LEARNING CATEGORY-SPECIFIC 3D SHAPE MODELS FROM WEAKLY LABELED 2D IMAGES

Dingwen Zhang[1,2] Junwei Han[1], Yang Yang[1], and Dong Huang[2]

[1]Northwestern Polytechnical University, [2]Carnegie Mellon University

IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 HONOLULU 2017

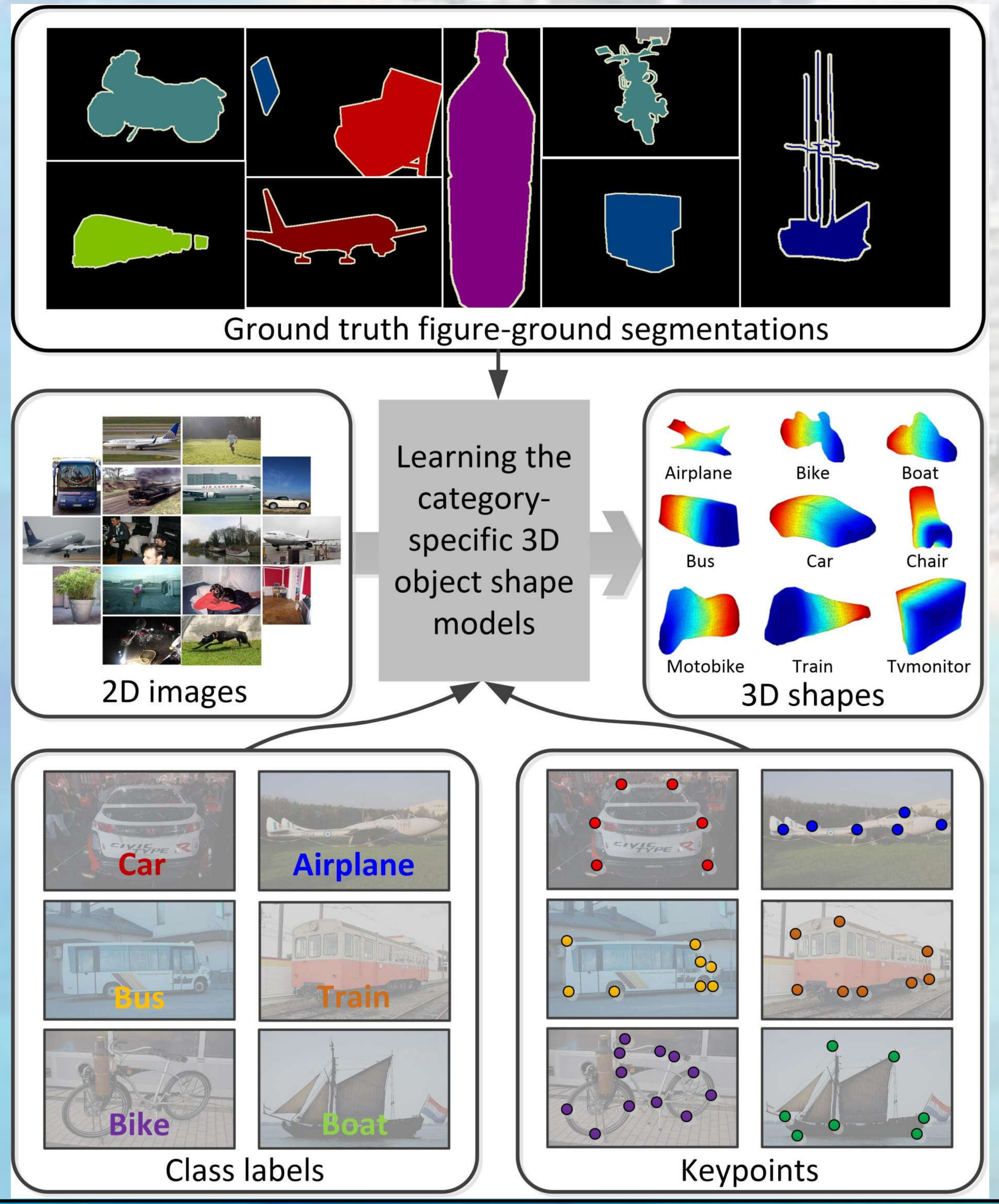
Problem

NEW

Weakly Annotated 2D Images: Images only annotated with the class labels and a small number of keypoints; the object segmentation masks (the most time consuming for 2D manual annotation) are not needed.

Goal: learn 3D shape models from weakly labeled 2D images; Reconstructing **PASCAL** objects using the learnt 3D models!.

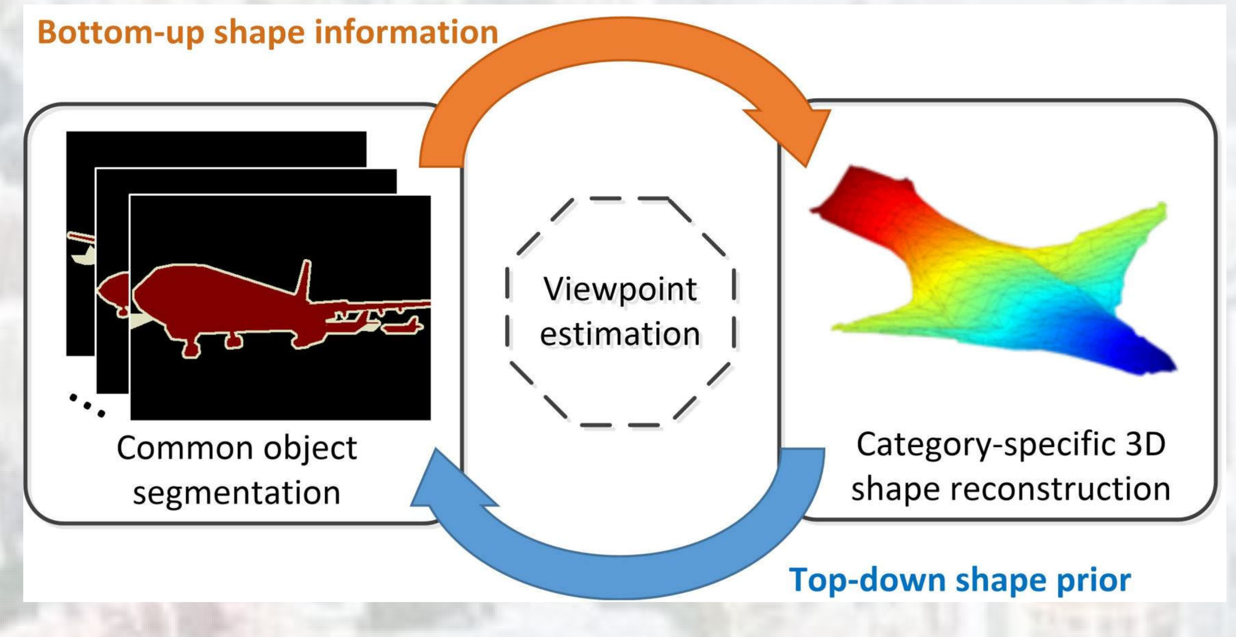
Alleviate human labor: Ground truth figure-ground segmentations (256.1s per image with LabelMe) and 3D shape training data (CAD).



Analysis

Solution: Jointly address two sub-tasks:

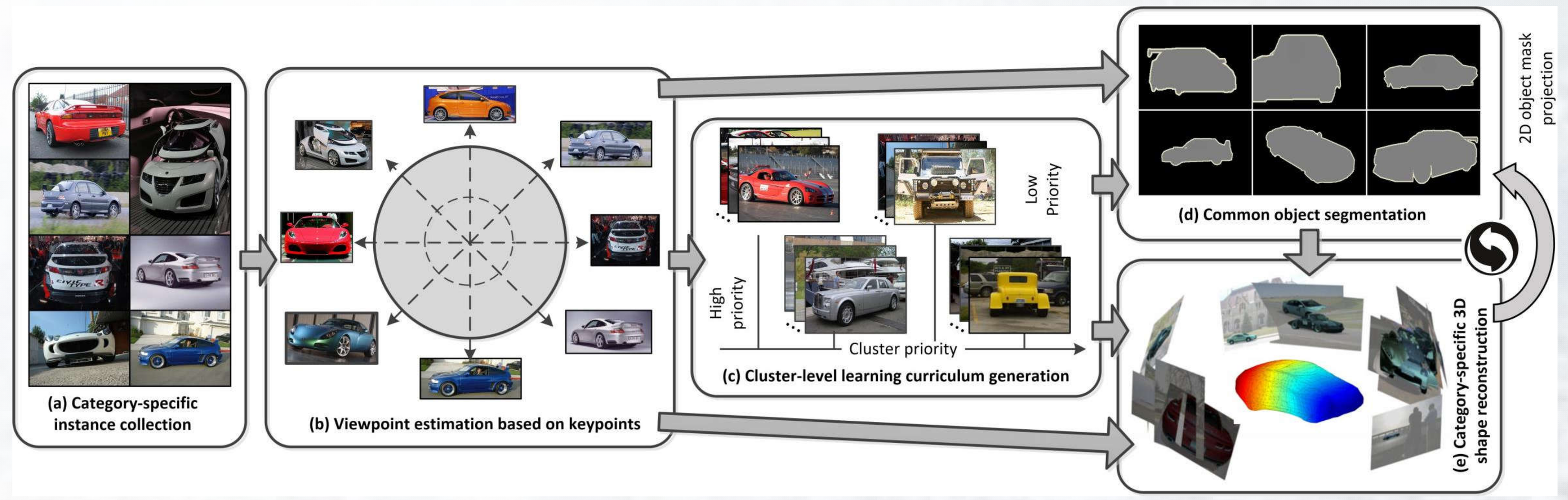
- **common object segmentation**, i.e., segmenting the common objects appearing in the image collection of a certain object category.
- **Category-specific 3D object shape reconstruction**, i.e., learning the category-specific 3D shape models for the co-occurring objects.



Relationship: they can work compatibly and help each other:

- The figure-ground object masks generated by **common object segmentation** helps providing bottom-up shape cues for building **category-specific 3D shape models**.
- The 3D shape models built by **category-specific 3D shape reconstruction** provides helpful yet under-explored top-down priors for **common object segmentation**.

Methods



Input: Weakly labelled images (images & keypoints)

Step 1: Estimating viewpoints based on the given keypoints via NRSfM.

Step 2: Generating Cluster-Level Learning Curriculum (**Two-stage clustering based on K-means** & **Generating learning curriculum by considering shape completeness and appearance compactness**.)

Step 3: Initializing 2D object segmentation masks.

Step 4: **Category-specific 3D reconstruction:**

$$\min_{\overline{\mathbf{Sh}},\mathcal{V},\alpha} E_{lc}(\overline{\mathbf{Sh}},\mathcal{V}) + E_{pd}(\alpha,\mathcal{V}) + \sum_n \left(E_{sc}(\mathbf{Sh}_n,\mathbf{O}_n,\pi_n) + E_{ns}(\mathbf{Sh}_n)\right),$$

$$s.t.\ \ \mathbf{Sh}_n = \overline{\mathbf{Sh}} + \sum_k \alpha_n^k \mathbf{V}_k,$$

Step 5: **Co-segmentation under 3D prior:**

$$\min_{L} E_I(\tau,p;A_\tau) + E_W(\tau,p,q;I_{\tau,p},I_{\tau,q}) + E_{TD}(\tau,p;\mathbf{SM},\mathbf{PM}),$$

$$E_{TD}(\tau,p;\mathbf{SM},\mathbf{PM}) = -\log p(I_{\tau,p}|\mathbf{SM},p) - \log p(I_{\tau,p}|\mathbf{PM},p),$$

Step 6: move to **Step 4** until converge.

Output: The learnt **category-specific 3D object model** and the **object segmentation masks**.

Experiments

Dataset: Sublet from **PASCAL VOC 2012**!

Evaluation of 3D Shape Models

Setting: Use the learnt 3D shape models to generate object mesh and depth map for each test image by following the standard pipeline of Tulsiani's TPAMI16.

Table 2. Comparing the learnt 3D shape models obtained the proposed approach with the Baslines and state-of-the-arts (STAs) in terms of the Mesh error (the less the better).

	Categories→	aero	bike	boat	bus	car	chair	mbike	sofa	train	tv	mean
Baselines	RC w/o SG	2.04	4.09	4.29	3.21	2.34	3.36	2.34	6.36	8.83	9.49	4.64
	LN w/o CL	1.95	3.40	4.32	3.01	2.43	2.78	2.30	6.61	8.73	9.12	4.46
	OURS	1.87	3.00	4.15	2.96	2.24	2.32	2.22	5.83	8.01	8.31	4.09
STAs	Tulsiani's [25]	1.72	1.78	3.01	1.90	1.77	2.18	1.88	2.13	2.39	3.28	2.20
	Vicente's [27]	1.87	1.87	2.51	2.36	1.41	2.42	1.82	2.31	3.10	3.39	2.31
	Twarog's [26]	3.30	2.52	2.90	3.32	2.82	3.09	2.58	2.53	3.92	3.31	3.03
	OURS	1.87	3.00	4.15	2.96	2.24	2.32	2.22	5.83	8.01	8.31	4.09

Evaluation of Object Segmentation Masks

Table 3. Comparing the segmentation results of our approach and other baselines and STAs in terms of the IOU (the higher the better).

	Categories→	aero	bike	boat	bus	car	chair	mbike	sofa	train	tv	mean
Baselines	RC w/o SG	0.714	0.572	0.669	0.753	0.790	0.673	0.717	0.794	0.678	0.741	0.710
	LN w/o CL	0.726	0.596	0.647	0.814	0.756	0.663	0.713	0.784	0.687	0.752	0.714
	OURS	0.737	0.614	0.673	0.825	0.794	0.720	0.738	0.865	0.692	0.771	0.743
STAs	Quan's [24]	0.729	0.481	0.644	0.764	0.788	0.608	0.743	0.831	0.666	0.648	0.690
	Chen's [9]	0.684	0.544	0.585	0.739	0.749	0.650	0.654	0.891	0.670	0.723	0.689
	Joulin's [18]	0.279	0.336	0.239	0.378	0.319	0.236	0.334	0.435	0.363	0.260	0.318
	OURS	0.737	0.614	0.673	0.825	0.794	0.720	0.738	0.865	0.692	0.771	0.743

- **RC w/o SG**: Reconstruction without segmentation, i.e., directly using the initial segmentation masks.
- **LN w/o CL**: learning without curriculum, i.e., using all training images in each learning iteration.