universität freiburg

MAX PLANCK INSTITUTE FOR INFORMATICS

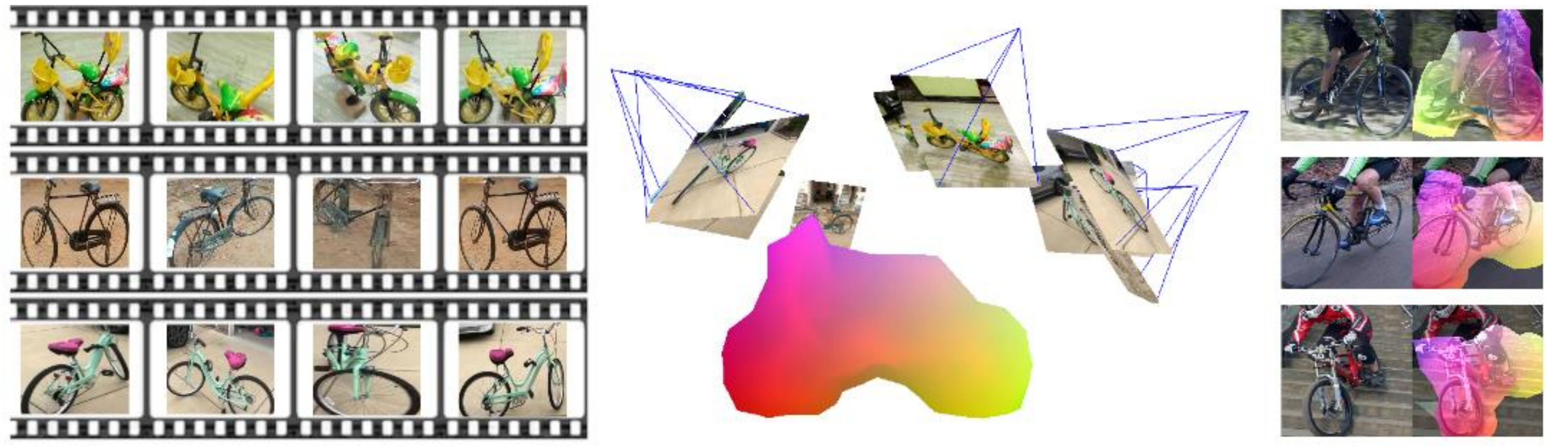
Unsupervised Learning of Category-Level 3D Pose from Object-Centric Videos

Leonhard Sommer[1], Artur Jesslen[1], Eddy Ilg[2], Adam Kortylewski[1,3]
[1]University of Freiburg [2]Saarland University [3]Max Planck Institute for Informatics

CVPR SEATTLE, WA JUNE 17-21, 2024

Motivation

- **Goal: Unsupervised Learning of Category-Level 3D Pose from Object-Centric Videos**



Unaligned Object-Centric Videos

Self-Supervised Alignment & Feature Learning

In-the-Wild 3D Pose Estimation

- **Challenges**

(1) Structure-from-Motion aligns only camera poses in one video.

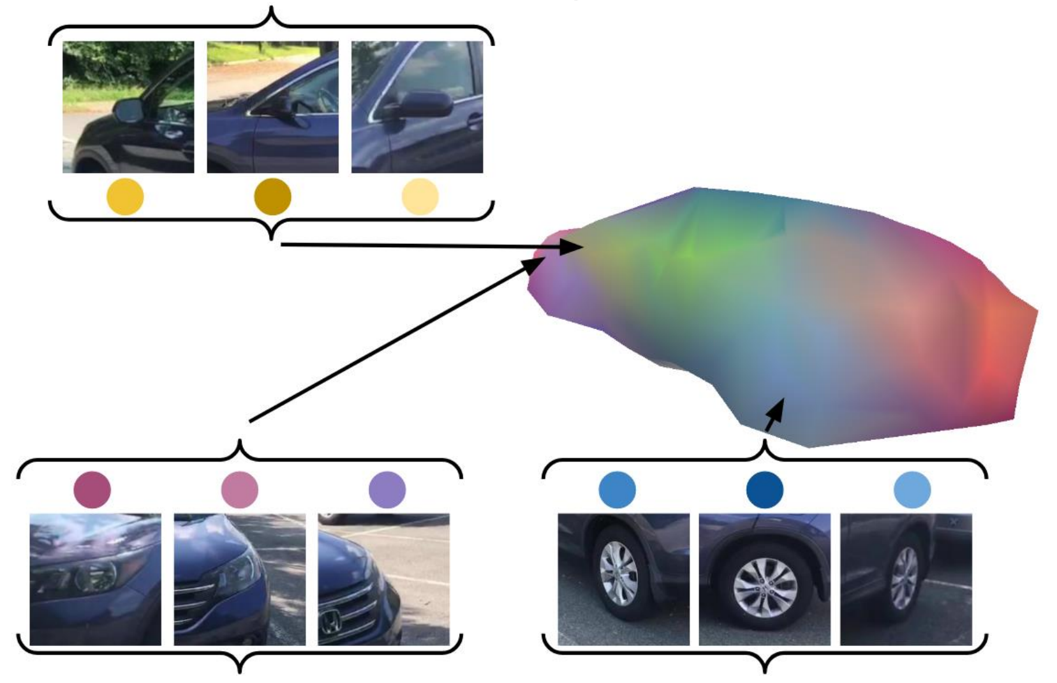
→ Novel alignment across videos via self-supervised surface features.

(2) Domain gap between object-centric videos and in-the-wild images.

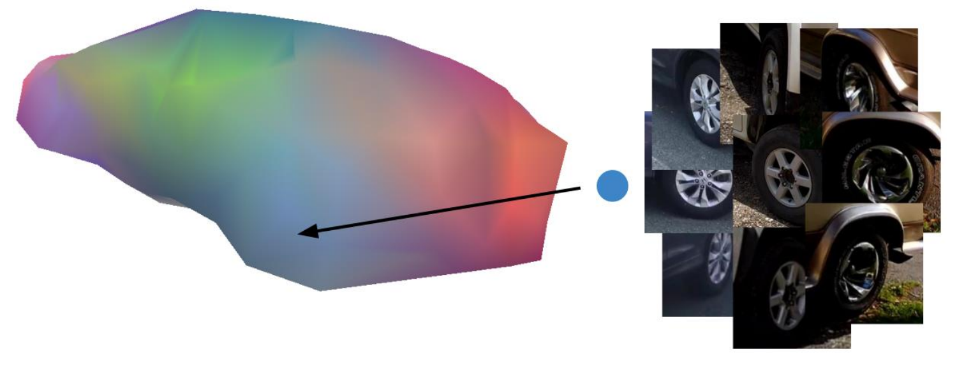
→ Learning category-level surface features and impose compositionality.

Our Representations: Surface Features

(1) From Single Video



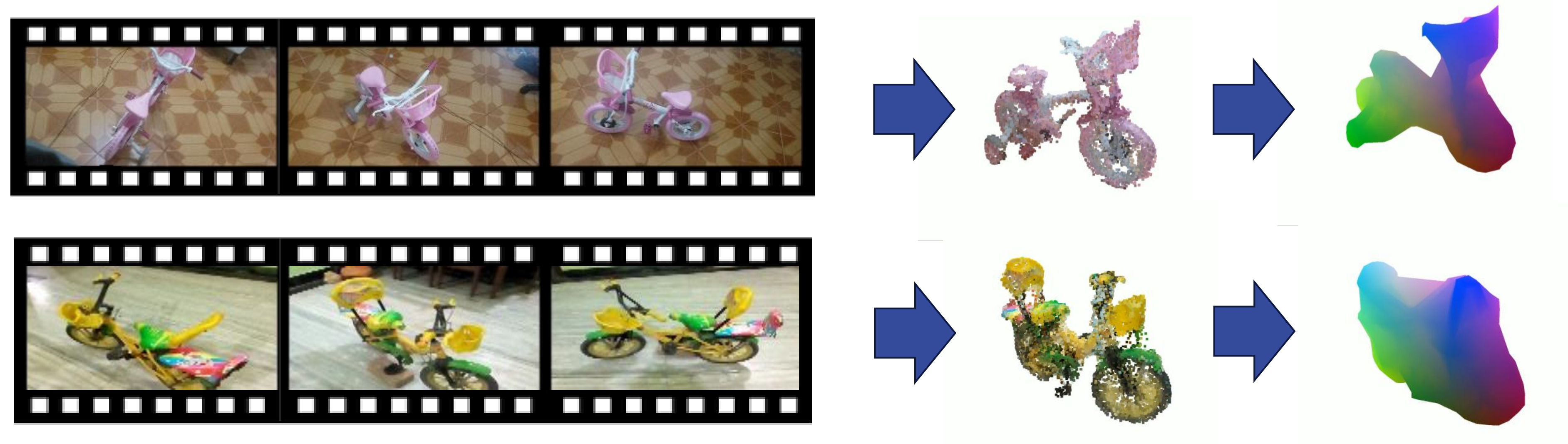
(2) Category-Level from Many Videos



- Viewpoint-dependent features (DINO).
- Viewpoint-invariant features facilitate 3D pose estimation.

(1) Method: Video Alignment

(1) Obtain surface features representation per video $S = \{V, F\}$.



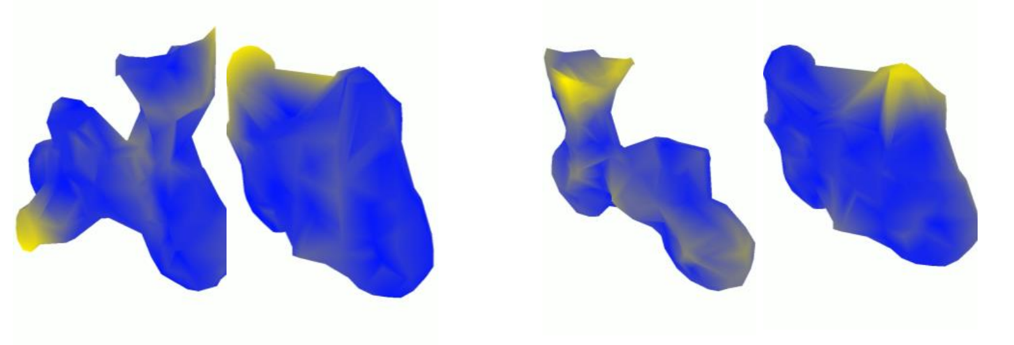
(2) Divide many-to-one alignment into many one-to-one alignments.

(3) Minimize one-to-one weighted distance.

$$\min_{\mathrm{T}} \mathcal{D}(S, \bar{S}, \mathrm{T}) = (1-\alpha)\mathcal{D}_{\text{geo}}(S, \bar{S}, \tau) + \alpha \mathcal{D}_{\text{app}}(S, \bar{S}, \tau)$$

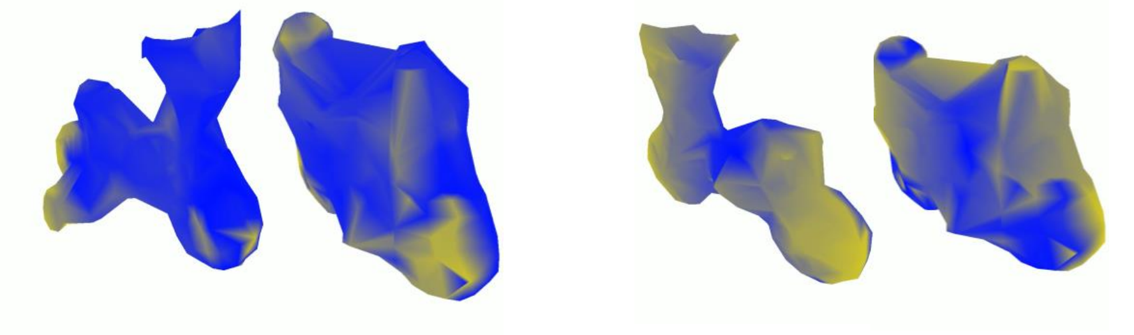
Geometric Distance

- Weighted Chamfer Distance.
- $\mathcal{D}_{\text{geo}}(S, \bar{S}, \tau) = \sum_{\mathrm{v}_i \in \mathrm{V} \cup \bar{\mathrm{V}}} \sigma(i, \tau)\, ||\mathrm{v}_i - \mathrm{v}_{\chi(\mathrm{v}_i)}||_2$



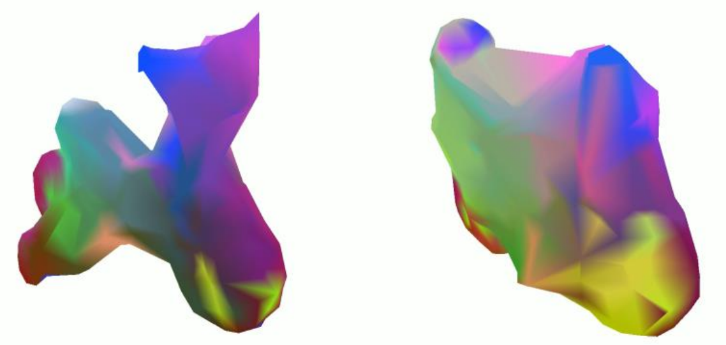
Appearance Distance

- Weighted appearance correspondences distance.
- $\mathcal{D}_{\text{app}}(S, \bar{S}, \tau) = \sum_{\mathrm{v}_i \in \mathrm{V} \cup \bar{\mathrm{V}}} \sigma(i, \tau)\, ||\mathrm{v}_i - \mathrm{v}_{\psi(\mathrm{v}_i)}||_2$



Appearance Correspondences

- $\psi(\mathrm{v}_i, \mathrm{f}_i) = \begin{cases} \operatorname{argmin}_{j \in 1 \ldots |\bar{\mathrm{V}}|} \min_{k,l} ||\mathrm{f}_j^k - \mathrm{f}_i^l||, & \text{for } \mathrm{v}_i \in \mathrm{V} \\ \operatorname{argmin}_{j \in 1 \ldots |\mathrm{V}|} \min_{k,l} ||\mathrm{f}_j^k - \mathrm{f}_i^l||, & \text{for } \mathrm{v}_i \in \bar{\mathrm{V}} \end{cases}$

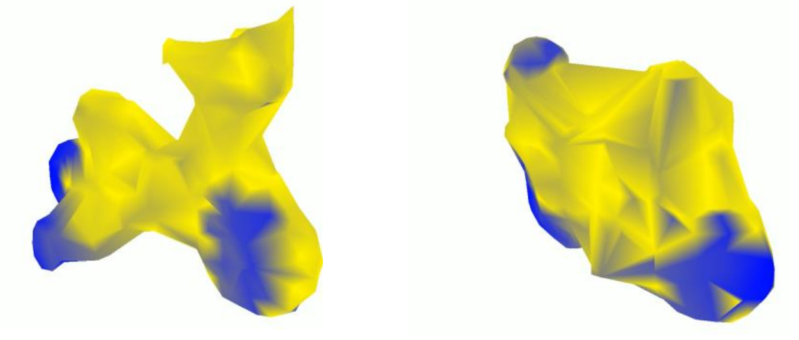


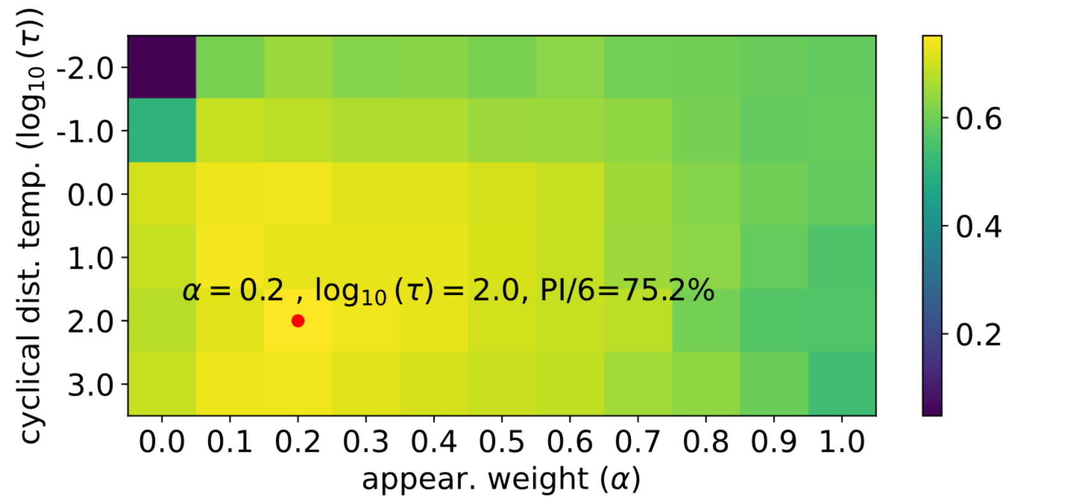
Optimum

- Geometric correspondences weighted four times as much as the appearance correspondences.
- Only weak cycle-distance weighting.

Cycle-Distance Weighting

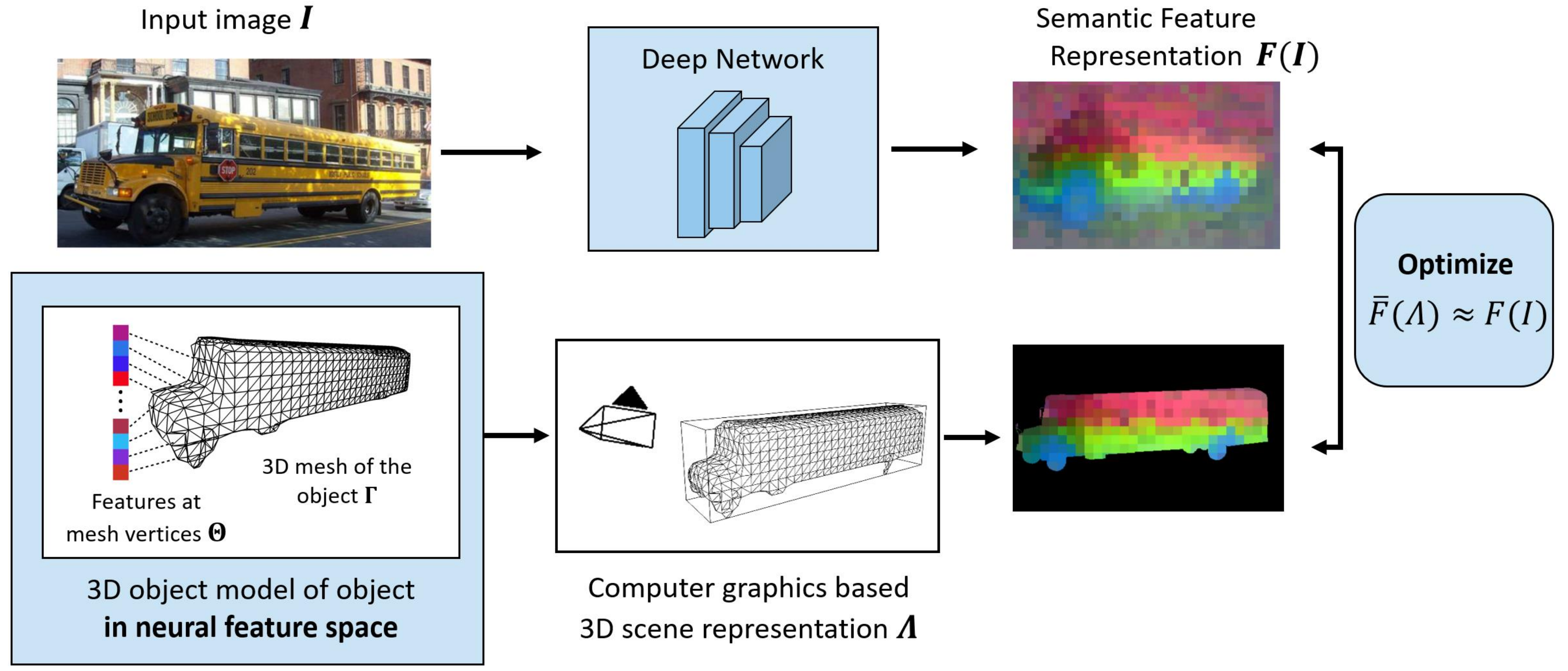
- $\mathrm{d}_{\text{cycle}}(\mathrm{v}_i, \mathrm{f}_i) = ||\mathrm{v}_i - \mathrm{v}_{\psi(\mathrm{v}_j, \mathrm{f}_j)}||_2$, with $j = \psi(\mathrm{v}_i, \mathrm{f}_i)$
- $\sigma(i, \tau) = \texttt{Softmax}\left(-\frac{\mathrm{d}_{\text{cycle}}(\mathrm{v}_i, \mathrm{f}_i)}{\tau \max_{\mathrm{v}_i, \mathrm{v}_j \in \mathrm{V}} ||\mathrm{v}_i - \mathrm{v}_j||_2}\right)$





cyclical dist. temp. ($\log_{10}(\tau)$)
-2.0
-1.0
0.0
1.0
2.0
3.0
$\alpha = 0.2$, $\log_{10}(\tau) = 2.0$, PI/6=75.2%
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
appear. weight (α)
0.6
0.4
0.2

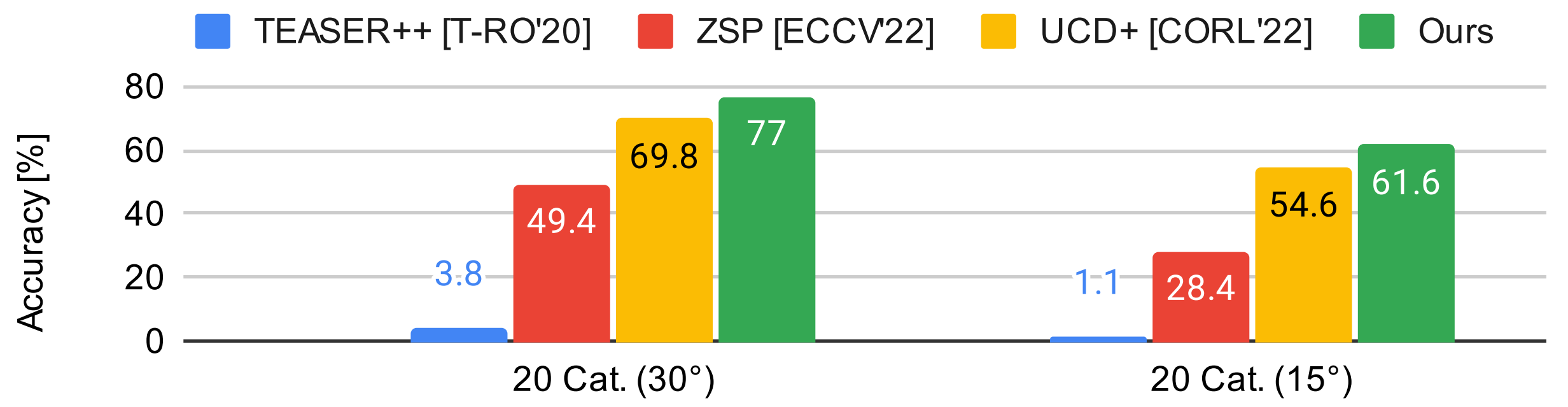
(2) Method: Learning Category-Level Surface Features



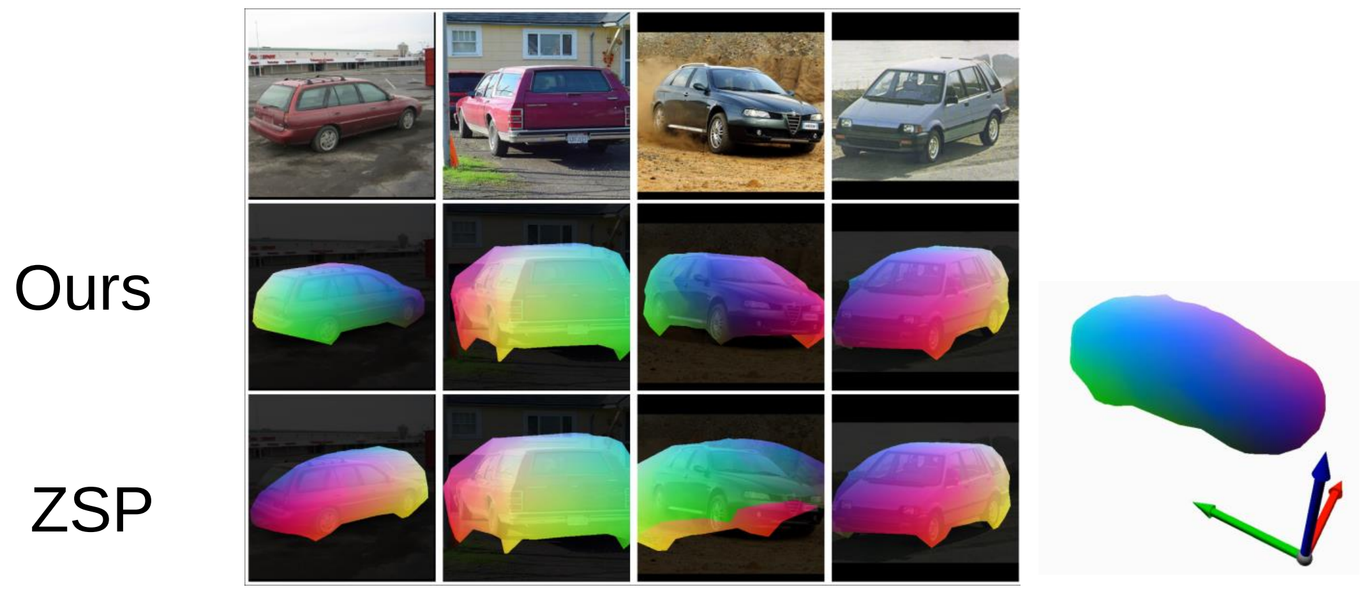
Input image I
Deep Network
Semantic Feature Representation $F(I)$
Optimize
$\bar{F}(\Lambda) \approx F(I)$
Features at mesh vertices Θ
3D mesh of the object Γ
3D object model of object in neural feature space
Computer graphics based 3D scene representation Λ

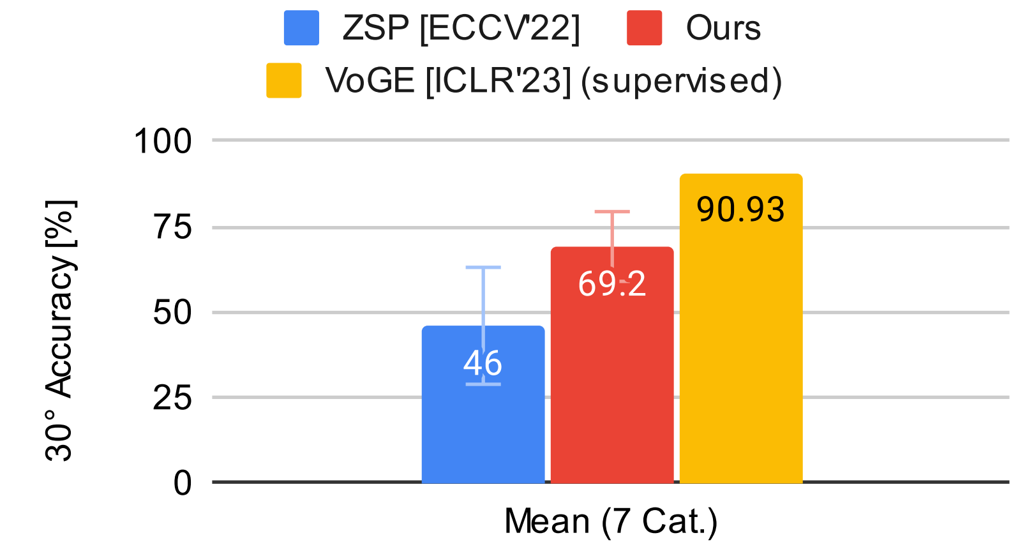
Results

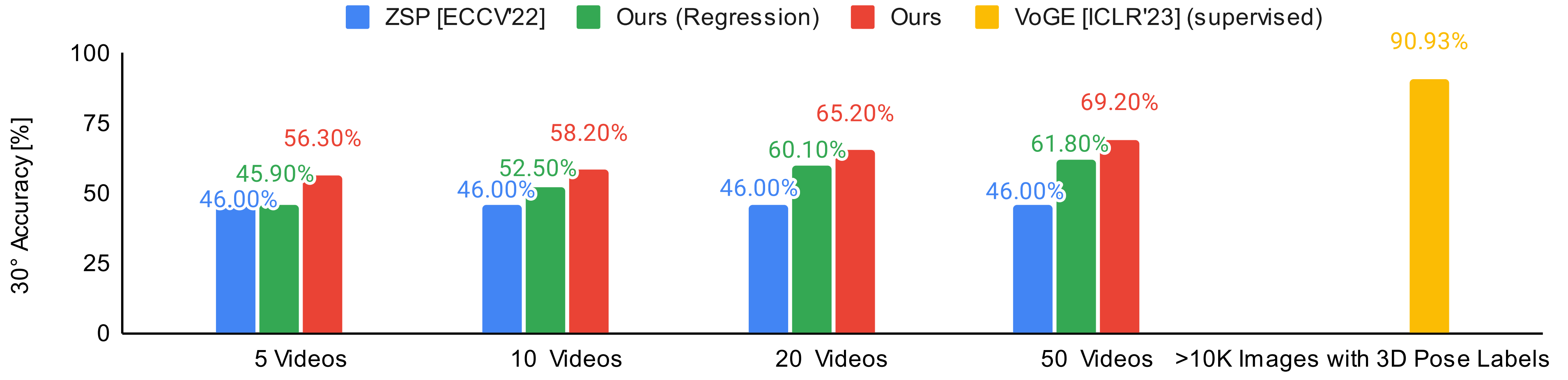
Video Alignment – CO3D



In-the-Wild 3D Pose Estimation – PASCAL3D+







Acknowledgements Adam Kortylewski acknowledges support for his Emmy Noether Research Group funded by the German Research Foundation (DFG) under Grant No. 468670075.