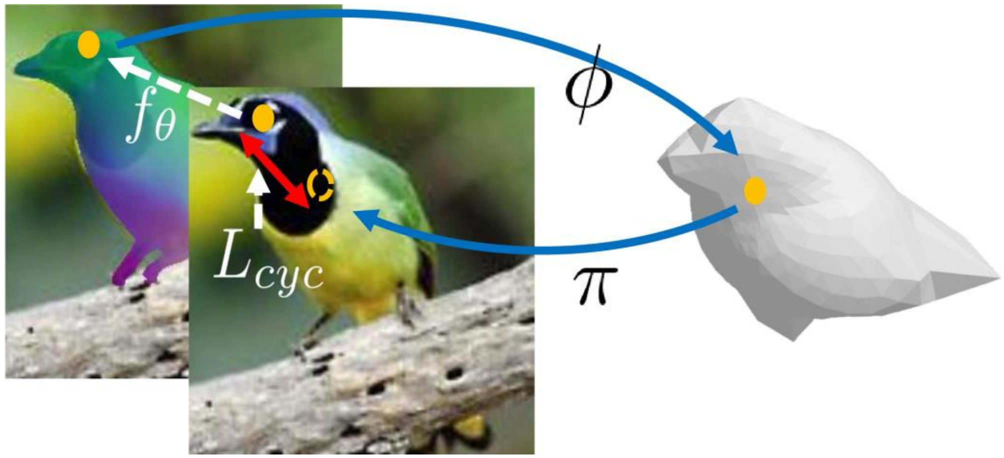
To The Point: Correspondence-driven monocular 3D category reconstruction

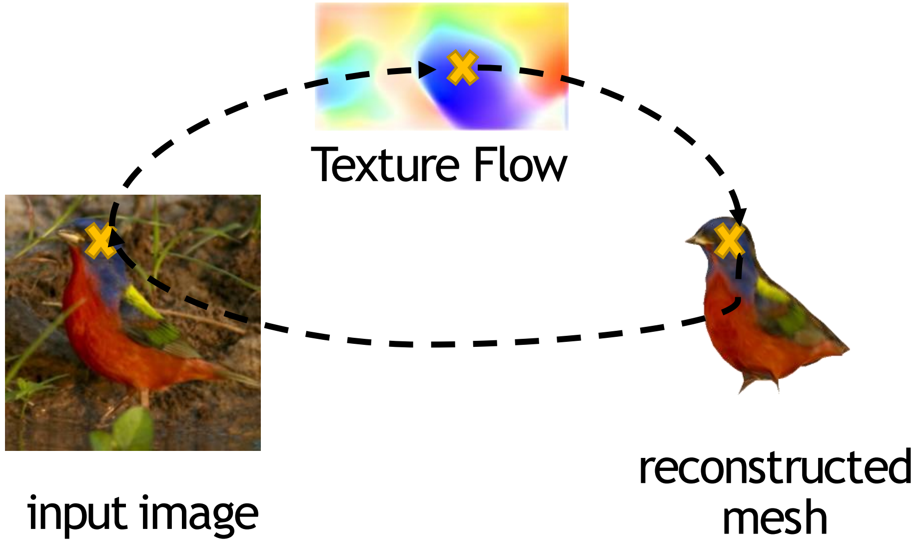
Filippos Kokkinos
Ph.D. student @ UCL

Motivation

Geometric Correspondence



Appearance Correspondence



Correspondences for 3D reconstruction:

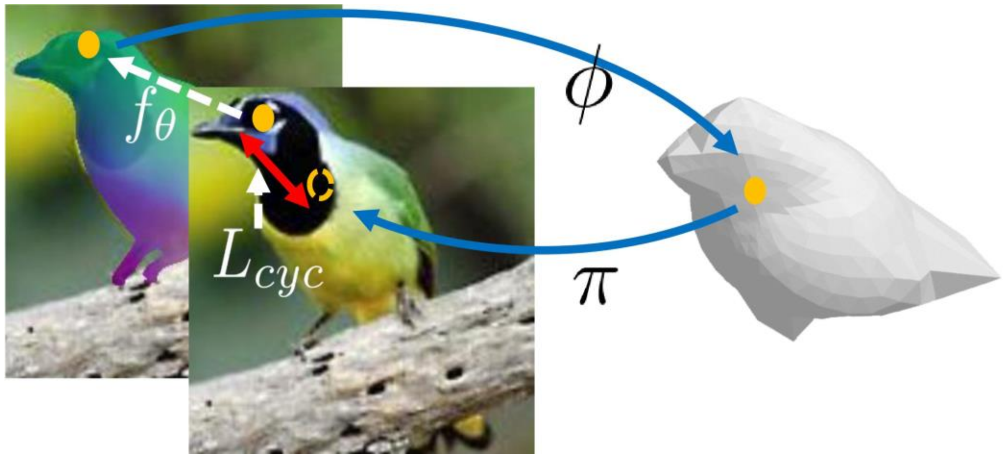
- Canonical Surface Mapping via Geometric Cycle Consistency, Kulkarni, N., Gupta, A., & Tulsiani, S. ICCV (2019)
- ACSM: Articulation-Aware Canonical Surface Mapping, N. Kulkarni, A. Gupta, D. F. Fouhey, S. Tulsiani, CVPR (2020)

Correspondences for texture flow:

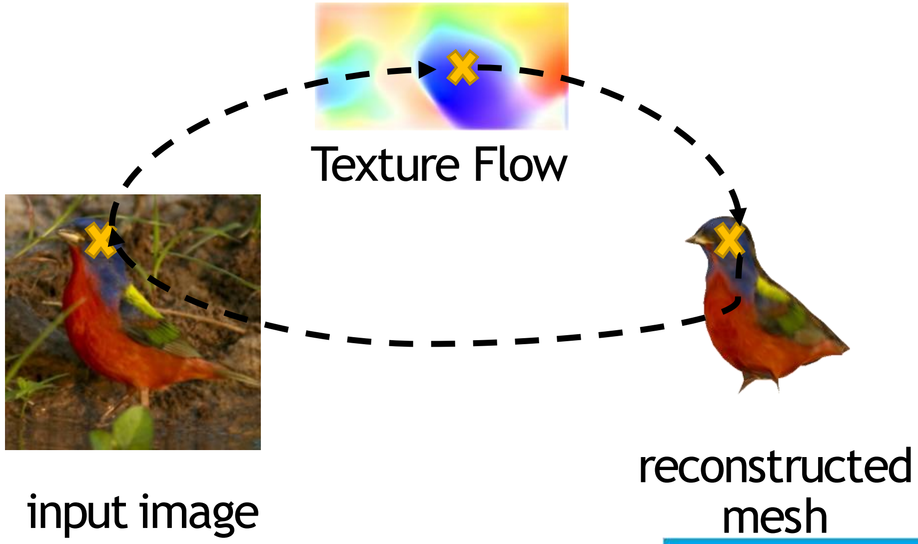
- Learning category-specific mesh reconstruction from image collections, Kanazawa, A., Tulsiani, S., Efros, A. A., & Malik, J. ECCV (2018)
- Self-supervised single-view 3d reconstruction via semantic consistency. Li, X., Liu, S., Kim, K., De Mello, S., Jampani, V., Yang, M. H., & Kautz, J. ECCV (2020)
- Online adaptation for consistent mesh reconstruction in the wild. Li, X., Liu, S., De Mello, S., Kim, K., Wang, X., Yang, M. H., & Kautz, J. NeurIPS (2020)

Shortcomings

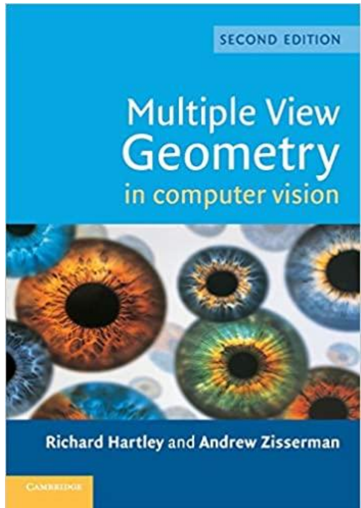
Geometric Correspondence



Appearance Correspondence



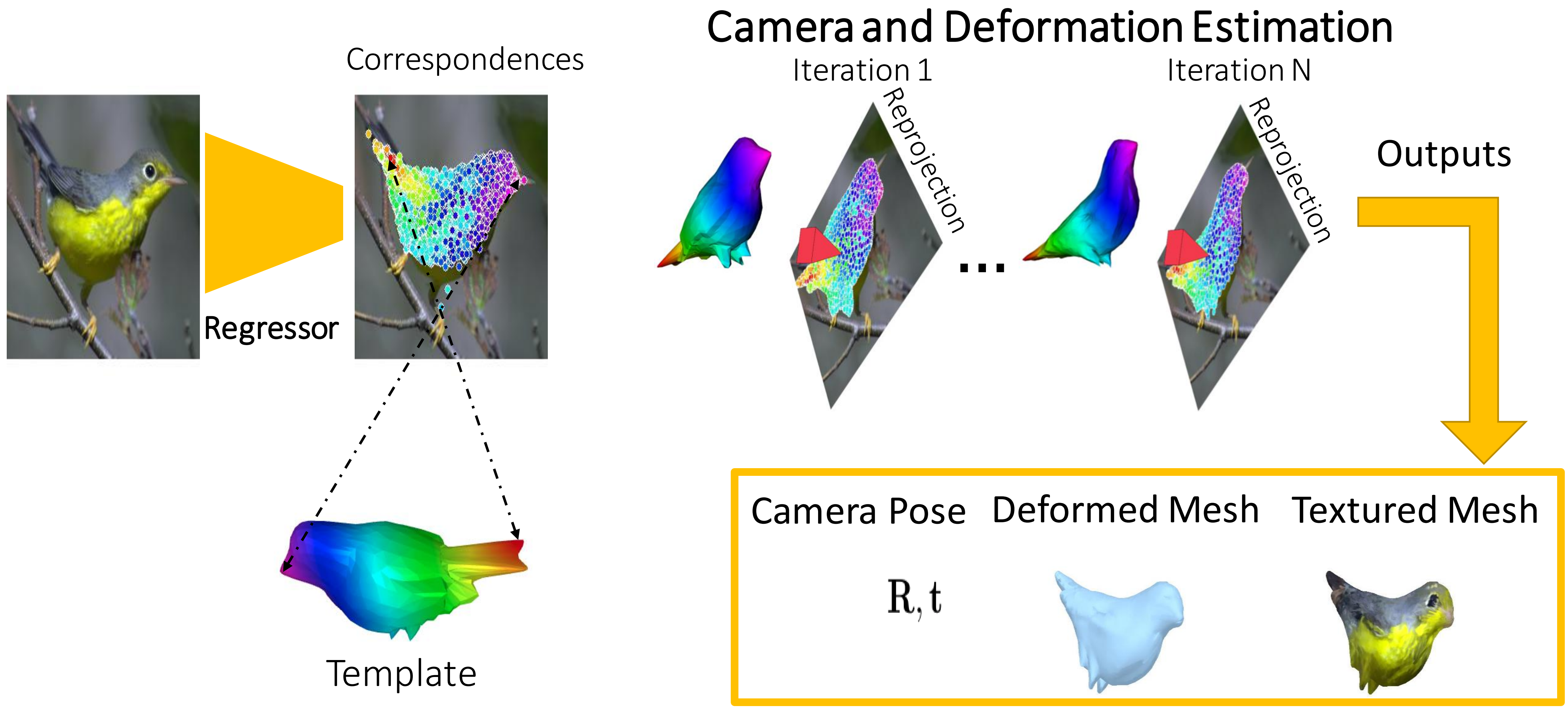
- Camera pose prediction is always a problem especially in self-supervised setups
- All non-keypoint based works use multiple hypothesis
- Given the correspondences, we can estimate pose using optimization



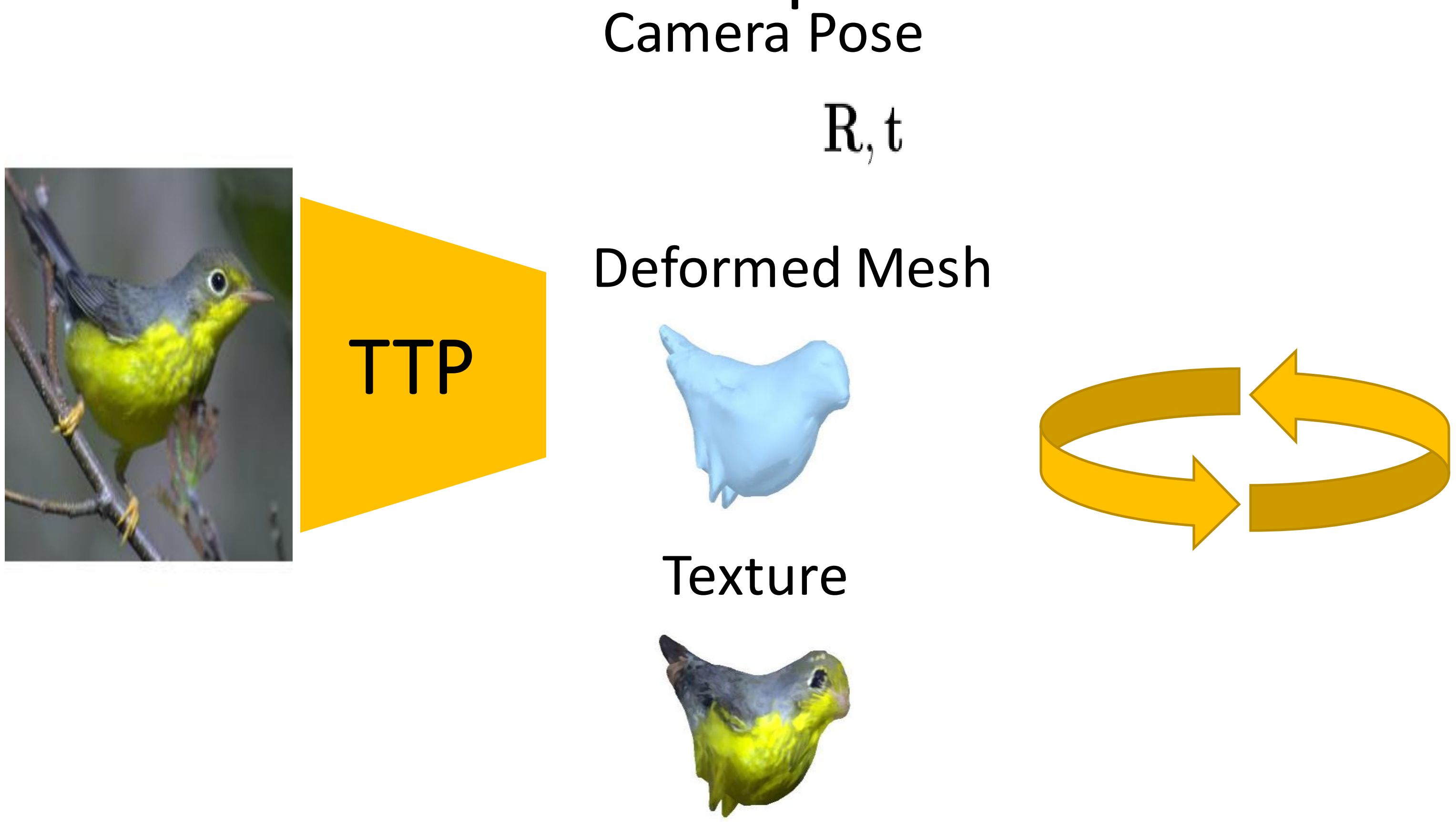
Contributions

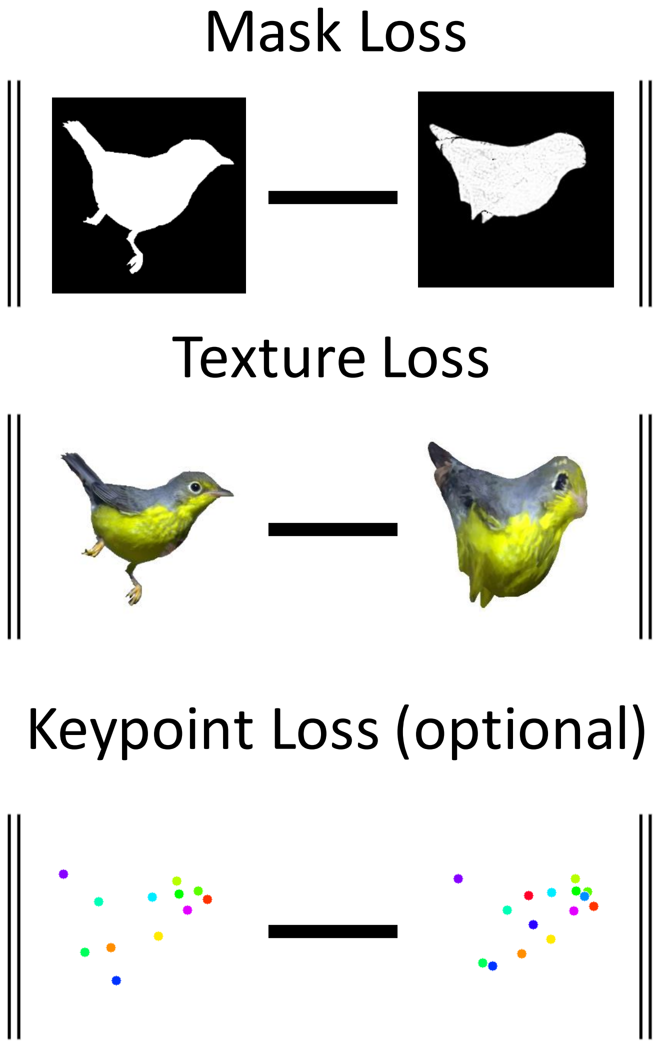
- Exploit correspondences to achieve 3D reconstruction
  - Regress only the correspondences using neural nets
  - Camera estimation via minimization of re-projection error
  - Deformation estimation via minimization of re-projection error
- Avoid:
  - Multiple hypothesis for camera pose and deformation
  - Separate branches and deep networks

Correspondence-driven self-supervision



Losses for self-supervision





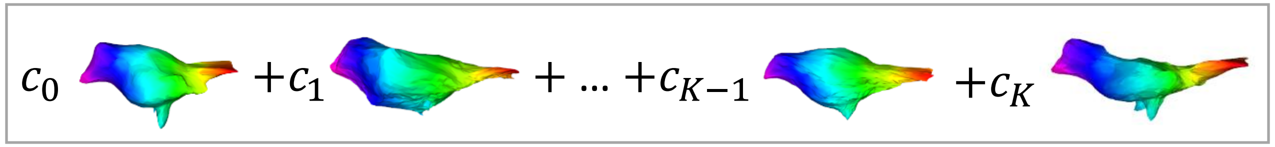
- As-Rigid-as-possible Regularization on Shape
- Equivariance Constraint

3D Mesh Deformation

- Mesh Deformation using a linear learnable Basis $\mathbf{B}$: $\mathbf{V} = \mathbf{T} + \mathbf{Bc}$

Template $\mathbf{T}$

Learned Basis $\mathbf{B}$



- Basis $\mathbf{B}$ treated as network parameters
    - Learned with self-supervision over the dataset

Estimation of Pose and Deformation

- Mesh Deformation using a linear learnable Basis $\mathbf{B}$: $\quad \mathbf{V} = \mathbf{T} + \mathbf{B}\mathbf{c}$

- Camera is a scaled Orthographic projection containing $\mathbf{R}$, $\mathbf{t}$ and scale s

$$\hat{\mathbf{u}} = \pi(\mathbf{V}) = \mathbf{C}(\mathbf{R}\mathbf{V} + \mathbf{t}) \qquad \mathbf{C} = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \end{bmatrix}$$

Per Instance Optimization

- Mesh Deformation using a linear learnable Basis $\mathbf{B}$: $\quad \mathbf{V} = \mathbf{T} + \mathbf{Bc}$
- Camera is a scaled Orthographic projection containing $\mathbf{R}$, $\mathbf{t}$ and scale s

$$\hat{\mathbf{u}} = \pi(\mathbf{V}) = \mathbf{C}(\mathbf{RV} + \mathbf{t}) \qquad \mathbf{C} = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \end{bmatrix}$$

- Reprojection Cost:

$$l(\mathbf{R}, \mathbf{t}, \mathbf{c}) = \sum_{i=1}^{N} \underbrace{v_i}_{\text{visibility}} \|\underbrace{\mathbf{u}_i}_{\text{regressed points}} - \underbrace{\boxed{\hat{\mathbf{u}}_i(\mathbf{C}, \mathbf{R}, \mathbf{t})}}_{\text{projected points}}\|_2^2 + \underbrace{\boxed{\gamma\|\mathbf{c}\|_2^2}}_{\text{basis reg.}}$$

- Per instance parameters: $\mathbf{C}$, $\mathbf{R}$, $\mathbf{t}$ and $\mathbf{c}$

Optimization In the Loop

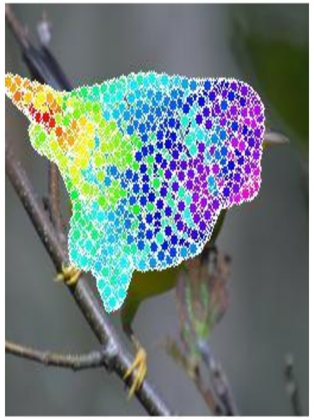
- Given the correspondences compute camera pose and deformation
  - Solved with Alternating Optimization using LBFGS

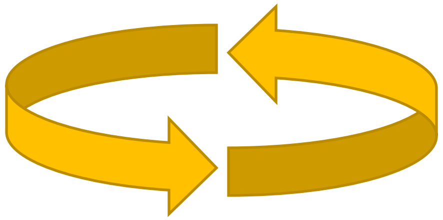
$l =$

Regressed Points

$-$

Initialization





$$l(\mathbf{R},\mathbf{t},\mathbf{c}) = \sum_{i=1}^{N} v_i \left\|\mathbf{u}_i - \hat{\mathbf{u}}_i(\mathbf{C},\mathbf{R},\mathbf{t})\right\|_2^2 + \gamma\|\mathbf{c}\|_2^2$$

<u>Camera step:</u>

$$\hat{\mathbf{R}}^t, \hat{\mathbf{t}}^t = \arg\min_{\mathbf{R},\mathbf{t}} l\left(\mathbf{R},\mathbf{t},\hat{\mathbf{c}}^t\right)$$

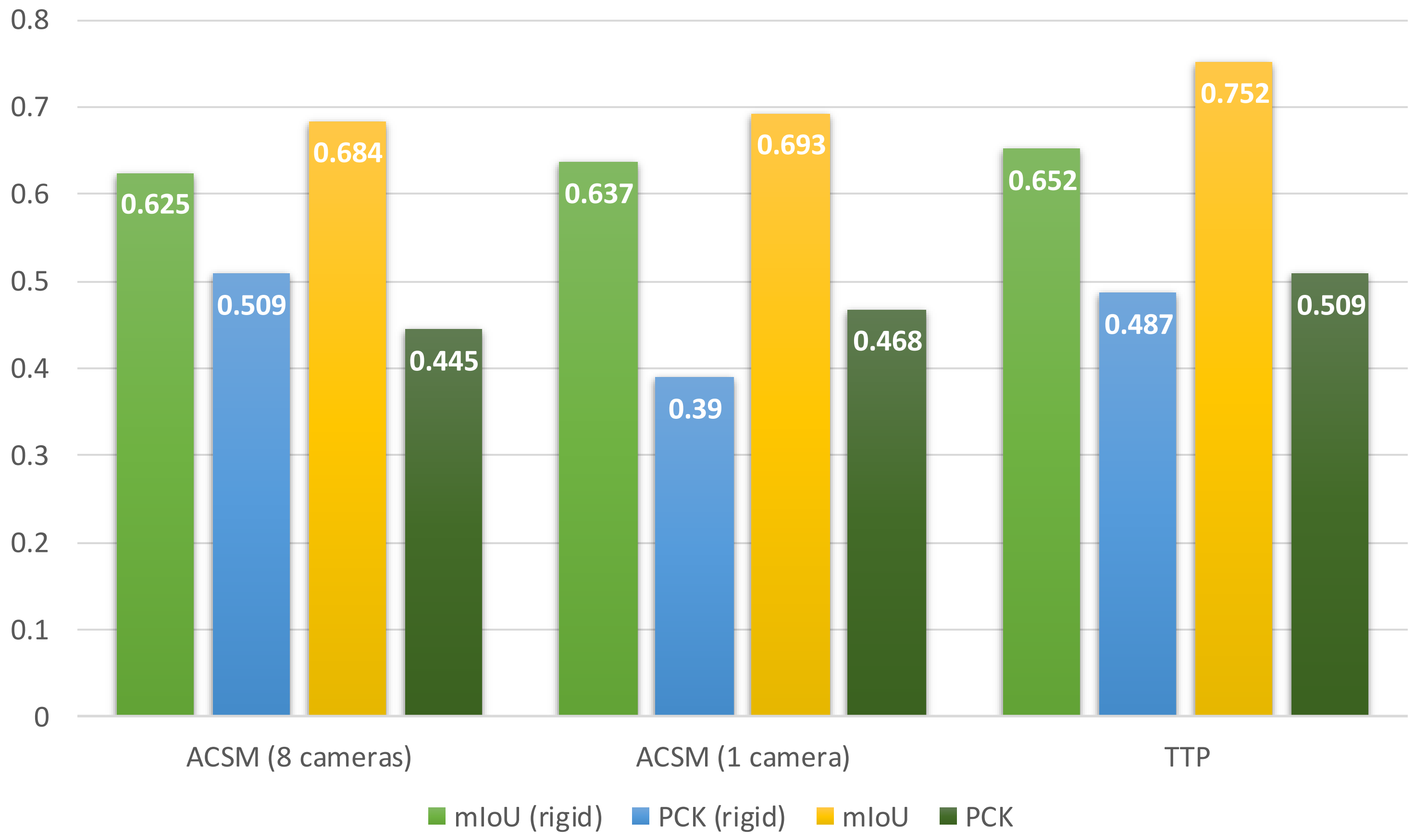
$$\text{subject to } \mathbf{R} \in SO(3)$$

<u>Deformation step:</u>

$$\hat{\mathbf{c}}^{t+1} = \arg\min_{\mathbf{c}} l\left(\hat{\mathbf{R}}^t, \hat{\mathbf{t}}^t, \mathbf{c}\right)$$

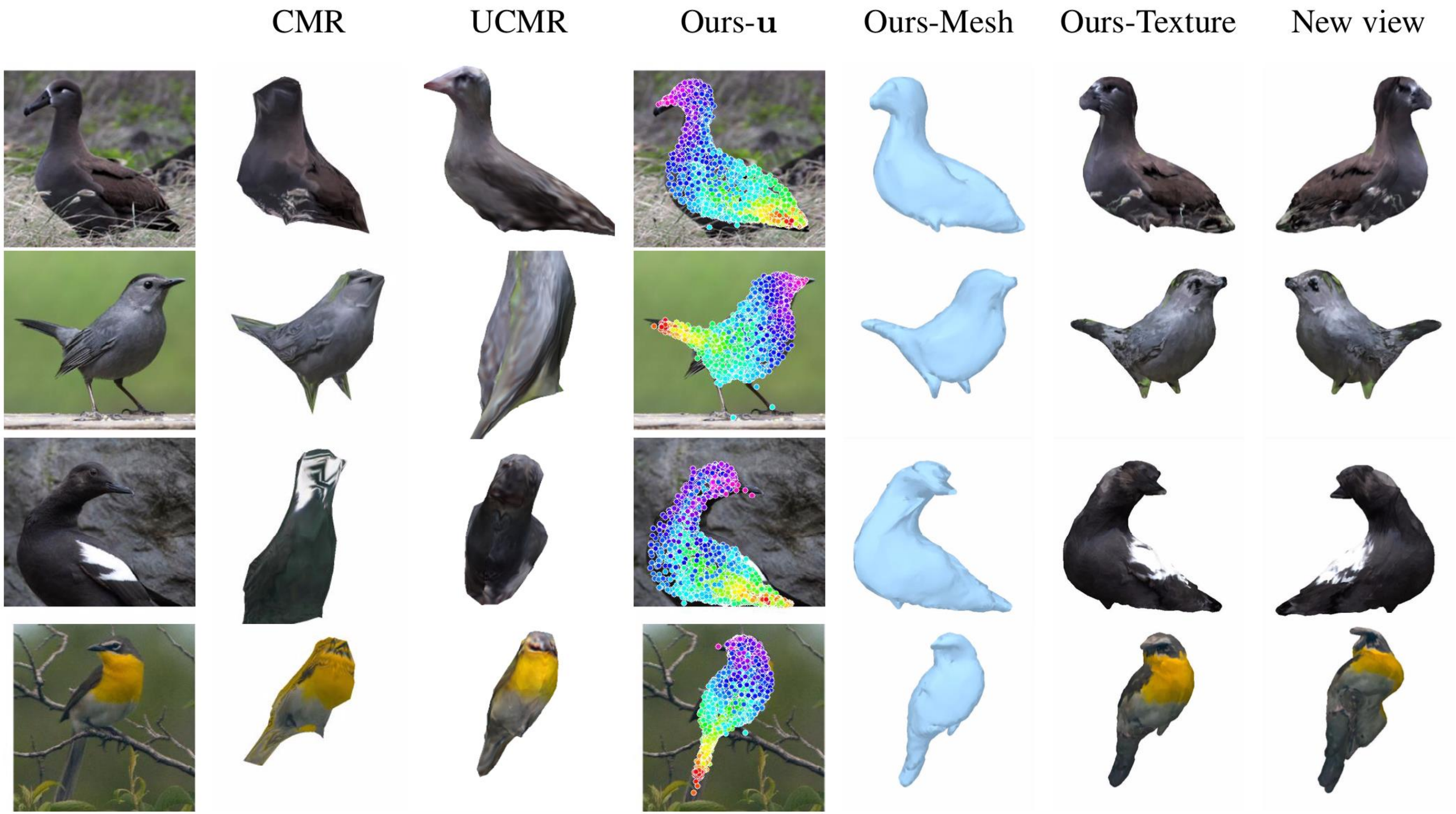
- Differentiation through the optimization layer is achieved with implicit function theorem

Evaluation – Bird Reconstruction

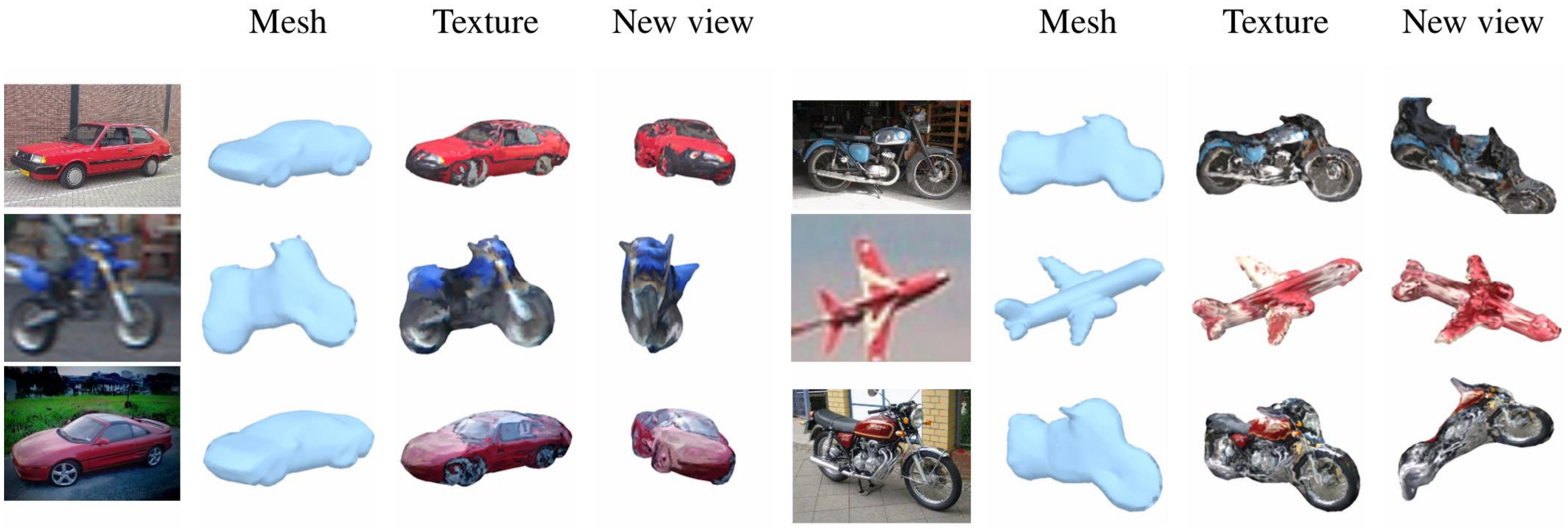


PCK accuracy scores are normalized for visualization purposes.

Qualitative Results



PASCAL 3D reconstructions



Thank you!

fkokkinos.github.io/to_the_point/