Learning Category-Specific 3D Shape Models from Weakly Labeled 2D Images

Dingwen Zhang[1,2], Junwei Han[1]*, Yang Yang[1], Dong Huang[2]
[1]Northwestern Polytechnical University [2]Carnegie Mellon University
{zhangdingwen2006yyy,junweihan2010,Tp030ny}@gmail.com,dghuang@andrew.cmu.edu

Abstract

Recently, researchers have made great processes to build category-specific 3D shape models from 2D images with manual annotations consisting of class labels, keypoints, and ground truth figure-ground segmentations. However, the annotation of figure-ground segmentations is still labor-intensive and time-consuming. To further alleviate the burden of providing such manual annotations, we make the earliest effort to learn category-specific 3D shape models by only using weakly labeled 2D images. By revealing the underlying relationship between the tasks of common object segmentation and category-specific 3D shape reconstruction, we propose a novel framework to jointly solve these two problems along a cluster-level learning curriculum. Comprehensive experiments on the challenging PASCAL VOC benchmark demonstrate that the category-specific 3D shape models trained using our weakly supervised learning framework could, to some extent, approach the performance of the state-of-the-art methods using expensive manual segmentation annotations. In addition, the experiments also demonstrate the effectiveness of using 3D shape models for helping common object segmentation.

1. Introduction

Nowadays, learning object models to recognize object categories, discover object locations, and segment object masks from given images has been widely studied and could almost achieve performance that approaches the human expectation. However, another interesting and meaningful problem—constructing rich internal representation, such as the depth information and 3D poses, of the presented objects—still remains to be challenging and understudied. To address this problem, one promising solution is to build 3D shape models of the objects appearing in the given images, which could be potentially utilized to generate the depth maps [28], perform pose correspondence [39], and construct the specific instances in each single image [17].

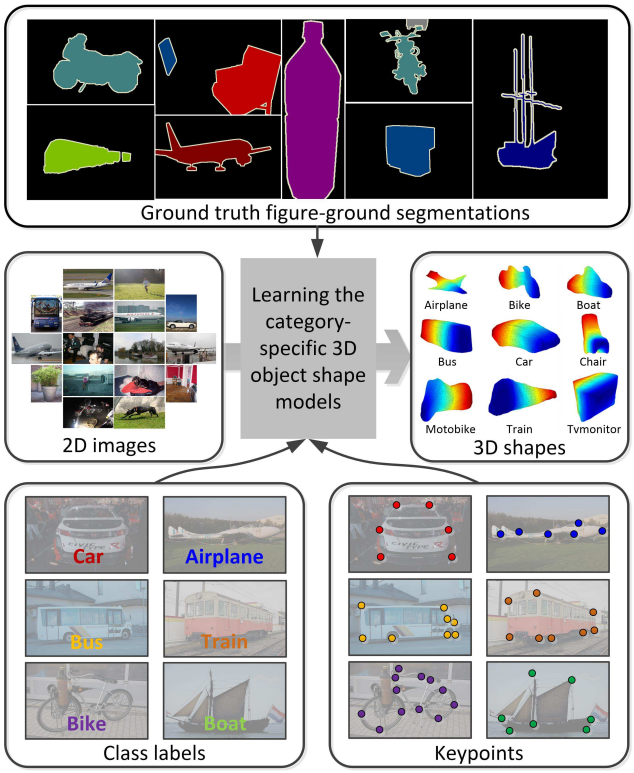

Figure 1. Illustration of the most recent progress [20, 7] in learning 3D shape models from 2D images, which could infer 3D shape models by using the 2D annotations, including the class labels, ground truth figure-ground segmentations, and a small set of keypoints, while the extra 3D shape training data are not required. Compared with these methods, this paper aims to learning 3D shape models even without using the ground truth figure-ground segmentations, which could significantly alleviate the manual annotation efforts for learning 3D shape models.

So how to construct the 3D shape models? In some early approaches, such as [22, 26], a precise 3D shape model of the target object is manually provided in advance. In other words, the 3D shape models used by these approaches to reconstruct the instances in each given image are obtained directly by the human design. Another group of approaches were proposed to leverage the 3D shape training data (e.g. those obtained by 3D scanning) to reconstruct the object-

*Corresponding author.

s in the given images via prototype alignment [3, 28], using morphable models [1, 40] or deep neural networks [10]. Kolev et al. [21] attempted to estimate the 3D geometry of the object under the guidance of user-interaction. However, their approach could only work on the calibrated images captured under manually controlled environment. Similarly, [25] also needed to constrain the input images from fixed camera locations. By relying on manual design, 3D scanning, or manually controlled imaging environment to obtain the 3D shape models, it is infeasible to use the aforementioned approaches for the images in the wild, which often contain unknown object categories and various image scenes. To solve this problem, some most recent approaches [7, 20] have made efforts to learn 3D shape models without using any 3D shape training data. As shown in Fig. 1, these approaches could successfully build 3D object shape models by only relying on the 2D image data manually annotated with class labels, ground truth figure-ground segmentations, and a small set of keypoints, which provides a solution to reconstruct dense, per-object 3D shapes for images in popular object detection datasets, e.g., PASCAL VOC [14] and ImageNet [11].

Along this line of research, this paper makes efforts to further significantly alleviate the burden of providing manual annotations for learning 3D shape models. As shown in Fig. 1, although the annotations of image labels and keypoints nowadays could be easily crowdsourced with Amazon Mechanical Turk by requiring only a few clicks per image, providing ground-truth annotations of the figure-ground segmentation masks still remains to be labor-intensive and time-consuming. Thus, we propose to make the earliest attempt to study how to learn 3D shape models from weakly labeled 2D images[1] which are defined as:

Definition (Weakly labeled 2D images:) *Images only annotated with the class labels and a small number of keypoints while the object segmentation masks (the most time-consuming for 2D manual annotation [2]) are not needed.*

Apparently, learning 3D shapes from such weakly labeled 2D data tends to be much more challenging. However, it is of great significance as it could lead 3D shape modeling to a unprecedented cheap fashion and thus facilitate large-scale practical applications.

For learning category-specific 3D shape models from weakly labeled 2D images, we propose to jointly address two sub-tasks simultaneously: 1) segmenting the common objects appearing in the image collection of a certain object category (i.e., common object segmentation), and 2) learning the category-specific 3D shape models for the co-occurring objects of the image collection (i.e., category-specific 3D shape reconstruction). Essentially, there is delicate relationship between these two tasks, which, however, still remains to be under-explored. To our best knowledge, both of common object segmentation and category-specific 3D shape reconstruction need to leverage the global shape information from multiple images rather than just processing each singe image separately. By exploring the global shape information from the image collections, it would inevitably suffer from large intra-class variations in terms of varying shapes, textures, sizes, and viewpoints (see Fig. 2). Thus, to better capture the global shape information, both of them need to carefully explore the low-frequency base shape and simultaneously handle the high-frequency details. Besides such common properties, it is more interesting that these two tasks could actually work compatibly and help each other (see Fig. 3):

Figure 2. Examples from the PASCAL VOC dataset to illustrate that practical image collections always exhibit significant intra-class variability, making it challenging for both common object segmentation and category-specific 3D shape reconstruction.

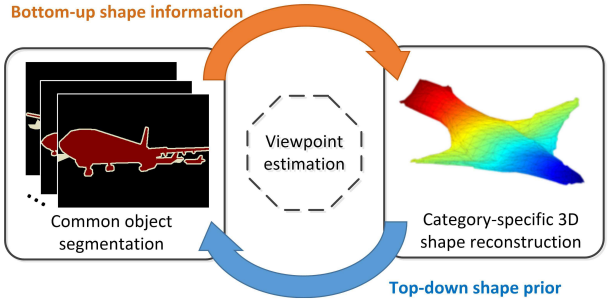

Figure 3. Through viewpoint estimation, the two sub-tasks discussed in this work (i.e., common object segmentation and category-specific 3D shape reconstruction) could help each other by providing useful information to the other.

Common object segmentation helps category-specific 3D shape reconstruction: The figure-ground object masks generated by common object segmentation could help providing informative bottom-up shape cues for building category-specific 3D shape models.

Category-specific 3D shape reconstruction helps common object segmentation: The 3D shape models built by category-specific 3D shape reconstruction technology could provide helpful yet under-explored top-down priors for common object segmentation.

[1] The "weak label" defined here is different from those in [35, 19, 15].

[2] According to our statistics, the time cost for manually annotating class labels, keypoints, and segmentation masks are around 1.2s, 4.4s, and 256.1s per image, respectively.

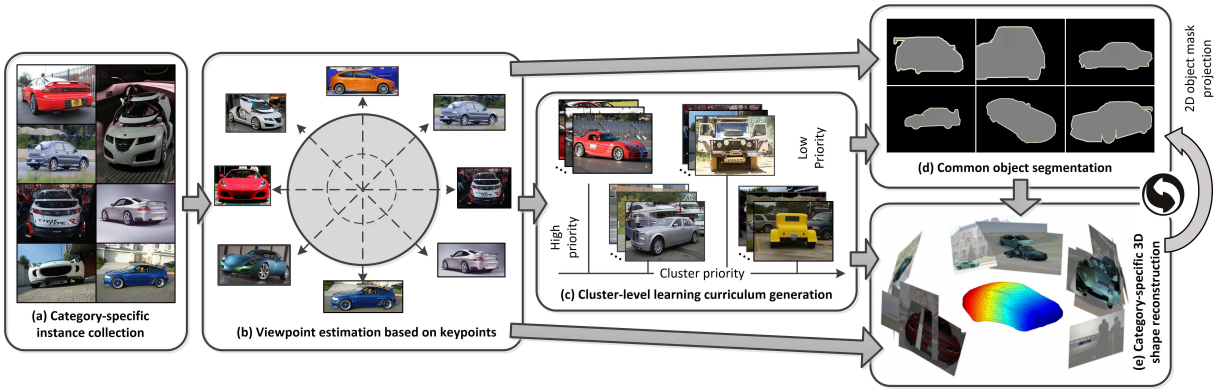

Figure 4. The proposed framework for learning category-specific 3D shape models from weakly labeled 2D image data.

Based on the above observation and discussion, we propose a novel framework to jointly conduct common object segmentation and category-specific 3D shape reconstruction, which leads to the implementation of learning category-specific 3D shape models from weakly labeled 2D images. As shown in Fig. 4, we first collect the category-specific instances by using the provided class labels and keypoints. Then, with the help of the annotated keypoints, we adopt the technique of recovering non-rigid 3D shape from image streams (NRSfM) [6] to estimate the camera viewpoint parameters for the training instances in each object category. Afterwards, inspired by the most recent progress in curriculum learning [8, 33], we design a cluster-level learning curriculum to guide the learning of the category-specific 3D shape models. Basically, we first decompose the category-specific instance collection into subgroups and then build a learning curriculum to encourage subgroups with more compact appearance and complete shape to be learnt with higher priority. Afterwards, the learner would gradually infer the object segmentation masks and category-specific 3D shape models along the established learning curriculum in an iterative fashion. Specifically, we first infer the object masks within the subgroups with high priority. Then, we use the obtained object masks to reconstruct coarse category-specific 3D shape models. The obtained 3D shape models could in turn provide the top-down prior for common object segmentation via a viewpoint guided-2D mask projection. Finally, the whole learning framework could obtain meaningful results including the segmentation masks and the category-specific 3D shape models.

We have three major contributions in this paper:

- We make the earliest effort to learn category-specific 3D shape models only from weakly labeled 2D images, which could largely save the time and labor for providing the figure-ground segmentation annotations manually. It is also of great significance to lead 3D shape modeling to a unprecedented cheap fashion and thus facilitate large-scale practical applications.
- By discovering the underlying relationship between the problems of common object segmentation and category-specific 3D shape reconstruction, we propose a novel framework for jointly solving these two problems along a learning curriculum, which gradually realizes the learning of category-specific 3D shape models from weakly labeled 2D images.
- Comprehensive experiments have been conducted to demonstrate the effectiveness of the proposed framework. Encouragingly, category-specific 3D shape models trained using our weakly supervised framework approach the performance of some state-of-the-art methods using a large amount of manual segmentation-level annotation. In addition, we also demonstrate the effectiveness by using 3D shape models for helping common object segmentation.

2. The Proposed Approach

2.1. Viewpoint Estimation

In order to estimate the camera viewpoint parameters for all the training instances in the category-specific instance collection, we follow [20] to adopt the NRSfM approach [6]. Here the category-specific instance collection is obtained by cropping the images with the corresponding class labels by using the rectangles to enclose the annotated keypoints. Given K_p keypoint correspondences per instance $n \in 1, 2, \cdots, N$, where N is the total number of instances, the NRSfM algorithm is used to maximize the likelihood of the following formulation:

$$\begin{aligned} &\mathbf{P}_n = s_n \mathbf{R}_n \mathbf{W}_n + \mathbf{1}^T \mathbf{T}_n + \mathbf{H}_n, \\ &\mathbf{W}_n = \overline{\mathbf{W}} + \sum\nolimits_k \mathbf{U}_k z_n^k, \\ &z_n^k \sim \mathcal{N}(0,1), H_{n,\iota} \sim \mathcal{N}(0,\sigma^2), \\ &k \in [1,m], \iota \in [1,K_p], \\ &s.t. \quad \mathbf{R}_n \mathbf{R}_n^T = \mathbf{I}_2, \end{aligned} \tag{1}$$

where I_2 indicates the 2×2 identity matrix, $\mathbf{P}_n$ is the provided keypoints which could also be formulated as the 2D projection of the 3D keypoints $\mathbf{W}_n$ with the white noise $\mathbf{H}_n$ and the camera parameters containing the orthographic projection matrix $\mathbf{R}_n$, scale s_n and 2D translation $\mathbf{T}_n$. The $\mathbf{W}_n$ is parameterized as a factored Gaussian with a mean shape $\overline{\mathbf{W}}$, m basis vectors $\mathcal{U} = \{\mathbf{U}_1, \mathbf{U}_2, \cdots, \mathbf{U}_m\}$ and the latent deformation parameters $\mathbf{z}_n$. We follow [6] and [20] to adopt the EM-PPCA algorithm to maximize the likelihood of the above formulation. With the input data $\mathbf{P}_n$, the algorithm infers the 3D keypoints $\mathbf{W}_n$ for all training instances as well as the project function $\pi_n \equiv \{s_n, \mathbf{R}_n, \mathbf{T}_n\}$.

2.2. Cluster-Level Learning Curriculum

2.2.1 Two-Stage Clustering

As we know, both common object segmentation and category-specific 3D shape reconstruction need to acquire the global shape information from the category-specific instance collection. However, directly exploring the entire category-specific instance collection can hardly acquire strong global shape information due to large intra-class variations in terms of different viewpoints and varying shapes, textures, and sizes (see Fig. 5). Inspired by the recent work [9], we build the priors based on the viewpoint-specific visual subgroups through a two-stage clustering strategy in order to gradually decompose the entire image collection into multiple clusters with much lower intra-group variations. Consequently, we could easily capture the meaningful priors from such visual subgroups and use them to build the global shape information in an effective way.

Specifically, in the first stage, we utilize the estimated camera parameters $\{s_n, \mathbf{R}_n, \mathbf{T}_n\}$ to describe each object instance and adopt the K-means clustering method to separate the entire category-specific instance collection into K_c viewpoint-specific clusters, i.e., $\{C_1, C_2, \cdots, C_{K_c}\}$. As shown in Fig. 5, the clusters obtained by this step usually contain the instances with similar viewpoints. Thus it could alleviate learning ambiguity caused by viewpoint variations. The next stage is to further alleviate the intra-class variations caused by other factors, such as the varying shapes, textures, and sizes. Here we adopt the seed-based clustering approach [9] to generate a set of subgroups within each viewpoint-specific cluster due to its superior capability in grouping visually coherent instances together. In this stage, for each viewpoint-specific cluster, we first use each instance as a seed and then build groups by detecting similar instances from the rest data. This is implemented by training the exemplar detectors eLDA [16] based on the HOG feautres of each instance, and then using each detector to group similar instances by selecting the top K_e detections with the highest scores. Thus, K_e is the number of instances in each subgroup. Suppose the c-th viewpoint-specific cluster $C_c, c \in [1, K_c]$ contains n_c instances, we could finally obtain $K_g = \sum_{c=1}^{K_c} n_c$ subgroups $\{G_1, G_2, \cdots, G_{K_g}\}$.

2.2.2 Cluster-Level Object Co-Segmentation

After obtaining the subgroups via the proposed two-stage clustering, we adopt the Seed Segmentation approach [9] to initialize the segmentation masks of the instances. Basically, the problem is casted as a classical graph cut problem [32] to label every pixel in every input image as the foreground or background, which can be solved by minimizing the energy function with an image-level unary potential term, a cluster-level unary potential term, and a pairwise potential term.

2.2.3 Learning Curriculum Generation

To guide the subsequent learning procedure in an effective way, we design a learning curriculum to gradually adapt the faithful knowledge from "easy" to "hard" training samples. This is closely related to the field of curriculum learning, which was originally proposed in [5] and has been successfully used in other applications like object detection and recognition [8, 38]. In our framework, the "easy" training samples are the ones with more compact appearance and complete shape masks, which mainly encode the low-frequency base shape and thus should be learnt with higher priority. One the contrary, the "hard" training samples are the ones with more diverse appearance and sometimes noisy shape masks, which might contain the high-frequency details but need to be learnt in the more "mature" stage.

To measure the appearance compactness of the g-th subgroup $G_g, g \in [1, K_g]$, we train K_e eLDA detectors to score each instance. Specifically, for the τ-th instance in G_g, we can obtain K_e detection scores $\{S_g^{\tau,1}, \cdots, S_g^{\tau,K_e}\}$. Then we binarize these detection scores to obtain the hit numbers $\{h_g^{\tau,1}, \cdots, h_g^{\tau,K_e}\}$ by using the threshold t. Afterwards, we use the mean hit numbers of all the instances in G_g as the compactness score CP_g. For obtaining the shape completeness scores SC_g for each subgroup, we compute the Pearson Linear Correlation Coefficient (PLCC) [27] of the mask-based distance matrix $\mathbf{D}_g^{mask}$ and the image-based distance matrix $\mathbf{D}_g^{img}$, which is based on the assumption that complete shape masks and the instance images should show similar feature-similarity distributions in subgroup with high shape completeness. Specifically, we extract the HOG features of the shape masks and the instance images, respectively, and use them to generate $\mathbf{D}_g^{mask}$ and $\mathbf{D}_g^{img}$ based on the Euclidean distance. As the HOG features extracted from each instance image capture the complete shape/contour information of the corresponding object instance, our assumption could work in real cases. Finally, the learning priority of each subgroup is obtained by $LP_g = SC_g \times CP_g$.

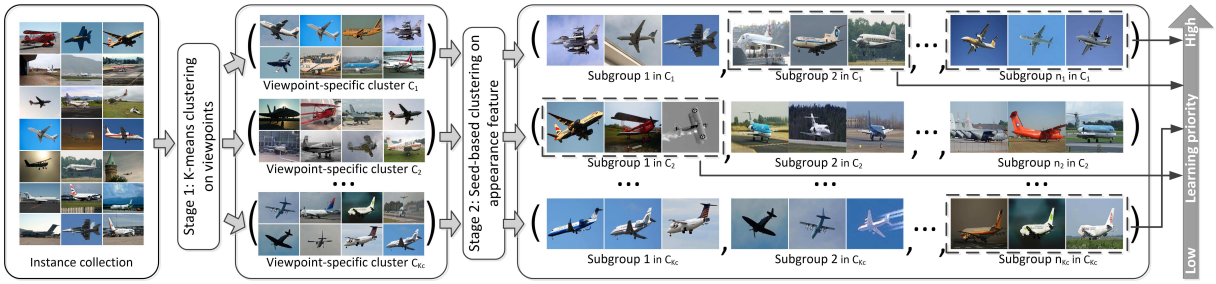

Figure 5. Examples to illustrate the generation of the cluster-level learning curriculum.

As shown in Fig. 5, the subgroups with higher learning priority tend to be the more "easy" ones that should be leant at earlier iterations. Thus, along this learning curriculum, we design a five-round learning iteration, where we start by using the subgroups of the top 30% learning priority to infer the category-specific 3D shape models and then gradually involve richer knowledge from more subgroups, i.e., additional 5% after each iteration, into the learning procedure, which could improve both the segmentation masks and the category-specific 3D shape models.

2.3. Joint Common Object Segmentation and Category-Specific 3D Shape Reconstruction

By discovering the underlying relationship between the problems of common object segmentation and category-specific 3D shape reconstruction and observing that they can provide useful information for each other, we propose to jointly conduct common object segmentation and category-specific 3D shape reconstruction in each learning iteration.

2.3.1 Category-specific 3D Shape Reconstruction

Based on the estimated camera projection parameters, keypoint correspondences, and the common object segmentation masks on the selected training subgroups, we follow [20] to build deformable 3D shape models from the object silhouettes. Specifically, the 3D shape models are formulated as $\mathcal{M} = (\overline{\mathbf{Sh}}, \mathcal{V})$, which consists of a mean shape $\overline{\mathbf{Sh}}$ and a set of deformation bases $\mathcal{V} = \{\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_m\}$. It could be inferred via the following energy function:

$$\begin{aligned} &\min_{\overline{\mathbf{Sh}},\mathcal{V},\alpha} E_{lc}(\overline{\mathbf{Sh}}, \mathcal{V}) + E_{pd}(\alpha, \mathcal{V}) \\ &\qquad + \sum\nolimits_n (E_{sc}(\mathbf{Sh}_n, \mathbf{O}_n, \pi_n) + E_{ns}(\mathbf{Sh}_n)), \\ &s.t. \quad \mathbf{Sh}_n = \overline{\mathbf{Sh}} + \sum\nolimits_k \alpha_n^k \mathbf{V}_k, \end{aligned} \tag{2}$$

where $\mathbf{O}_n$ and $\mathbf{Sh}_n$ denote the instance silhouettes and 3D shape of the n-th instance, respectively, α is the deformation parameter. The local consistency term $E_{lc}(\overline{\mathbf{Sh}}, \mathcal{V})$, which is used to restrict arbitrary deformations, is defined as:

$$\begin{aligned} E_{lc}(\overline{\mathbf{Sh}}, \mathcal{V}) = &\sum\nolimits_i \sum\nolimits_{j \in N(i)} ((\parallel \overline{Sh}_i - \overline{Sh}_j \parallel - \delta)^2 \\ &+ \sum\nolimits_k \parallel V_{k,i} - V_{k,j} \parallel^2), \end{aligned} \tag{3}$$

where δ represents the mean squared displacement between the neighboring points $N(\cdot)$, which encourages all faces to have similar sizes, $V_{k,i}$ is the i-th point in the k-th basis, $\overline{Sh}_i$ is the i-th point in $\overline{Sh}$.

$E_{ns}(\mathbf{Sh}_n)$ is the normal smoothness term, which places a cost on the variation of normal directions in a local shape neighborhood as shape change tends to be locally smooth. Specifically, it is formulated as:

$$E_{ns}(\mathbf{Sh}_n) = \sum\nolimits_i \sum\nolimits_{j \in N(i)} (1 - \overrightarrow{\mathcal{N}}_{n,i} \cdot \overrightarrow{\mathcal{N}}_{n,j}), \tag{4}$$

where $\overrightarrow{\mathcal{N}}_{n,i}$ denotes the normal for the i-th point in $\mathbf{Sh}_n$. It is computed by fitting planes to local point neighborhoods.

$E_{sc}(\mathbf{Sh}_n, \mathbf{O}_n, \pi_n)$ is the shape consistency term:

$$\begin{aligned} E_{sc}(\mathbf{Sh}_n, \mathbf{O}_n, \pi_n) = &\sum\nolimits_{Ch^{mask}(p)>0} \triangle^1(p; \mathbf{O}_n) \\ &+ \sum\nolimits_{p \in \mathbf{O}_n} \triangle^2(p; \pi_n(\mathbf{Sh}_n)), \end{aligned} \tag{5}$$

where Ch^{mask} refers to the Chamfer distance of the binary mask of silhouette $\mathbf{O}_n$, $\triangle^1(p; \mathbf{O}_n)$ indicates the squared average distance of pixel p to its nearest neighbors in set $\mathbf{O}_n$, $\triangle^2(p; \pi_n(\mathbf{Sh}_n))$ indicates the squared average distance of pixel p to its two nearest neighbors in the 2D projection π_n of shape $\mathbf{Sh}_n$, $\pi_n(\mathbf{Sh}_n) = s_n \mathbf{R}_n \mathbf{Sh}_n + \mathbf{1}^T \mathbf{T}_n$. The first term enforces the predicted shape to project inside its silhouette, while the second term encourages the points on the silhouette to pull nearby projected points towards them.

The last term $E_{pd}(\alpha, \mathcal{V})$ is used to penalize the L_2 norm of the deformation parameter α in order to prevent unnaturally large deformations. Specifically, it is defined as:

$$E_{pd}(\alpha, \mathcal{V}) = \sum\nolimits_n \sum\nolimits_k \parallel \alpha_n^k \mathbf{V}_k \parallel_F^2 . \tag{6}$$

As the objective in (2) is highly non-convex and non-smooth, we follow [12] to initialize mean shape with a soft visual hull, which is computed by using the selected training instances. The deformation bases and deformation weights are initialized randomly.

2.3.2 Common Object Segmentation Using 3D Shape Projection Prior

Once we obtain the category-specific 3D shape models, we use them to provide informative top-down priors for guiding the common object segmentation in the next learning iteration. Specifically, for each training cluster containing K_e instance images, our goal is to label each pixel to be foreground $l_{\tau,p} = 1$ or background $l_{\tau,p} = 0$, where p denotes the pixel location in image $\tau, \tau \in [1, K_e]$. Such labeling problem can be solved by minimizing an energy function over pixels and labels:

$$\begin{aligned} E_I(\tau, p; A_\tau) + E_W(\tau, p, q; l_{\tau,p}, l_{\tau,q}) \\ + E_{TD}(\tau, p; \mathbf{SM}, \mathbf{PM}), \end{aligned} \tag{7}$$

which mainly contains three terms. The first term $E_I(\tau, p; A_\tau)$ is the unary potential from an appearance model specific to the instance image τ:

$$E_I(\tau, p; A_\tau) = -\log p(l_{\tau,p}; \mathbf{x}_{\tau,p}, A_\tau), \tag{8}$$

where $p(l_{\tau,p}; \mathbf{x}_{\tau,p}, A_\tau)$ evaluates how likely a pixel with its RGB color feature $\mathbf{x}_{\tau,p}$ is to take label $l_{\tau,p}$, according to the appearance model A_τ. Here A_τ consists of two Gaussian mixture models (GMM) over the RGB color space as defined in [9], i.e., one for the foreground (when $l_{\tau,p} = 1$) and the other for the background (when $l_{\tau,p} = 0$). The appearance model is learnt using the pixels inside and outside the segmentation masks inferred from the previous iteration.

The second term $E_W(\tau, p, q; l_{\tau,p}, l_{\tau,q})$ is the pairwise potential defined as:

$$E_W(\tau, p, q; l_{\tau,p}, l_{\tau,q}) = \delta(l_{\tau,p} \neq l_{\tau,q}) e^{-\beta \|\mathbf{x}_{\tau,p} - \mathbf{x}_{\tau,q}\|^2}, \tag{9}$$

which penalizes two pixels (p and q) when they are assigned with different labels but having similar features.

The third term $E_{TD}(\tau, p; \mathbf{SM}, \mathbf{PM})$ is the top-down prior term introduced to help common object segmentation, which encourages the obtained segmentation masks within each subgroup to be consistent. This is modeled as the top-down shape priors over pixels:

$$\begin{aligned} E_{TD}(\tau, p; \mathbf{SM}, \mathbf{PM}) = \\ - \log p(l_{\tau,p} | \mathbf{SM}, p) - \log p(l_{\tau,p} | \mathbf{PM}, p), \end{aligned} \tag{10}$$

where **SM** indicates the average segmentation mask of the subgroup. Thus $-\log p(l_{\tau,p}|\mathbf{SM}, p)$ denotes the prior probability that each pixel belongs to the foreground or background, given the pixel location and **SM**. Similarly, **PM** indicates the average projection shape mask of the subgroup obtained by using the 3D shape model $\mathcal{M} = (\overline{\mathbf{Sh}}, \mathcal{V})$ and the deformation parameter α from the previous iteration:

$$\begin{aligned} \mathbf{PM} &= \frac{1}{\tau} \sum\nolimits_\tau (s_\tau \mathbf{R}_\tau \mathbf{Sh}_\tau + \mathbf{1}^T \mathbf{T}_\tau) \\ \mathbf{Sh}_\tau &= \overline{\mathbf{Sh}} + \sum\nolimits_k \mathbf{V}_k \alpha_\tau^k, \end{aligned} \tag{11}$$

where the parameters $\{s_\tau, \mathbf{R}_\tau, \mathbf{T}_\tau\}$ are obtained from Sec. 2.1. Thus $-\log p(l_{\tau,p}|\mathbf{PM}, p)$ could effectively introduce the top-down prior provided by the category-specific 3D shape models to common object segmentation, and the entire energy (7) can be connivently minimized by using the graph-cut algorithm [32].

3. Experiments

3.1. Experimental Settings

We performed our experiments on the dataset collected in [29]. It contains 10 rigid object categories selected from the challenge PASCAL VOC 2012 benchmark [13]. We used the publicly available category-specific keypoints during the learning phase and adopted the ground-truth segmentation masks for evaluation. For evaluating the expressiveness of our learned 3D models, we adopted the 3D CAD models provided by the PASCAL3D+ dataset [34]. During training, we only used the images containing one instance. The additional localization scheme, which can apply the trained 3D shapes to segment multiple instances in singe images, is a promising future direction but beyond the scope of this paper.

For comprehensively demonstrating the effectiveness of the proposed approach, we conducted experiments to evaluate the obtained category-specific 3D shape models and the common object segmentation masks. Specifically, we quantify the quality of the obtained 3D models on the test set based on two metrics. The first one is the mesh error metric, which is computed as the Hausdorff distance between the predicted mesh and the ground-truth mesh [2]. The second one is the depth map error, which is measured as the mean absolute distance between the reconstructed depth and the ground truth depth. It can reflect the quality of the reconstructed visible object surface [29]. For quantifying the quality of the obtained segmentation masks, we adopted the standard intersection-over-union (IOU) metric by comparing each segmentation mask and the corresponding ground-truth mask. During our implementation, we set K_c around 4 and K_e as 3. In addition, we followed the previous work [29] to set m as 5.

The experiments were run on a 24-core Lenovo Server with an Intel Xeon CPU of 2.8-GHz and 64-GB RAM. Our method takes 8.93 hours for training, which is slower than [29] (4.39 hours). Notice that the latter needs much more time for manually labelling the ground truth segmentation masks. For test, our method takes 38s per image, which is the same as [29].

3.2. Evaluation of 3D Shape Reconstruction

In this section, we first conducted experiments to demonstrate that the proposed framework can effectively learn the 3D shape models from weakly labeled 2D images by com-

Table 1. Comparing the learnt 3D shape models obtained the proposed approach with the weakly supervised baseline methods in terms of the Mesh error and Depth error (the lower the better).

	Categories→	aero	bike	boat	bus	car	chair	mbike	sofa	train	tv	mean
Mesh	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
Depth	RC w/o SG	10.77	14.73	17.13	18.51	11.22	10.72	11.73	26.60	37.50	36.84	19.58
	LN w/o CL	10.77	13.79	17.44	16.55	11.21	10.72	11.29	28.00	36.46	29.57	18.58
	OURS	10.68	13.53	17.03	18.06	10.28	11.07	11.18	27.32	36.23	26.39	18.18

Table 2. Comparing the learnt 3D shape models obtained the proposed approach with the state-of-the-arts (STAs) in terms of the Mesh error and Depth error (the lower the better). Notice that all the STAs require stronger supervision than the proposed approach.

	Categories→	aero	bike	boat	bus	car	chair	mbike	sofa	train	tv	mean
Mesh	Tulsiani's [29]	1.72	1.78	3.01	1.90	1.77	2.18	1.88	2.13	2.39	3.28	2.20
	Vicente's [31]	1.87	1.87	2.51	2.36	1.41	2.42	1.82	2.31	3.10	3.39	2.31
	Twarog's [30]	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
Depth	Tulsiani's [29]	9.51	9.27	17.20	12.71	9.94	7.78	9.61	13.70	31.58	8.78	13.01
	Vicente's [31]	10.05	9.28	15.06	18.51	8.14	7.98	9.38	13.71	31.25	8.33	13.17
	Barron's [4]	13.52	13.79	20.78	29.93	22.48	18.59	16.80	18.28	40.56	20.18	21.49
	OURS	10.68	13.53	17.03	18.06	10.28	11.07	11.18	27.32	36.23	26.39	18.18

paring our work with the baseline methods of "RC w/o SG" and "LN w/o CL". Specifically, "RC w/o SG" (Reconstruction without segmentation) directly utilized the initial co-segmentation masks of all the training images to reconstruct the category-specific 3D shape models, while "LN w/o CL" (learning without curriculum) jointly conducted category-specific 3D shape reconstruction and common object segmentation without adopting the learning curriculum. The experimental results are reported in Table 1. From Table 1, we have two observations: 1) Learning curriculum is important for coping with the learning ambiguity when reconstructing the category-specific 3D shapes from weakly labeled 2D images (see the comparison between OURS and "LN w/o CL"). 2) Category-specific 3D shape reconstruction and common object segmentation could potentially help each other to improve the learning performance (see the comparison between "RC w/o SG" and "LN w/o CL").

We also compared the 3D shape models learnt by using the proposed approach with those from several state-of-the-art methods, including Tulsiani's [29], Vicente's [31], Twarog's [30], and Barron's [4]. When compared with our approach, all of the state-of-the-art methods need to additionally use a large amount of manually labeled segmentation masks. However, from the experimental results in Table 2, we observe that our approach achieves encouraging performance which approaches to some state-of-the-art methods using much stronger supervision in the categories like "boat" and "aero". Our approach even beats Barron's [4] in the categories like "car" and "mbike". Our approach does not perform well on the categories like "train" and "sofa", due to the lack of sufficient training data.

3.3. Evaluation of Common Object Segmentation

In this section, we conducted experiments to demonstrate that the proposed framework can also segment common objects effectively by leveraging the learnt 3D shape models. Firstly, we compared the proposed approach with another two baseline frameworks, i.e., "RC w/o SG" and "LN w/o CL". The experimental results are reported in the top part of Table 3. As can be seen, the proposed approach significantly outperforms these two baselines. In addition, being consistent with the expressiveness of the learned 3D models, "LN w/o CL" can outperform "RC w/o SG" by jointly inferring the category-specific 3D shapes and common object segmentation masks. However, it still achieves worse performance than the proposed approach due to the lack of an effective learning curriculum.

We also compared the segmentation masks obtained by using the obtained 3D shape models to segment the common objects via (7) with those obtained from three state-of-the-art (STA) object co-segmentation methods, including Quan's [24], Chen's [9], and Joulin's [18]. For fair comparison, we also provided the keypoint information to help the STA methods, i.e., by using these methods to segment on the cropped instance images. From the experimental results in the bottom part of Table 3, we observe that the proposed approach outperforms all the compared STAs, especially for the categories like "bike" and "chair". Thus, the experiment demonstrates the effectiveness by using the top-down priors from the 3D shape models (even learnt from weakly labeled images) for helping common object segmentation. Finally, we show some experimental results in Fig. 6.

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

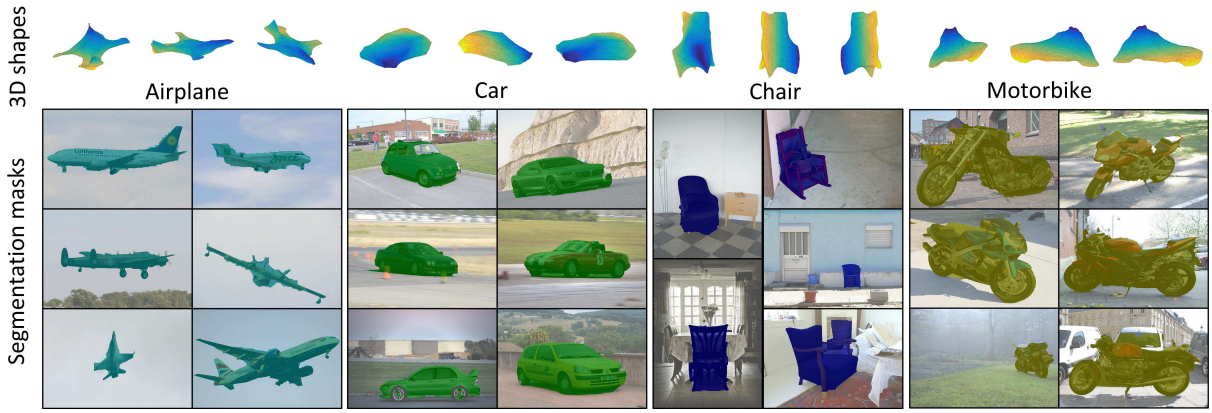

Figure 6. Examples of the 3D shape models (shown in different views) and segmentation masks obtained by the proposed approach.

3.4. Discussion

In this section, we further compared and discussed the proposed method with the state-of-the art method [31]. According to our statistics (see Sec. 1), annotating figure-ground segmentations takes 97.9% of the entire human effort for annotating 2D images. It means that using our approach can save such amount of human effort for 3D shape reconstruction, which is significant. We also conducted the following experiments: 1) We randomly selected 10% data for testing and used different percentages of data for training. The blue curve in Fig. 7 demonstrates that our approach boosts the learning performance along with the increase of the weakly labeled training data. 2) We also used 20% fully labeled data (annotated with the additional segmentation masks) for training 3D shapes using [31] and tested on the same set of testing data. As shown in Fig. 7, although the annotation cost (human labor) for 90% weakly labeled data is actually much less than that for 20% fully labeled data, training our model based on the former even outperforms training [31]'s model based on the latter. In a sense, this experiment demonstrates the potential value for applying our approach in large-scale image data, where human can only annotate the segmentation masks for a small part of them.

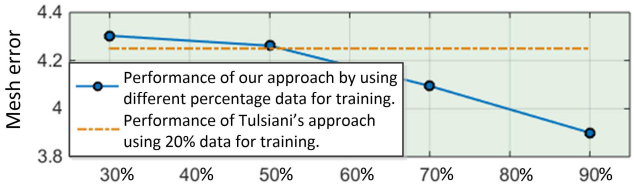

Figure 7. Performance comparison of the proposed approach using different amount of weakly labeled training data and Tulsiani's approach [31] using 20% fully labeled training data.

4. Conclusion

This paper has proposed to learn category-specific 3D shape models under the supervision of weakly labeled 2D images without using any manually annotated segmentation masks. It significantly saves the time and labor for providing manual annotations, which leads 3D shape modeling to a unprecedented cheap fashion. In this paper, we implemented it by establishing a novel framework to jointly conduct common object segmentation and category-specific 3D shape reconstruction along a cluster-level learning curriculum. Comprehensive experiments on the PASCAL VOC dataset have demonstrated the effectiveness of the proposed framework as well as the potential value for applying it in large-scale image data. In the future, we will introduce more effective co-saliency detection [37, 36] or co-segmentation methods [24] in this problem and use some better 3D object detection pipelines like [23].

Acknowledgement: This work was supported in part by the National Science Foundation of China under Grant 61473231 and Grant 61522207, in part by the Excellent Doctorate Foundation of Northwestern Polytechnical University, and in part by the China Scholarships Council under Grant 201506290113.

References

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. *ACM TOG*, 24(3):408–416, 2005.

[2] N. Aspert, D. Santa Cruz, and T. Ebrahimi. Mesh: measuring errors between surfaces using the hausdorff distance. In *ICME*, 2002.

[3] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014.

[4] J. T. Barron and J. Malik. Color constancy, intrinsic images, and shape estimation. In *ECCV*, 2012.

[5] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009.

[6] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, 2000.

[7] J. Carreira, S. Vicente, L. Agapito, and J. Batista. Lifting object detection datasets into 3d. *IEEE TPAMI*, 38(7):1342–1355, 2016.

[8] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *ICCV*, 2015.

[9] X. Chen, A. Shrivastava, and A. Gupta. Enriching visual knowledge bases via object discovery and segmentation. In *CVPR*, 2014.

[10] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[12] C. H. Esteban and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. *CVIU*, 96(3):367–392, 2004.

[13] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2012 (voc2012), 2012.

[14] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[15] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE TGRS*, 53(6):3325–3337, 2015.

[16] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *ECCV*, 2012.

[17] Q. Huang, H. Wang, and V. Koltun. Single-view reconstruction via joint analysis of image and shape collections. *ACM TOG*, 34(4):87, 2015.

[18] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, 2010.

[19] A. Kanazawa, D. W. Jacobs, and M. Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *CVPR*, 2016.

[20] A. Kar, S. Tulsiani, J. Carreira, and J. Malik. Category-specific object reconstruction from a single image. In *CVPR*, 2015.

[21] K. Kolev, T. Brox, and D. Cremers. Fast joint estimation of silhouettes and dense 3d geometry from multiple images. *IEEE TPAMI*, 34(3):493–505, 2012.

[22] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *AI*, 31(3):355–395, 1987.

[23] B. Pepik, M. Stark, P. Gehler, T. Ritschel, and B. Schiele. 3d object class detection in the wild. In *CVPRW*, 2015.

[24] R. Quan, J. Han, D. Zhang, and F. Nie. Object co-segmentation via graph optimized-flexible manifold ranking. In *CVPR*, 2016.

[25] D. J. Rezende, S. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess. Unsupervised learning of 3d structure from images. In *NIPS*, 2016.

[26] L. G. Roberts. *Machine perception of three-dimensional soups*. PhD thesis, MIT, 1963.

[27] S. M. Stigler. Francis galton's account of the invention of correlation. *Statistical Science*, pages 73–79, 1989.

[28] H. Su, Q. Huang, N. J. Mitra, Y. Li, and L. Guibas. Estimating image depth using shape collections. *ACM TOG*, 33(4):37, 2014.

[29] S. Tulsiani, A. Kar, J. Carreira, and J. Malik. Learning category-specific deformable 3d models for object reconstruction. *IEEE TPAMI*, 2016.

[30] N. R. Twarog, M. F. Tappen, and E. H. Adelson. Playing with puffball: simple scale-invariant inflation for use in vision and graphics. In *ACM SPA*, 2012.

[31] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc. In *CVPR*, 2014.

[32] S. Vicente, V. Kolmogorov, and C. Rother. Graph cut based image segmentation with connectivity priors. In *CVPR*, 2008.

[33] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, Y. Zhao, and S. Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *arXiv preprint arXiv:1509.03150*, 2015.

[34] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014.

[35] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo. Semantic annotation of high-resolution satellite images via weakly supervised learning. *IEEE TGRS*, 54(6):3660–3671, 2016.

[36] D. Zhang, J. Han, J. Han, and L. Shao. Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining. *IEEE TNNLS*, 27(6):1163–1176, 2016.

[37] D. Zhang, J. Han, C. Li, J. Wang, and X. Li. Detection of co-salient objects by looking deep and wide. *IJCV*, 120(2):215–232, 2016.

[38] D. Zhang, D. Meng, L. Zhao, and J. Han. Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. In *IJCAI*, 2016.

[39] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang, and A. A. Efros. Learning dense correspondence via 3d-guided cycle consistency. In *CVPR*, 2016.

[40] M. Z. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3d representations for object recognition and modeling. *IEEE TPAMI*, 35(11):2608–2623, 2013.