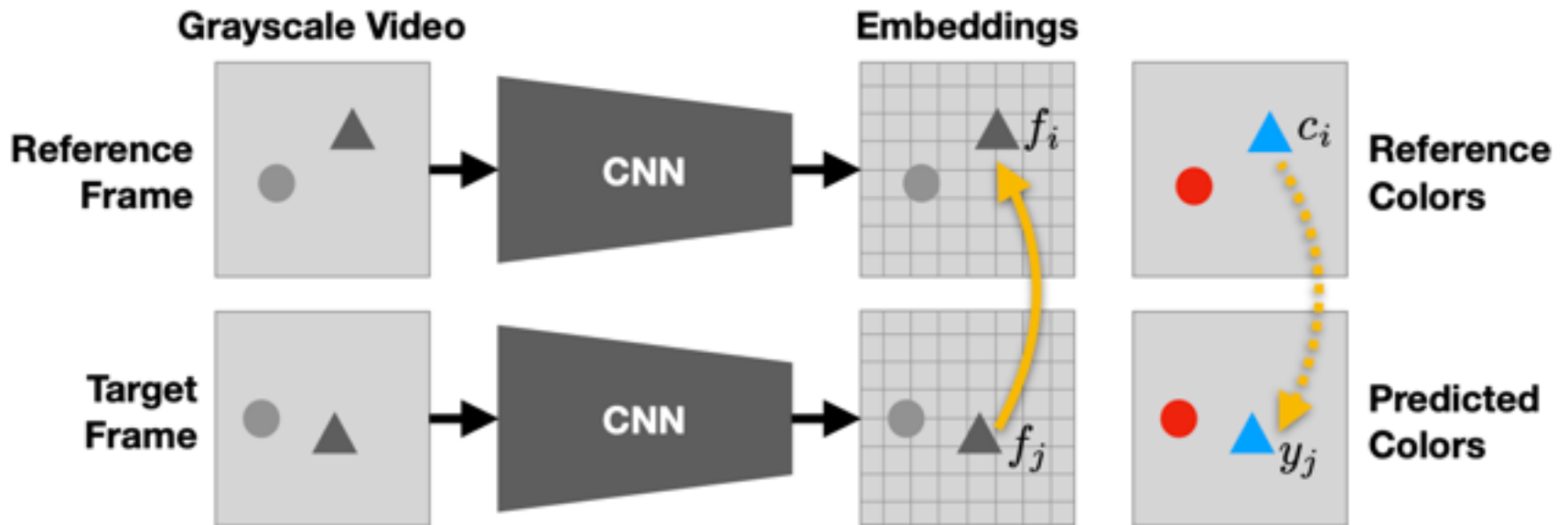
Technical Approach

- The model learns to copy the colors from the reference frame into the target frame.



Technical Approach

- The model learns to copy the colors from the reference frame into the target frame.



Learning to Point


- How is the pointer mechanism for finding correspondences across two frames implemented?

$$y_j = \sum_i A_{ij} c_i$$

Learning to Point

- How is the pointer mechanism for finding correspondences across two frames implemented?

$$y_j = \sum_i A_{ij} c_i$$



3 dimensional RGB color prediction for pixel j in the target frame.

Learning to Point

- How is the pointer mechanism for finding correspondences across two frames implemented?

$$y_j = \sum_i A_{ij} c_i$$

**Probabilistic similarity
between pixels j and i in the
reference and target frames.**

Learning to Point

- How is the pointer mechanism for finding correspondences across two frames implemented?

$$y_j = \sum_i A_{ij} c_i$$



3 dimensional RGB color value for pixel i in the reference frame.

Pixel-level Similarity

- As is standard similarity between pixels in reference and target frames is computed using normalized dot product.

Feature vector for pixel i in the reference frame.

Feature vector for pixel j in the target frame.

$$A_{ij} = \frac{\exp(f_i^T f_j)}{\sum_k \exp(f_k^T f_j)}$$

Loss Function

- The authors quantize the raw RGB color values across the dataset via k-means clustering (using 16 clusters).
- Afterward, standard cross-entropy loss is used to optimize the network.

$$\min_{\theta} \sum_j \mathcal{L}(y_j, c_j)$$

Experiments

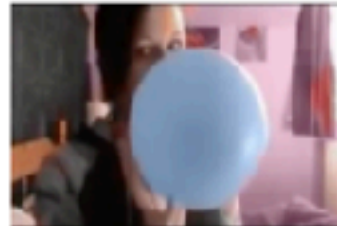
Reference Frame



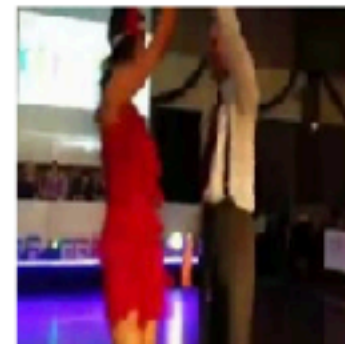
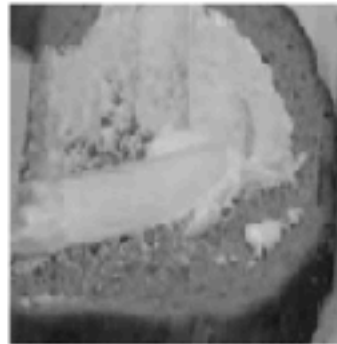
Future Frame (gray)



Predicted Color



True Color



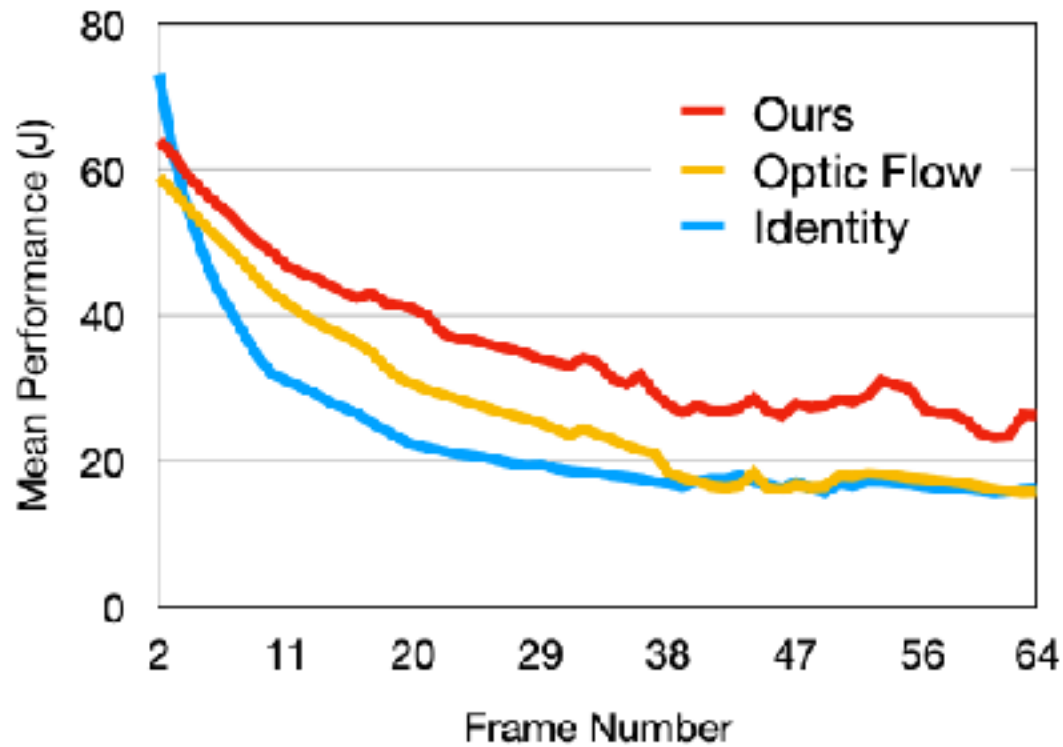
Video Object Segmentation

- Video object segmentation Results on the DAVIS 2017 validation set.

Method	Supervised?	Segment Boundary	
Identity		22.1	23.6
Single Image Colorization		4.7	5.2
Optical Flow (Coarse-to-Fine) [59]		13.0	15.1
Optical Flow (FlowNet2) [23]		26.7	25.2
Ours		34.6	32.7
Fully Supervised [46, 47]	✓	55.1	62.1

Video Object Segmentation

- Video object segmentation average performance versus time in the video.

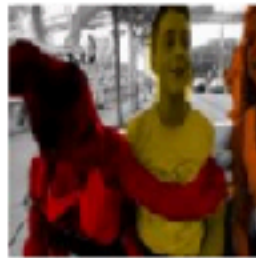
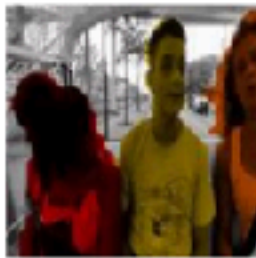


Video Object Segmentation

Inputs



Predicted Segmentations

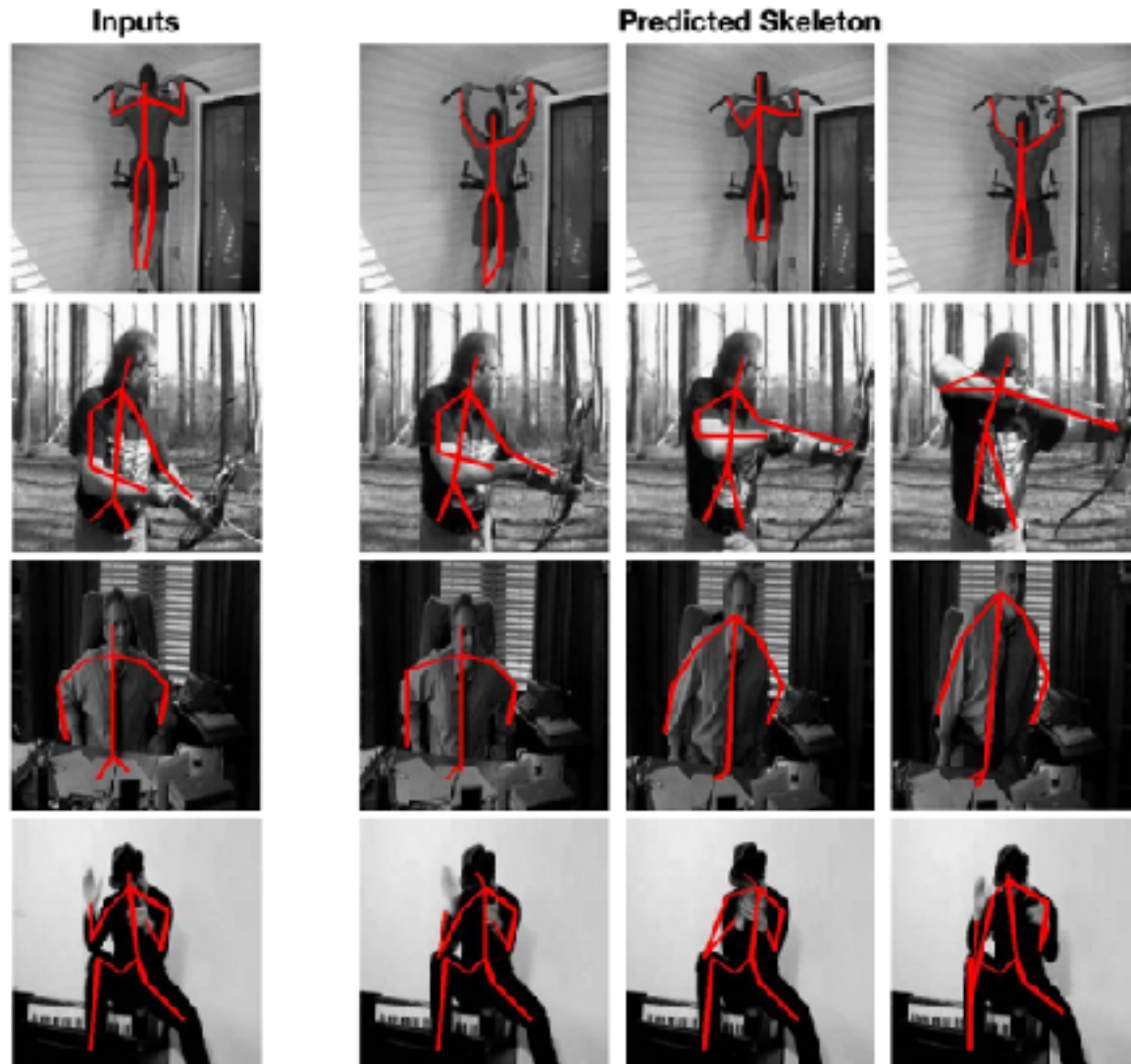


Human Pose Tracking

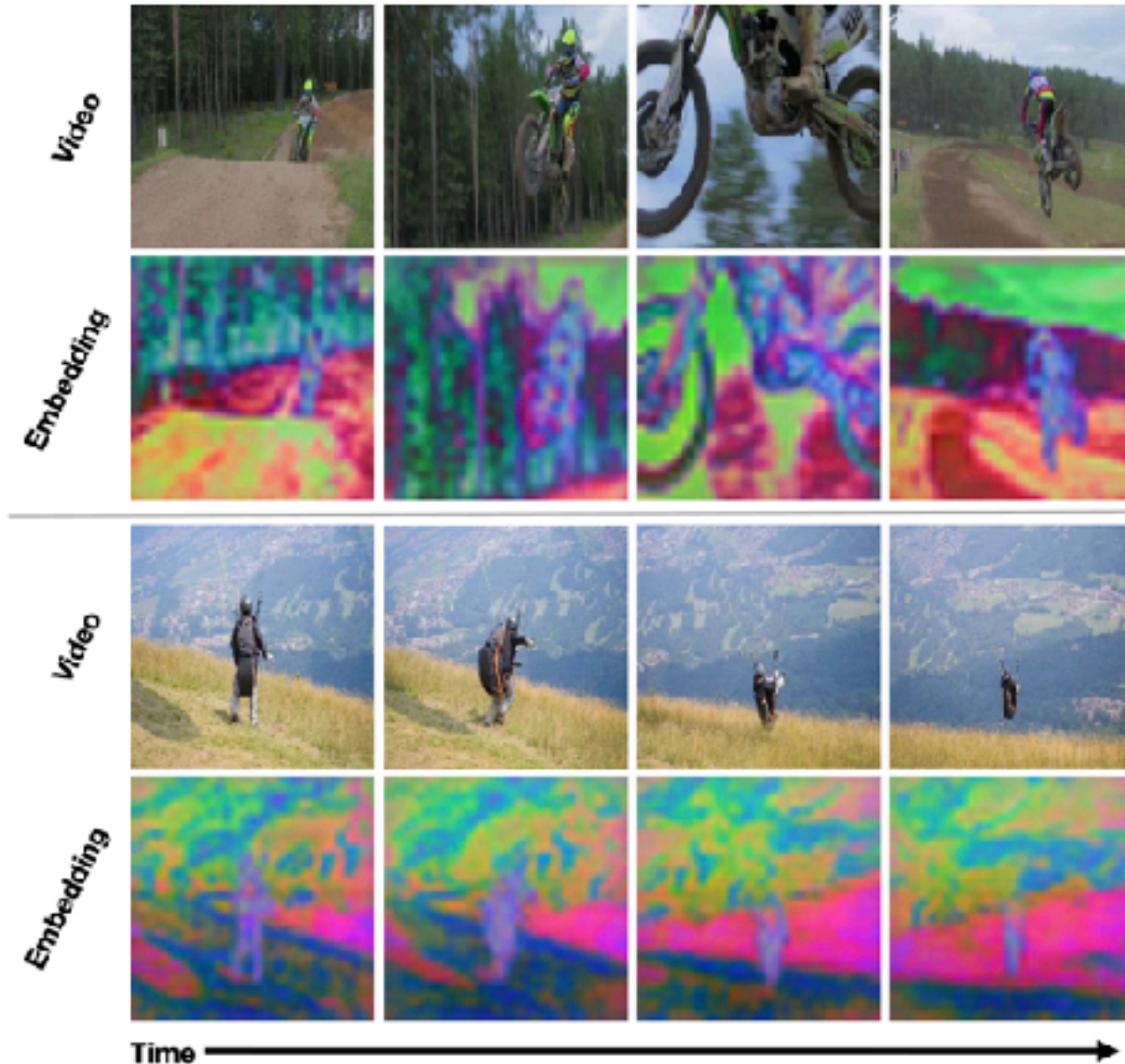
- Evaluated on the JHMDB (a subset of HMDB) validation set.
- A small scale dataset consisting of <10K videos of humans performing various actions.

Method	PCK@.1	PCK@.2	PCK@.3	PCK@.4	PCK@.5
Identity	43.1	64.5	76.0	83.5	88.5
Optical Flow (FlowNet2) [23]	45.2	62.9	73.5	80.6	85.5
Ours	45.2	69.6	80.8	87.5	91.4

Human Pose Tracking



Learned Embeddings



Contributions

- A clever colorization-based learning objective for learning a good representation to track objects.
- A very simple, yet effective technical approach.
- Convincingly outperforms simple baselines for this problem.

Discussion Questions

- What happens if two objects in the video have the same color?

Discussion Questions

- What happens if two objects in the video have the same color?
- Why is color discretization needed during the loss function computation?

Discussion Questions

- What happens if two objects in the video have the same color?
- Why is color discretization needed during the loss function computation?
- Supervised vs self-supervised?