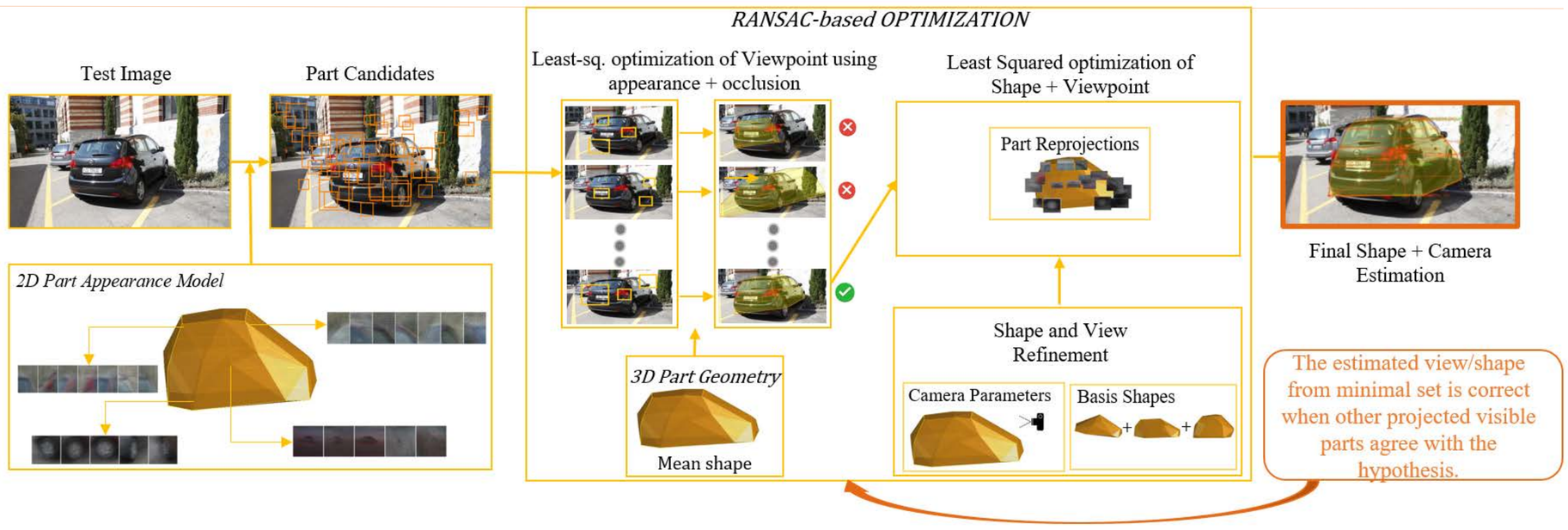
Class-Specific Object Pose Estimation and Reconstruction using 3D Part Geometry

Arun CS Kumar[1], András Bódis-Szomorú[2], Suchendra Bhandarkar[1], Mukta Prasad[3]
[1]University of Georgia, [2]ETH Zürich, [3]Trinity College Dublin

GOAL

Class-specific reconstruction, viewpoint estimation and detection

- Deformable linear subspace, fine-grained, part-based model
- learnt from real image sequences (*no CAD models*)
- Geometric occlusion - cognizant reasoning (analytic optimization of shape + view with RANSAC) instead of regression.



MODEL

- A Shape S=[$\mathbf{s}_1$ … $\mathbf{s}_P$] represents P semantic 3D parts on object shape instance, s.t.
  - $S = \sum_{l=0}^{L} \alpha_l B_l$ represents deformation in a linear subspace of L basis shapes learnt from training.
- The *2D* projection of a 3D part location $\mathbf{s}_p$ is given as:

$$\hat{x} = \rho(C \cdot S_p) = \rho\left(C \sum_{l=1}^{L} \alpha_l b_{lp}\right)$$

- The loss function for computing Shape S and camera matrix C from visible image information:

$$L(\{\alpha_l\}, C) = \sum_{p=1}^{P} v\left(s_p, C\right) \cdot \left\| x_p - \rho(C \cdot s_p) \right\|^2, \qquad s_p = \sum_{l=1}^{L} \alpha_l b_{lp},$$

where $v(\mathbf{s}_p, C) \in \{0, 1\}$ is binary visibility of $\mathbf{s}_p$ in camera C.

- This paper: Minimize L *w.r.t.* the shape coefficients $\{\alpha_l\}$ and projection parameters C, using occlusion-aware, analytic optimization in a RANSAC framework.

RELATED WORK

- [Pepik et al. ECCV 12] "3D2PM – 3D Deformable Parts Model"
- [Hejrati et al. CVPR 14] "Analysis by Synthesis"
- [Tulsiani et al. CVPR 15] "Viewpoints and Keypoints"
- [Zia et al. PAMI 13, CVPR 14]

Most methods learn from labour-intensive CAD models (using similar models for result rendering too), often resorting to prediction using regression, ConvNets or sampling-intensive inference. We show the power of a simpler, automatic image-data-driven approach, that interleaves analytic optimization of geometric + deformation + occlusion properties with a RANSAC scheme.

DATASET

- *RealCar dataset:* 35 image sequences (between 30 and 115 images per sequence), taken around unique and distinct instances of cars, with challenging viewing conditions.
- *SfM* (*123DCatch*) is used to estimate the 3D car shapes and full camera matrices .

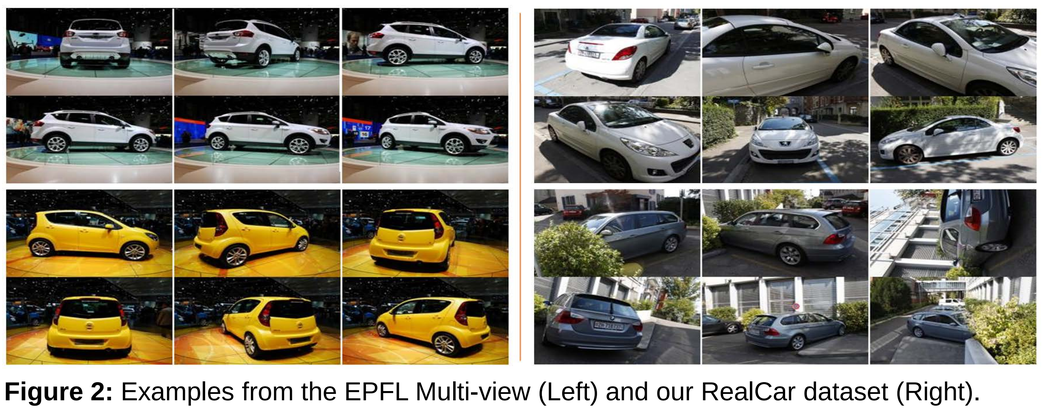


Figure 2: Examples from the EPFL Multi-view (Left) and our RealCar dataset (Right).

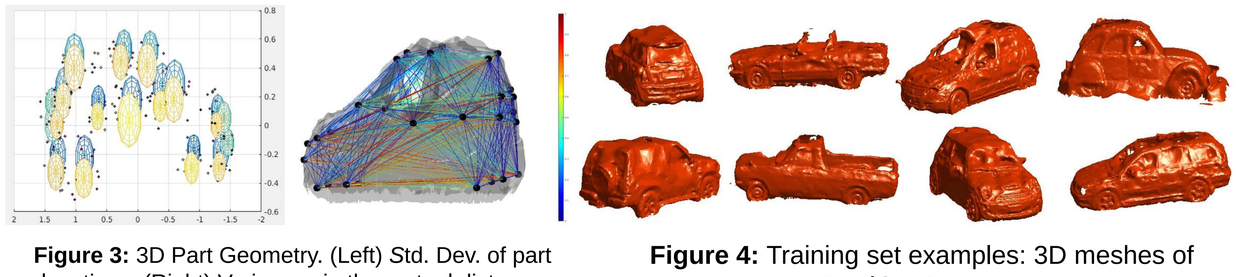


Figure 3: 3D Part Geometry. (Left) Std. Dev. of part locations. (Right) Variances in the mutual distance between each part pair overlaid on mean shape

Figure 4: Training set examples: 3D meshes of RealCar Dataset.

Algorithm 1: Shape recovery, pose estimation, detection

- *Part Detection:* SVM part appearance classifiers on *conv5 feature pyramids,* Non-Maxima Suppression + Platt's scaling.
- *Viewpoint Estimation:* Fit camera params for mean shape (|min set|=3, Algo. 2).
- *Viewpoint and Shape Refinement:* Repeat RANSAC-based routine by incrementally optimizing loss *w.r.t.* the 5 deformation bases and camera params (|min set| = 5).
- *Object Mask*: by projecting estimated deformable shape under estimated camera.

Algorithm 2: RANSAC-based function estimation

- Do *Part detection* to obtain possible candidates.
- **For** N iterations do
  - Randomly sample jointly visible, minimal set of unique parts.
  - Fit unknown parameters by minimizing projection loss.
  - Check for inliers within threshold τ_1
  - If number of inliers greater than threshold τ_2 then store them
- Re-estimate params by minimizing proj. loss for the set with maximum inliers.

EVALUATION

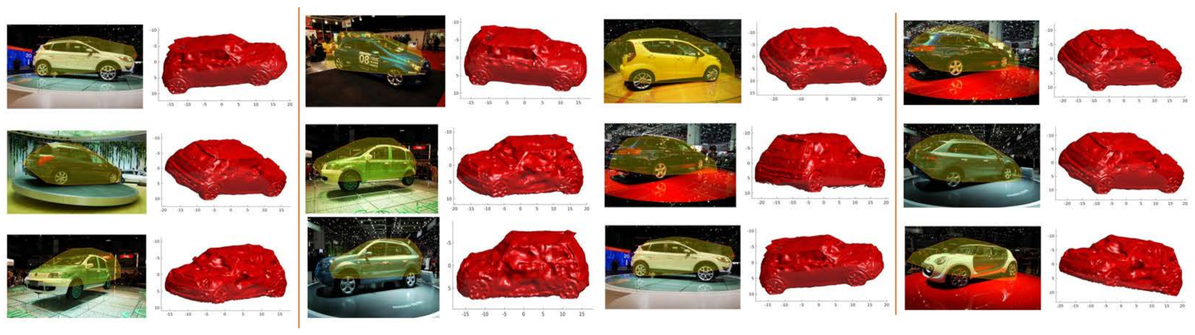


Figure 5: EPFL Cars : ***Odd columns***: Test image + Viewpoint/Shape estimations overlaid, ***Even columns*:** Viewpoint Estimate using a sample mesh. Quantitative results below.

ϑ	Our RealCar Dataset						EPFL Multi-view Cars Dataset						
	(Ours) Training Set			(Ours) Test Set			(Ours)			3D²PM[13]		Ozuysal et al. [28]	Glasner et al. [29]
	MPPE	MAE	MAE[1]	MPPE	MAE	MAE[1]	MPPE	MAE	MAE[1]	MPPE	MAE	MPPE	MAE
π/4	93.8	13.0	16.1	86.1	14.1	18.3	59.9	17.4	31.5	78.5	12.9	-	24.8
π/6	89.4	11.9	14.9	79.1	12.4	16.3	50.1	13.6	22.7	75.5	9.0	-	-
π/8	83.9	11.1	13.2	71.3	10.9	14.5	40.5	10.7	17.3	69.8	7.2	41.6	-
π/9	78.3	10.3	12.3	65.2	9.9	13.5	36.7	9.6	15.1	71.8	6.2	-	-
π/18	46.6	5.5	6.9	43.5	6.2	6.6	19.2	4.8	8.0	45.8	5.2	-	-

Figure 6: Improvement due to the Viewpoint and Shape refinement step

SUMMARY

- Good results for class-specific shape and view reconstruction on state-of-the-art.
- Fine-grained parts and linear subspace representation model deformation effectively and true to life, but with far fewer vertices than full shape.
- Our method fast but still efficient and also models the projection process more accurately than regression, refinement step clearly helps.
- Aim: To explore unsupervised methods with more image evidence.